COUNCIL OF EUROPE CONSEIL DE L'EUROPE

**December 2004** **DGIV/EDU/LANG (2004) 13**

# Reference Supplement

**to the**

**Preliminary Pilot version of the Manual for**

***Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment***

Language Policy Division, Strasbourg

# CONTENTS

# Foreword

The Language Policy Division of the Council of Europe in Strasbourg has published a "Preliminary Pilot Version of a Proposed Manual: "Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF)" DGIV/EDU/LANG(2003) 5 in order to assist member States, national and international providers of examinations in relating their certificates and diplomas to the *Common European Framework of Reference for Languages*.

This Reference publication accompanies the Pilot Manual. Its aim is to provide the users of the Pilot Manual with additional information which will help them in their efforts to relate their certificates and diplomas to the CEF.

During the work on the Pilot Manual it was agreed that the Reference Supplement would contain three main components: *quantitative and qualitative considerations in relating certificates and diplomas to the CEF and different approaches in standard setting.*

Dr. Norman Verhelst (member of the Authoring Group for the Manual), Dr. Jayanti Banerjee (Lancaster University) and Dr. Felianka Kaftandjieva (University of Sophia) undertook to write the various sections of the Reference Supplement and Dr. Sauli Takala to edit the publication. The authors have revised their contributions on the basis of comments from the editor. There have also been some comments from the other members of the Authoring Group and from the ad hoc advisory group. However, the authors have final responsibility for their texts.

The authors' goal has been to try to make their contributions as readable as possible. They have avoided technical language (formulas, symbols etc) as far as possible and provided concrete examples, figures and tables to illustrate the exposition. However, demanding subject matter cannot be simplified beyond a certain point without risking oversimplification. Indeed, one of the authors' main concerns has been to caution about oversimplifications that many "rules of thumb" imply. The authors have, by contrast, tried to promote thoughtful application of various methods and approaches. With some effort, all persons working in language testing and assessment will be able to grasp the essentials and will have gained a deeper understanding of how to construct better tests and examinations and especially how to assess their quality. They will also be more aware of the complexities involved in relating certificates and diplomas to the CEF.

Section A of the Reference Supplement provides a short overview of the linking process. This section is drawn from the Manual and is provided to help readers remind themselves of the approach proposed.

Dr. Felianka Kaftandjieva has written Section B on Standard setting. She has done considerable amount of work on standard setting specifically in relation to the CEF.

In Section B, the author notes that the link between language examinations and the Common European Framework for Language (CEF) can be established in at least three different ways:

- direct linkage to the CEF scales of language proficiency
- indirect linkage via linkage to some local scales of language proficiency which have already been linked to the CEF scales
- indirect linkage via equation to an existing test already linked to the CEF scales.

Whatever approach is adopted in the particular concrete situation, the author stresses that the linkage always requires standard setting and thus standard setting is a key element in the linkage process. Section B underlines the potentially very high stakes of the examinations for the examinees, and seeks to promote better understanding by providing a review of the current status of standard setting, its theoretical framework and still unresolved issues. Section B does this by:

- giving a brief overview of the main trends in the development of standard setting methodology
- describing the major unresolved issues and controversial points
- discussing some of the major factors that affect standard setting decisions and their quality
- presenting some of the most common methods for standard setting
- outlining the validation process and providing evaluation criteria for the technical quality of the standard setting
- describing the main steps in standard setting procedures, and
- presenting some basic recommendations and guidelines for standard setting.

It will be obvious from the thorough review in Section B that there are several possible approaches for standard setting in relation to CEF and the approach presented in the Manual is not the only appropriate one. Whatever approach is chosen, the validity of the claimed linkage depends on how well the various activities were carried out and how thoroughly and appropriately the results are reported.

Section C, written by Dr. Norman Verhelst, gives an overview of the main concepts and theoretical foundations of Classical Test Theory (CTT). Classical Test Theory has been used for more than fifty years as a guide for test constructors to understand the statistical properties of test scores, and to use these properties to optimise the quality of the test under construction in a number of ways. Section C reviews the main issues of Classical Test Theory and shows what can and cannot be expected from CTT. First, some basic concepts are presented followed by a discussion of procedures which are used in the framework of Classical Test Theory.

As the author's goal has been to make the text as accessible as possible for the non-technical reader, the first two sections (Basic Concepts and Procedures) do not contain any formulae. However, the author notes that as CTT is a statistical theory, it is not possible to present and discuss it in great depth without having recourse to the exact and compact mode of expression provided by mathematical formulae and, therefore, reference is made to formulae in a more technical section. These more technical sections are stand-alone elements, and follow the main text in the order they are referred to.

Section D, on qualitative analysis methods, is written by Dr. Jayanti Banerjee. The chapter provides an extensive overview of the range of qualitative methods available for investigating test quality. It demonstrates a large variety of options available and explains the key features of each, covering the following topics: an overview of qualitative methods, verbal reports, diary studies, discourse/conversation analysis, analysis of test language, data collection frameworks, task characteristic frameworks, questionnaires, checklists and interviews.

In addition, examples of research using the methods are provided to illustrate how specific qualitative methods have been implemented.

The author suggests that many of the methods described could also be used as part of standard setting procedures and illustrates this in sub-section 6: Using qualitative methods in standard setting.

The author concludes that qualitative methods have considerable potential to explain and augment the statistical evidence we gather to assess test quality. Many of the methods are complementary and can be used for the triangulation of data sources. The importance of the validity and generalizability of the data collection methods is stressed in order to legitimise the inferences drawn from them.

Section E, by Dr. Norman Verhelst, deals with Generalizability Theory and contains four parts. The first two parts give a non-technical introduction into generalizability theory. In the third and fourth sections the same problems are treated in a somewhat more technical way.

The author notes that a very basic term of Classical Test Theory is not well defined: reference is made to repeated observations under 'similar' conditions, but 'similar' is not defined precisely.

A traditional way of controling for systematic effects is to try to standardize test administration as far as possible and feasible. Generalizability Theory was launched in the early 1970s to provide a method for assessing the effect of various factors on the measurement results. In the theory, measurements are described in terms of the conditions where they are observed. A set of conditions that belong together is called a facet. In this way, items and raters are facets of the measurement procedure.

Two important conditions in language testing are dealt with in more detail: the one-facet crossed design (persons by items) and the two-facet crossed design (persons by items by raters), and the possible application of Generalizability Theory in deciding on the optimal number of items and raters is demonstrated.

The author also discusses a problem which is commonly overlooked in using Generalizability Theory: typically every rater rates the same performances of the students to the task instead of every student generating an independent response for each rater. Yet, the design is treated as a two facet crossed design, which is not the case. This leads, in fact, to two different sources of measurement error: one attached to the student-task combination and one attached to the rater. This is a fundamental difference with the crossed model.

Section F, by Dr. Norman Verhelst, deals with a topic which has been a subject of discussion and debate in language testing for some time: *is language competence a unitary (unidimensional) or a multidimensional phenomenon*? If a test consists of several subtests, is it meaningful to report a single score or should test scores be reported separately for each subtest (in a profile)? Section F presents Factor Analysis - a well-established method (developed more than a hundred years ago) to test the dimensionality of the test in order to decide whether to report results using a single score or several scores. The author notes that although factor analysis was not defined originally as such, the model fits very well in the family of IRT-models discussed in Section G.

Section G, also by Dr Norman Verhelst, deals with the relatively more recent Item Response Theory (IRT). It consists of four non-technical sections (containing no formulae) where basic notions of IRT are explained and discussed. A number of notions and techniques are then discussed in a more formal and technical style. The author has strived to avoid the use of formulae as much as possible, making extensive use of graphical displays. To help the reader in constructing graphs using his/her own materials and using modern computer technology, a special section has been added with a step by step explanation of how most of the graphs in the section were produced.

Whereas the basic notion in Classical Test Theory is the true score (on a particular test), in Item Response Theory (IRT) the concept to be measured (in our case, language proficiency) is central in the approach.

Basically, this concept is considered an unobservable or latent variable, which can be of a qualitative or a quantitative nature. If it is qualitative, persons belong to (unobserved) classes or types (of language proficiency); if it is quantitative, persons can be represented by numbers or points on a line. Only the latter case is dealt with in Section G.

One of the most attractive advantages of IRT is the possibility to carry out meaningful measurement in incomplete designs: it is possible to compare test takers with respect to some proficiency even if they did not all take the same test. This happens in Computer Adaptive Testing (CAT), where the items are selected during the process of test taking so as to fit optimally with the level of proficiency as currently estimated during test taking. Incomplete designs are also used in paper-and-pencil formats. Use of IRT methods requires a lot of technical know-how. This is sometimes packed in attractive software, and some users of this software may think that the problem is nothing more than technical know-how. The author warns that this is a naive way of thinking: the advantages of IRT are available if, and only if, the theoretical assumptions on which the theory is built are fulfilled. Therefore it is the responsibility of all users applying IRT to check these assumptions as carefully as possible. IRT methods are more powerful than methods based on classical test theory, but they may mistakenly be considered a methodology that ensures high quality assessment. The author, who has co-authored a very powerful IRT- programme called OPLM (One Parameter Logistic Model), warns against over-optimism which may be promoted by some enthusiastic proponents of IRT: " … using an IRT-model does not convert a bad test into a good one. A careless construction process cannot be compensated by a use of the Rasch model; on the contrary, the more carelessly the test is composed, the greater the risk that a thorough testing of the model assumptions will reveal the bad quality of the test." One practical consequence is that a separate assessment of the test reliability is always needed (preferably before IRT modeling) since it cannot be inferred from statistical tests of goodness-of-fit provided by software.

The originally planned Section H on Test Equating will appear later in the Revised Reference Supplement.

As editor of the Reference Supplement I am confident that it will prove very useful for the language testing and assessment community in general. It contains information which is not readily available in the mainstream language testing literature. More specifically it will provide good support for those who wish to contribute to the development of the Manual by providing feedback, by piloting the Manual and by writing case studies on some aspects or the whole process of linking examinations to the CEF and hopefully it will contribute to improvement of language testing quality.

Feedback and comments on the Reference Supplement are invited. Please contact Johanna Panthier at Johanna.Panthier@coe.int

December, 2004

Sauli Takala

# Section A

# Overview of the Linking Process

The Manual for relating examinations to the Common European Framework of Reference for Languages (CEFR) presents four inter-related sets of procedures that users are advised to follow in order to design a linking scheme in terms of self-contained, manageable activities. All of the activities carried out in all four sets of procedures contribute to the validation process.

> **Familiarisation:** a selection of activities designed to ensure that participants in the linking process have a detailed knowledge of the CEFR. This familiarisation stage is necessary at the start of both the Specification and the Standardisation procedures
>
> In terms of validation, these procedures are an indispensable starting point. An account of the activities taken and the results obtained is an essential preliminary component of the validation report.
>
> **Specification**: a self-audit of the coverage of the examination (content and tasks types) profiled in relation to the categories presented in CEFR Chapter 4 "Language use and the language learner" and CEFR Chapter 5 "The user/learner's competences." As well as serving a reporting function, this exercise also has a certain awareness-raising function that may assist in further improvement in the quality of the examination concerned.
>
> These procedures assure that the definition and production of the test have been undertaken carefully, following good practice.
>
> **Standardisation**: suggested procedures to facilitate the implementation of a common understanding of the "Common Reference Levels" presented in CEFR Chapter 3. Standardised exemplars will be provided to assist training in the standardisation of judgements.
>
> These procedures assure that judgements taken in rating performances reflect the constructs described in the CEF, and that decisions about task and item difficulty are taken in a principled manner on the basis of evidence from pre-testing as well as expert judgement.
>
> **Empirical Validation**: the collection and analysis of test data and ratings from assessments to provide evidence that both the examination itself and the linking to the CEFR are sound. Suggestions and criteria are provided for adequate and credible validation appropriate for different contexts.
>
> These procedures assure that the claims formulated through Specification and Standardisation ("test-under-construction") can indeed be confirmed when the examination is administered in practice ("test-in-action") and data on how persons belonging to the target population behave when the test is so administered becomes available.

Relating examinations to the CEFR can best be seen as a process of "building an argument" based on a theoretical rationale. As noted above, the central concept within this process is "validity".

Evidently it is first necessary to ensure **Familiarisation** with the CEFR (Chapter 3) before linking can effectively be undertaken.

Then before an examination can be linked to an external framework like the CEFR (external validity), it must demonstrate the validity of the construct, and the consistency and stability of the examination (internal validity). To prove internal and external validity, quantitative and qualitative methods can be combined. **Specification** (Chapter 4) can be seen as a qualitative method: providing evidence through content-based arguments. The actions which result in filling in forms A1 and A3-A7 in Chapter 4 focus

on the *internal* validity of the examinations. Forms A2 and A8-A20 focus in a *qualitative* way on the external validity. There are also quantitative methods for content validation but this Manual does not require their use.

**Standardisation** (Chapter 5) involves both qualitative and simple quantitative procedures - through training and comparison with calibrated test samples and performances - to prove external validity. While the activities are mainly qualitative in orientation, quantitative evidence of the degree of success in the standardisation of judgements is also required.

Finally, **Empirical Validation** (Chapter 6) uses quantitative procedures based on data collection and analysis to demonstrate firstly "internal validity" and secondly "external validity". Chapter 6 demonstrates that proper empirical validation requires considerable psychometric know-how, just as test construction does. If such experience is not available to the examination providers, it is recommended that they arrange sufficient training or obtain the services of a qualified psychometrician.

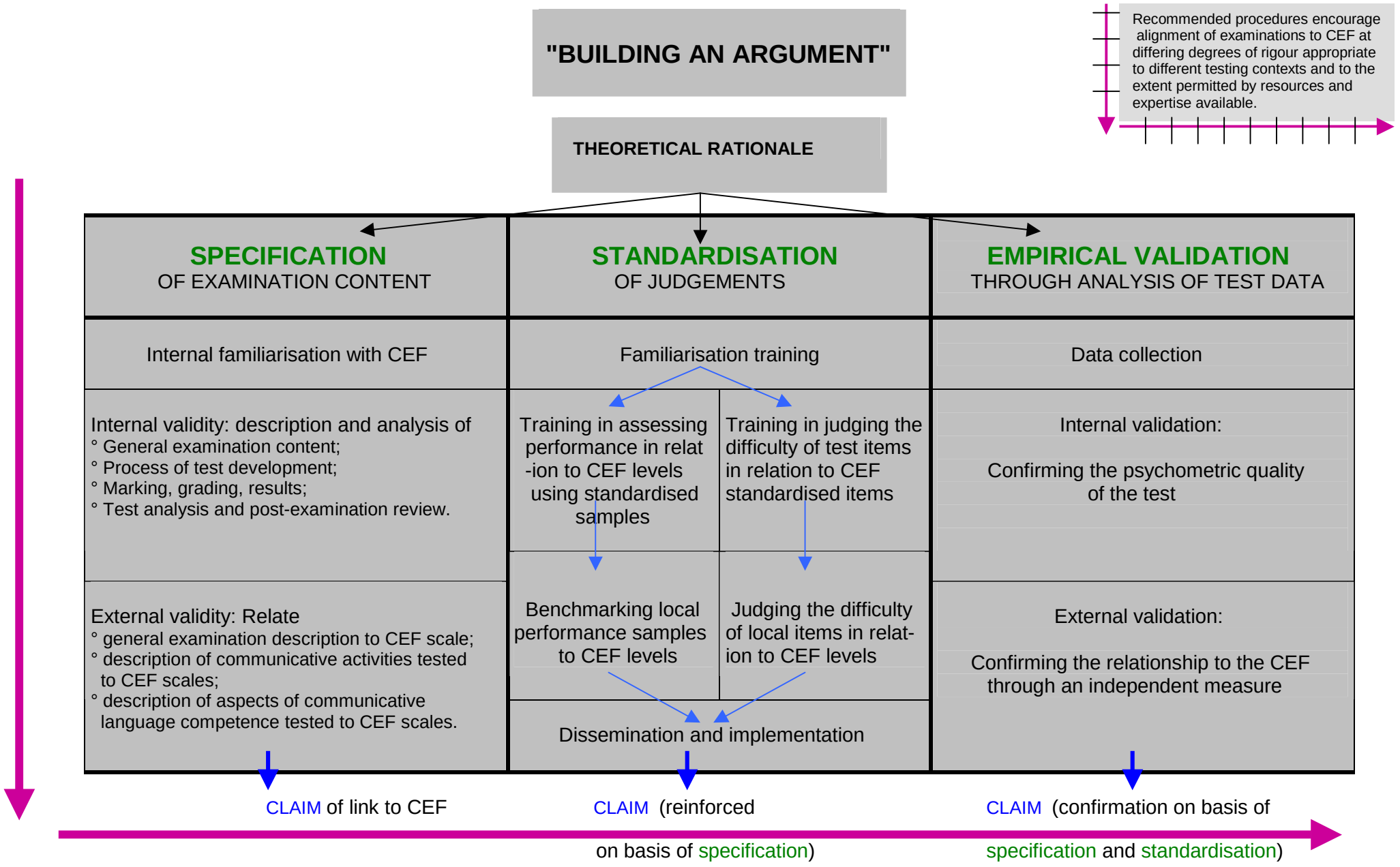
The approach adopted in this process is an inclusive one. The recommended procedures in each of the chapters mentioned above encourage alignment of examinations to the CEFR with differing degrees of rigour appropriate to different testing contexts. The Manual aims to encourage the application of principles of best practice even in situations with modest resources and expertise available. First steps may be modest, but the aim is to help examination providers to work within a structure, so that later work can build on what has been done before, and a common structure may offer the possibility for institutions to more easily pool efforts in certain areas.

The recommended techniques are organised in a logical order in such a way that all users will be able to follow the same broad approach. Users are encouraged to start with Familiarisation and are guided through the options offered by the techniques for each of Specification, Standardisation and Empirical. They are asked to identify, from the range of techniques and options offered and similar techniques in the literature, those most appropriate and feasible for their context.

Not all examination providers may consider they can undertake studies in all of the areas outlined above. Some institutions in “low-stakes” contexts may decide to concentrate on specification and standardisation, and may not be able to take the process to its logical conclusion of full-scale empirical validation as outlined in internationally recognised codes and standards for testing and measurement. However, it is highly recommended that even less well-resourced examination providers should select techniques from all three areas. The linking of a qualification to the CEFR will be far stronger if the claims based on test specifications and their content are supported by both standardisation of judgements and empirical validation of test data. Every examination provider - even examination providers that have only limited resources or countries that have decentralised traditions - should be able to demonstrate in one way or another through a selection of techniques both the internal quality and validity of their examination and its external validity: the validity of the claimed relationship to the CEFR.

The different elements in the linking scheme outlined above are shown in Figure 1.1.

**FIGURE 1.1: VISUAL REPRESENTATION OF PROCEDURES TO RELATE EXAMINATIONS TO THE CEF**



## SECTION B

## S T A N D A R D  S E T T I N G

**Felianka Kaftandjieva**

**University of Sofia**

*Si duo faciunt idem, non est idem.*
*If two people do the same thing, it is not the same.*
Terentius

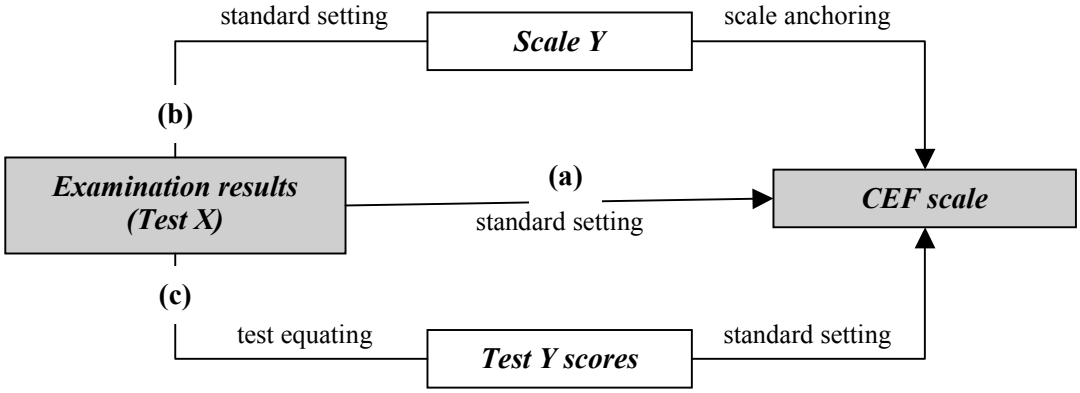
The linkage between language examinations and the Common European Framework for Language (CEF) means the establishment of a correspondence between examination results and CEF levels of language proficiency. This correspondence can be established in at least three different ways:

**a.** Direct linkage to the CEF scales of language proficiency

**b.** Indirect linkage via linkage to some local scales of language proficiency which has already been linked to CEF scales

**c.** Indirect linkage via equation to an existing test already linked to the CEF scales

**Fig. 1. Linkage Process**




As can be seen in Fig. 1, irrespective of the approach adopted in the particular concrete situation, the linkage always requires standard setting at a certain point. In other words, standard setting is at the core of the linkage process. Furthermore, bearing in mind the potentially very high stakes of the examinations for the examinees, the need for a more detailed review on the current status of standard setting, its theoretical framework and still unresolved issues is evident. In order to fill this need the current chapter sets the following objectives:

- to give a brief overview of the main trends in the development of standard setting methodology,
- to delineate the major unresolved issues and controversial points,
- to discuss some of the major factors affecting standard setting decisions and their quality,
- to present some of the most common methods for standard setting,
- to outline the validation process and provide evaluation criteria for the technical quality of the standard setting,
- to describe the main steps in standard setting procedures, and

- to submit some basic standard setting recommendations and guidelines.

**1. Basic Terminology**

The term '*standard setting*' in the field of educational measurement refers to a decision making process aiming to classify the results of examinations in a limited number of successive levels of achievement (proficiency, mastery, competency).

Two other terms which comprise the word 'standard' are closely related to *standard setting* and occasionally are used as counterparts although they are not synonyms (Hansche, 1998; Hambleton, 2001). These two terms are: content standards and performance standards. *Content standards* refer to the curriculum and answer the question: WHAT someone should know and be able to do as a result of a specific course of instruction? *Performance standards* on the other hand are "explicit definitions of what students must do to demonstrate proficiency at a specific level on the content standards" (CRESST Assessment Glossary, 1999) and answer the question: HOW good is good enough?

Hansche (1998) defines performance standards as a system including performance levels, performance descriptors, exemplars of student work at each level, and *cut-off scores* that separate the adjacent levels of performance. Therefore there is a symbiotic relationship between performance standards and cut-off scores where each cut-off score can be considered as "… an operational version of the corresponding performance standard" (Kane, 2001). Standard setting is usually focused on the establishment of these cut-off points on the scale, and hence it is closely affiliated to performance standards. There is also an indirect connection between standard setting and content standards, since performance standards are always related to some specific content standards.

It should be mentioned, however, that performance standards are not always defined as successive intervals on the scale in which examination results are presented and therefore they do not require an establishment of cut-off points on a continuum scale. Sometimes performance standards are presented only as verbal descriptions delineating different performance categories (Hambleton, 2001, p. 92). In language testing it usually takes place when productive skills like writing and speaking have been assessed. In such cases the examinees can be classified by raters directly into one of the six CEF performance levels matching examinee performance to the verbal descriptors of the corresponding CEF scale of language proficiency. In the current Manual this process is described in detail in Chapter 5 as **Benchmarking Performances –** a special case of a standard setting procedure, which requires no cut-off point establishment and therefore will not be discussed any further in the present chapter..

Alignment is another term which is very often used in connection with performance standards and standard setting. According to CRESST Assessment Glossary (1999) *alignment* is "the process of linking content and performance standards to assessment, instruction, and learning". Linn (2001) defines the alignment in narrower terms as "… the degree to which assessments adequately reflect standards". Hansche (1998), on the other hand, specifies two different dimensions of alignment: "(1) alignment of student, classroom, school, local, state, and national learning goals; and (2) alignment of content standards, curricula and instruction, performance standards, and assessments". It becomes evident from the definitions provided that alignment is closely related to validity in all its aspects: content, procedural, evidential and consequential basis.

A logical inference drawn on the above definitions of alignment is that standard setting is an integral part of the alignment process and as such is "… central to the task of giving meaning to test results and thus lies at the heart of validity argument" (Dylan, 1996).

Generally speaking, standard setting can be considered as a process of compressing the broad range of test scores into a limited number of rank-ordered categories (levels). Very often, especially in case of complex performance assessment, as it is usually the case with language assessment, standard setting is followed by another aggregation procedure aiming to combine the results of different performance

tasks (skills, dimensions) into a single score of overall performance. This procedure of combining the results of several standard setting procedures is called '*standard setting strategy'*. In spite of their great impact on the final decisions, standard setting strategies usually "… have received little attention in the testing literature thus far" (Haladyna & Hess, 2000, p. 130). Standard setting strategies are not the main focus of this chapter, either, but due to their significance to the consequences of standard setting they will be briefly described here.

The term '*standard setting strategy*' refers to the decision rule applied to combine the scoring results of a number of tasks (subtests, skills, traits) into a single score, usually expressed in terms of performance levels. In the educational setting the most often applied standard setting strategies are conjunctive, compensatory, and mixed strategies.

A *compensatory strategy* allows a high level of performance on one task (subtest, skill, trait) to compensate for a lower level of performance on some other task (subtest, skill, trait). The final decision in this case is based on the total score, and the compensatory strategy is, in fact, based on the assumption that '… the total score meaningfully reflects the construct' (Haladyna & Hess, 2000, p. 134). The reliability of the total score is usually higher than the reliability of its components especially if its components are highly inter-correlated, as is usually the case in the field of language testing. That is why many authors (Haladyna & Hess, 2000; Hambleton et al., 2000; Hansche, 1998) recommend the compensatory strategy to be preferred if other sound reasons do not entail the application of the conjunctive or mixed strategy.

A *conjunctive strategy* requires some a priori defined minimum level of performance to be reached on every single task (subtest, skill, trait) in order for the overall performance to be judged as satisfactory. Although "… the reliability data did not favor a conjunctive strategy" (Haladyna & Hess, 2000, p. 151), its use should be considered when each task (subtest, skill, trait) measures a unique aspect of the construct and the overall proficiency requires mastery on all components. More commonly such a situation arises in case of licensure and certification. For example to get a driver's license requires that someone should demonstrate both: (a) a satisfactory level of knowledge about the law as well as (b) a satisfactory level of driving skills, and a higher level on one of these two does not compensate for a low level on the other one.

If the different components are not equally important, then a mixed standard setting strategy might be implied. A *mixed* **(***hybrid***)** *standard setting strategy* requires a minimum level of performance on one or more tasks (subtest, skill, trait) allowing at the same time higher performance on some of the tasks to compensate for lower performance on some of the other tasks (Winter, 2001).

Another possible standard setting strategy, which is not typical for educational settings, is the *disjunctive standard setting strategy*, in which the satisfactory level of proficiency on only one task (sub-test, skill, trait) is considered enough for the overall satisfactory level of proficiency.

In discussing the choice of a standard setting strategy it should be mentioned that there is no best standard setting strategy. It is a matter of choice and whether the choice is good or bad depends entirely on the concrete circumstances and the consequences. In any case the consequential impact of the strategy choice should be explored before the final choice is made and the rationale for the strategy choice should be described and justified. The selection of standard setting strategy and its justification is an important and difficult issue, but it goes beyond the scope of this chapter and will not be discussed in the sequel.

## 2. Development of Standard Setting Methodology

As it was mentioned in the beginning, standard setting is a decision making process. With or without applying intentionally any specific methodology, human beings are involved in a number of decision

making processes on a daily basis. We constantly have to classify people and things and make choices, which only a posteriori, on basis of the consequences, can be judged to be good or bad choices. This is the reason for the roots of standard setting methodology to be traced by some authors back to ancient Egypt, China and the Old Testament (Green, 2000; Zieky, 2001).

Zieky distinguishes four distinct stages in the history of standard setting, which he called the ages of innocence, awakening, disillusionment, and realistic acceptance (cited in Stephenson et al., 2000). The long age of innocence ended in the mid 1950s. The period 1960-1980 was the era of awaking characterized by the invention a number of newly developed standard setting methods and extensive research. This era of awaking is closely connected with the rapid development of criterion-referenced testing.

The stage of disillusionment started with the first severe criticism, which came from Glass (1978) and concerns the arbitrary nature of standard setting. According to Glass (1978, p. 258) "… every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. But arbitrariness is no bogeyman, and one ought not to shrink from a necessary task because it involves arbitrary decisions. However, arbitrary decisions often entail substantial risks of disruption and dislocation. Less arbitrariness is safer".

Although Glass was villainized because of his strong criticism (Stone, 2002) his article had a great impact on the further development in the field of standard setting and led to a better understanding of the nature of the standard setting process.

Another effect of Glass's article is that his appeal to less arbitrariness has been repeated over the past 25 years by many other leading measurement specialists (Zieky, 2001). A quarter of a century after Glass, Linn (2003, p. 14) for example insists that: "Reports of individual student assessment results in terms of norms have more consistent meaning across different assessments than reports in terms of proficiency levels based on uncertain standards" and suggests "to shift away from standards-based reporting for uses where performance standards are not an essential part of the test use".

In response to Glass's criticism in 1978 Popham (1978, p. 298) argued that although standard setting is arbitrary it does not need to be capricious, but 20 years later he asserted that the main lessons he learned in a hard way were that "any quest for 'accurate' performance standard' is silly" (Popham, 1997). and that "the chief determiner of performance standards is not truth; it is consequences" (Popham, 1997).

The arbitrariness in fact is the Achilles' heel of standard setting and the most controversial issue. This fact is somewhat strange since the judgmental basis decision making as a whole is well recognized and does not provoke vehement discussions. There are three possible explanations for the causes of this long lasting debate on the arbitrary nature of standard setting.

- Firstly, the search for the absolute truth is somehow deep-seated in every human being. Epistemological anthropology reveals that the truth as such is not only a central concern of most cultures including pre-scientific ones, but also that "the desire for truth occupies a central role in workday cognitive practices such as magic, divination, and religion" (Goldman, 1999, p. 32).
- Secondly, the cut-off score establishment which usually follows the judgment process in many standard setting methods usually involves complex computational procedures aiming to aggregate expert judgments into a single cut-off score. In this way the judgmental character of the cut-off score is masked and "in turn gave the entire process a patina of professionalism and propriety" (Cizek, 2001, p. 7). In other words, the respect of numbers and the fact that the cut-off scores were established by a computer ('objectively'), not by a human being ('subjectively'), plays a practical joke in the interpretation of these cut-off scores.

- Thirdly, the every day decision making usually affects a limited number of people while standard setting has a great impact not only on the examinees being assessed, but also on further instructional and policy decisions. In other words, standard setting is a policy decision and as such it might become an object of criticism from all parties which had not been fully satisfied. According to Cizek (2001, p. 5) "standard setting is perhaps the branch of psychometrics that blends more artistic, political, and cultural ingredients into the mix of its products than any other".

The era of realistic acceptance started by 1983 when according to Zieky "setting cutscores has matured as a field" and transformed from "an esoteric topic limited to psychometricians or statisticians" to "a stuff of basic introductory text" in basic textbooks on educational measurement (Zieky, 2001, p. 25).

Summarizing Zieky's review (Zieky, 2001) of the evolution of standard setting development in the last 20 years the major changes are in the following directions.

### 2.1. CHANGES IN FOCUSES

- Increased emphasis on meeting rigorous cut-off scores

The shift from minimal competence testing to testing proficiency in more complex areas led to the development of more demanding tests and to the establishment of higher performance standards. Since higher performance standards lowered the pass rate, the demands for validity evidence concerning the established cutoff scores increased.

- Increased emphasis on the development of new standard setting methods

The switch from pass/fail decisions to multiple levels of proficiency on one hand and the increased use of performance assessment on the other hand called for the development of either new standard setting methods or modifications of the already existing methods in order to adjust them for the new conditions.

- Increased emphasis on the details of setting cut-off scores

The main shift in this direction was from comparative analysis of different standard setting methods toward more in depth analysis of the factors having greatest impact on the implementation of a given method. Research on the impact of different factors on the standard setting process still remains the central focus of the research agenda. Among the main factors affecting standard setting process are: (a) selection and number of judges involved in standard setting; (b) personal characteristics of judges (expertise, cognitive characteristics, decision making style, deliberation style, etc.); (c) amount and character of training; (d) social interaction in the group judgment; (e) type and amount of feedback, normative and impact data; and (f) number of iterative procedures.

- Increased concern about legal issues

The possibility (and the practice at least in the USA) for the cut-off scores of some high-stake examinations to be attacked on legal grounds increased the concern about legal issues and inspired the provision of more validity evidence especially in terms of adverse impact analysis (for a possible substantially different pass rate which works to the disadvantage of members of a race, sex, or ethnic group) and consequential validity arguments. The additional effect was that the need for providing legally defensible standards drew attention to better documentation on the standard setting procedures. More detailed descriptions of legal issues in standard setting can be found in Philips (2001), Carson (2001), Biddle (1993) and Cascio et al. (1988).

- Increased concern about fairness

Fairness of standard setting means that examinees who are on the same ability level will be classified into the same proficiency category irrespective of their gender, race, ethnicity, or disability. In other words, fairness means that in addition to the validity evidence about the whole population, validity evidence for each of the subpopulations is also needed.

## 2.2. CHANGES IN PROFESSIONAL STANDARDS IN TESTING

Every profession has its own Code of practice which includes a number of basic evaluation criteria of the quality of the work in this specific field. The *Standards for Educational and Psychological Testing* (AERA, NAPA, NCME) addresses professional and technical issues of test development and use in education, psychology and employment, and provides a number of definitive statements concerning the expected quality of the assessment instruments and they are the leading professionally recognized standards of sound testing practices within the educational measurement field.

The comparison of the standards concerning standard setting (Table 1) of the two consecutive editions of the *Standards for Educational and Psychological Testing* (1985 and 1999) reveals that the main changes are in the direction of:

(**a**) Increased number of technical standards about the quality of standard setting
The analysis of the standards in Table 1 shows that while the quality of standard setting in terms of standard error and validity of cut-off scores is mentioned only two times in the 1985 edition (Standards 2.10 and 5.11), in the 1999 edition the quality (reliability, standard error, stability, equivalence, agreement, pass rate, validity, etc.) of standard setting is mentioned in 7 standards (6.5, 4.20, 14.7, 1.7, 2.14, 2.15, 4.17);

(**b**) Greater attention has been paid to the content and procedural validity components
The content and procedural validity components are very vaguely mentioned in the 1985 edition (Standards 8.6, 6.9, 10.9, 5.11), whereas there are 11 standards (6.5, 4.4, 4.9, 4.19, 4.20, 14.7, 4.21, 1.7, 2.15, 6.12, 4.17) in the 1999 edition, which point out the rationale of the interpretations and the procedures for cut-off score establishment and validation.

(**c**) Clear requirements about detailed documentation of the standard setting procedures
Simply comparing the length of Standard 8.6 (Edition 1985) with the length of Standard 6.5 (Edition 1999) makes apparent the change toward a stronger emphasis on proper reporting. There are at least two more standards in the 1999 edition (Standards 4.19 and 1.7) which accentuate on the need of detailed documentation.

(**d**) Encouragement for broader use of empirical data in standard setting
There are at least 3 standards in the 1999 edition (4.20, 14.7 and 4.17) which recommend broader use of empirical data in standard setting.

(**e**) Recognized need of proper training of judges
There is no standard in the 1985 edition which refers to the training of judges while in the 1999 edition there are two standards (4.21 and 1.7) concerning the judgmental process and the training of judges.

**Table 1: Quality standards for standard setting**

| **Standards for Educational and Psychological Testing** | |
|---|---|
| **Edition 1985** | **Edition 1999** |
| *Standard 8.6:* Results from certification tests should be reported promptly to all appropriate parties, including students, parents, and teachers. The report should contain a description of the test, what is measured, the conclusions and decisions that are based on the test results, the obtained score, information on how to interpret the reported score, and any cut score used for classification. | *Standard 6.5:* When statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations are available, the information should be included in the test's documentation. When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules, information about raw scores and derived scores, normative data, the standard errors of measurement, and a description of the procedures used to equate multiple forms |

| | |
|---|---|
| *Standard 6.9:* When a specific cut score is used to select, classify, or certify test takers, the method and the rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. | *Standard 4.4:* When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified in the same manner as is done for derived score scales. |
| | *Standard 4.9:* When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained. |
| | *Standard 4.19:* When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented. |
| | *Standard 4.20:* When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria. |
| | *Standard 14.7:* If tests are to be used to make job classification decisions (e.g., the pattern of predictor scores will be used to make differential job assignments), evidence that scores are linked to different levels or likelihoods of success among jobs or job groups is needed. |
| *Standard 10.9:* A clear explanation should be given of any technical basis for any cut score used to make personnel decisions. Cut scores should not be set solely on the basis of recommendations made in the test manual. | *Standard 4.21:* When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way. |
| | *Standard 1.7:* When a validation rests in part of the opinion or decisions of expert judges, observers or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The description of procedures should include any training and instruction provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth. |
| *Standard 2.10:* Standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score. | *Standard 2.14:* Conditional standard errors of measurement should be reported at several score levels if constancy cannot be assumed. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score. |

| | |
|---|---|
| *Standard 1.24:* If specific cut scores are recommended for decision making (for example, in differential diagnosis), the user's guide should caution that the rates of misclassification will vary depending on the percentage of individuals tested who actually belong in each category. | *Standard 2.15:* When a test or combination of measures is used to make categorical decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two applications of the procedure, using the same form or alternate forms of the instrument. |
| *Standard 5.11:* Organizations offering automated test interpretation should make available information on the rationale of the test and a summary of the evidence supporting the interpretations given. This information should include the validity of the cut scores or configural rules and a description of the samples from which they were derived. | *Standard 6.12:* Publishers and scoring services that offer computer-generated interpretations of test scores should provide a summary of the evidence supporting the interpretations given. |
| | *Standard 4.17.* Testing programs that attempt to maintain a common scale over time should conduct periodic checks on the stability of the scale on which scores are reported. |
| | *Standard 13.6:* Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experience. |

### 2.3. CHANGES IN METHODOLOGIES

The changes in the methodology were introduced for several reasons:

**Firstly**, in the mid 1980s it became evident that different standard setting methods produce different cut-off scores. Summarizing the results of 12 comparative studies Jaeger (1989, p. 500) analyzed 32 pairs of cut-off scores (in terms of a number of correct items) set by different methods and found that the ratio of the larger to the smaller of the cut-off score in every pair varies between 1 and 42 with an average of 5.30. In other words, in general, the cut-off scores (number of correct items) set by two different standard setting methods applied to the same test and meant to lead to comparable classification decisions might differ drastically.

The critical role of the choice of a specific standard setting method on the resulting cut-off score made Jaeger recommend – instead of one standard setting method in any study – to apply a combination of several standard setting methods and to establish the final cut-off score after considering all resulting cut-off scores as well as all additional information available.

This suggestion makes sense, but it does not provide an answer to the question: How is it possible for different methods to produce so different results if they were designed for one and the same purpose – to determine the cut-off point between two levels of proficiency? In fact, Glass (1978, p. 249) asked the same question, and regarded such discrepancy ("a startling finding") between the results of different methods as "… virtually damning the technical work from which it arose". In response to Glass, Hambleton (1978, p. 283) did not find anything 'startling', since if "… directions to judges were different, and the procedures

differed, no one should expect the results from these two methods to be similar". Unfortunately, while this response is reassuring, it does not resolve the main issue. When we do shopping we do not expect different shop assistants to use the same scale, but we expect the weight of the same five apples to be the same (or at least comparable) irrespective of the scale used. Is it then so much to expect that one and the same examinee will be assigned to the same level of proficiency irrespective of the standard setting method applied? Zieky (2001, p. 35) mentioned that "if the methods gave different results, people believed that one or possibly both of the results had to be wrong, and there was no way to tell which one is wrong". I would add to this that it is not a question of beliefs, but deductive reasoning (if two cut-off scores represent the same standard on the same test they should be the same or at least about the same) and "people" should not be blamed for being reasonable.

The controversy concerning the existing standard setting methods and their drawbacks were one of the main drives for the development of new methods, hoping to find the best one.

**Secondly**, performance assessment gains increasing popularity and can be characterized with "complex and polytomous (more than two score points per task) scoring rubrics (i. e., criteria used for assigning scores to examinee responses to each task), multidimensionality in response data (tasks requiring multiple skills for successful completion), interdependencies in the scoring rubrics (e. g., being unable to complete a task because one part of it was missed), and low score generalizability at the task or exercise level (performing well on one group of tasks does not mean a high performance on another)" (Hambleton et all, 2000, p. 356). Most of the well known old standard setting methods are not well suited for these specific characteristics of performance assessment and therefore new standard setting methods are needed to meet the new requirements.

**Thirdly**, broader use of IRT modeling for test analysis, item bank building and development of computerized adaptive tests naturally lead to the invention of new standard setting methods based on IRT modeling.

In summary, changes in methodology in the last 20 years are mainly in three basic directions:

- Increased number of newly developed compromise standard setting methods, which in setting cut-off scores combine human judgment with empirical data.
- Development of standard setting methods appropriate for constructed response items and performance tasks
- Intensified research in the field of computerized adaptive and web-based testing and apposite standard setting methods

### 2.4. CURRENT UNDERSTANDING AND COMMON AGREEMENT

- Acceptance of the role of values

There is a broad consensus that standard setting is a judgmental task, and a policy decision and as such it "… is arbitrary in the sense that it reflects a certain set of values and beliefs and not some other set of values and beliefs" (Kane, 1994, p. 434). There is also an agreement that the arbitrariness in the sense that they are based on judgment does not mean arbitrariness in the sense of capriciousness (Popham, 1978; Kane, 1994; Hansche, 1998; Impara & Plake, 2000; Zieky, 2001; Linn, 2003).

Capricious or not, the arbitrary nature of performance standards in terms of their dependence on values makes them vulnerable to objections and rebuttals. That is why providing sufficient evidence for the credibility and defensibility of the established performance standards and cut-off scores becomes an immanent and one of the most important parts of the standard setting process. In other words, standard setting nowadays is considered as a development of policy and that this policy "… should be legitimate in the sense that it is established by a specified authority in a reasonable way, and the consequence of implementing the policy should be positive" (Kane, 2001, p. 85).

- Different standard setting methods yield different cut-off scores

It took some time for the specialists to overcome the shock and disconcertment when they discovered that not only different standards tend to produce different cut-off scores, but also the same method, applied to the same test might result in different standards when it was applied with different groups of judges. There is a number of reasons which might explain the discrepancies, but such results challenge the theoretical foundations of standard settings and calls for re-conceptualization of the nature of standard setting.

- Loss of belief in a true cut-off score

In the earlier ages of standard setting development there was a hope that the 'true' standard exist and the only task of standard setting is to discover the right answer. Starting with Glass (1978) a number of leading professionals in the field (i.e. Jaeger, 1989; Cizek, 1993; Kane, 1994; Popham, 1997; Hansche, 1998; Reckase, 2000; Zieky, 2001; Linn, 2003) oppose this view. According to Zieky (2001, p. 45) nowdays "there is general agreement that cutscores are constructed, not found. That is, there is no 'true' cutscore that researchers could find if only they had unlimited funding and time and could run a theoretically perfect study" or in Kane's words: "There is no gold standard. There is not even a silver standard" (Kane, 1994, p. 448-449). And since "the tacit parameter estimation paradigm is, as has been argued, unsatisfactory, a dramatically different paradigm is needed" (Cizek, 1993, p. 99).

According to this alternative conceptualization, proposed by Cizek (1993, p. 100), which is a generalization of one of the procedural definitions of measurement, "…the foundation – like the function – of standard setting rests simply on the ability of standard setters to rationally derive, consistently apply, and explicitly describe procedures by which inherently judgmental decisions must be made". As can be seen, the emphasis in this re-conceptualization of standard setting is on the procedural aspects of standard setting as well as on the quality and legitimacy of standard setting procedures applied. That is why, by analogy with legal practice, Cizek (1993, p. 100) suggests standard setting to be considered *as a psychometric due process.*

According to the Random House Webster's College Dictionary *a due process of law* is "the regular administration of a system of laws, which must conform to fundamental and generally accepted legal principles and be applied without favor or prejudice to all citizens". In conformity with this definition if *a due process of law* has to be defined with one word, this word should be '*fairness'*.

Considering standard setting as a psychometric due process on one hand underlines the judgmental nature of standard setting and reflects, on the other hand, all major changes in the focus of standard setting, namely, increased concerns about:

- the details of standard setting procedures,
- the legal issues, and
- fairness.

In addition, the new conceptual framework of standard setting re-directs the research efforts from estimations of 'true standards' toward "refining and elaborating the systems of rules for deriving and applying judgment", and "improving the acceptability and defensibility of the endeavor" (Cizek, 1993, p. 103). The pragmatism and rationality of Cizek's re-conceptualization of the nature of standard setting turn it into the prevalent new paradigm of standard setting.

The term 'true cut-off score' is still used occasionally, but with a different meaning. For example, according to Reckase (2000, p. 50-51) "There is no such thing as a true standard, but there is a theoretical cut-score that would be set by a judge if he or she totally understood the process, the test, the content, and the policy and had a true score on the test in mind as the standard. The question is whether the standard-setting method can recover the theoretical cut-score assuming a judge performed every task consistently and without error". In fact, Reckase's interpretation of the meaning of the term 'theoretical cut-score' is consistent with Jaeger's view that "a right answer does not exist, except, perhaps, in the minds of those providing judgments" (Jaeger, 1989, p. 492).

The other areas of general agreement according to Linn (2003, p. 8) are, that:

- The role of the judges, involved in the standard setting procedure is crucial, and therefore they have to be well trained and knowledgeable, as well as to represent diverse perspectives. In other words, to represent different sets of values and beliefs.
- In the light of the procedural aspect of standard setting as a due process the well prepared documentation about all steps of standard setting process serves as procedural evidence and contributes to the credibility of the established performance standards.

### 2.5. MAJOR ISSUES IN STANDARD SETTING

Irrespective of the areas of common agreement delineated above, standard setting remains the most controversial topic in the field of educational measurement.

A number of issues still wait to be properly resolved and require further research. Some of these issues will be discussed in more detail later in this chapter, but most of them deal with:

- Some details of the judgment process and factors which affect it
- Procedures for cut-off score establishment and their impact on the resulting cut-off scores
- Validation of standard setting and performance standards
- Advantages and disadvantages of different standard setting methods and the choice of the most appropriate one in a given situation.

## 3. Standard Setting Methods

The first standard setting method, known as Nedelsky's method, was published in 1954 (Nedelsky, 1954). Thirty two years later in one of the most cited and comprehensive reviews on standard setting Berk (1986) listed 38 different standard setting methods, describing in more detail and evaluating 23 of them on the basis of 10 criteria of technical adequacy and practicability. More recently Reckase (2000) in search for possible standard-setting methods to be applied for setting performance standards on the National Assessment of Educational Progress (NAEP), reviews 14 newly developed methods applying 4 evaluation criteria: (1) minimal level of distortion in converting judgments to a standard, (2) moderate to low cognitive complexity of the tasks judges are asked to perform, (3) acceptable standard errors of estimate for the cut-scores, and (4) replicable process for conducting the standard setting study (Reckase, 2000, p. 50). Another review, published in the same year (Hambleton et al., 2000) appraises 10 standard setting methods applicable to complex performance assessment with polytomous scoring.

Up to date there are over 50 different standard setting methods and for many of them a number of different modifications exists.

### 3.1. CLASSIFICATION SCHEMES

In order to deal and summarize the increasing number of standard setting methods different schemes for classifications have been suggested. Berk (1986, p.139) suggests a 3-category classification scheme in which methods are classified '… according to whether they are based entirely on judgment (judgmental), primarily on judgment (judgmental-empirical), or primarily on test-data (empirical-judgmental). This classification scheme is seldom used at present since with the development of standard setting methodology most of the methods incorporate both judgments and empirical data.

The most commonly used classification scheme nowadays is the one suggested by Jaeger (1989, p. 493) who splits the standard setting methods into two large groups:

- test-centered continuum models, and
- examinee-centered continuum models.

The basis for this classification is the focus of the judgment task. According to this classification, test-centered methods are those methods in which judges have to make judgments about the examination tasks, while examinee-centered methods are those in which judgments concern real examinees and/or their work products. Sometimes the methods focused on the examinee performance are separated in another category called 'performance-centered' (Haertel & Lorié, 2000). Although this classification scheme is still the most prevalent one, some of the newly developed methods do not fit the two-category scheme and require a third, complementary category usually under the name 'other methods', which includes methods focused on score distribution, methods based on decision theory or some statistical techniques like cluster analysis.

The limitations of Jaeger's classification scheme have led to development of new classification schemes. For example, Reckase (2001, pp. 46-49) suggests 3 different classification continuums: (a) the size or complexity of the judgment task; (b) the amount and type of the supporting information and feedback provided to judges; (c) the complexity of the method applied for cut-off score establishment. Hambleton et al. (2000, pp. 356-357) on the other hand, offered a six-dimension classification scheme:

1. Focus of panelists' judgments (tasks, examinees, work products, scored performances)
2. Judgment task presented to the panel
3. The judgmental process
4. Composition and size of the panel
5. Validation of the resulting standards
6. The nature of the assessment

These new classification schemes, however, are still in limited use and that is why the most popular Jaeger's scheme will be applied in this chapter.

### 3.2. OVERVIEW OF STANDARD SETTING METHODS

Each one of the existing standard setting methods has its advantages as well as a number of limitations. Therefore the decision which of them to be applied in a concrete situation, should be made only on the basis of thorough analysis of the pros and cons of each of them in the light of the state of affairs. Since an in depth description of all available standard setting methods is rather impossible within the framework of this chapter the table in the Appendix provides only a list of the 34 most popular methods with their main characteristics as well as the sources where a detailed description of the methods can be found. Based on the information in the table one will be able to select the most appropriate methods under the circumstances and then find the basic sources for a detailed description of the selected method.

The table in the Appendix includes 13 columns and the brief explanation of the content of these columns is as follows:

**Column 1** (***No***) provides the ID numbers for the methods listed in the table.

**Column 2** (***Method***) presents the names of the methods.

**Column 3** (***Source***) lists the main sources where the method is described. The complete bibliographical description of the sources is given in the References.

**Column 4** (***Test format***) describes the format of examination for which the method is appropriate.

**Column 5** (***Focus***) specifies the focus of the judgment task. The methods in the table are sorted on the basis of their focus and within each of the categories in this column the methods are ordered in alphabetical order. Roughly speaking the first 21 methods can be classified as test-centered methods. Method 22 (Multistage Aggregation) is a complex method which belongs to both categories (test-

centered and examinee-centered methods). The next 7 methods (23 – 29) belong to the group of examinee-centered methods, and the last 4 methods (31 – 34) do not fit Jaeger's classification scheme and therefore fall into the third category: 'other methods'. Method 30 also has more than one focus (items and populations) and can be considered either as a test-centered method or as belonging to the third category – 'other methods'.

**Column 6** (***Outcome***) describes the main outcomes of the accomplishment of the judgment task. The outcomes vary depending on the task and it focus. These outcomes might be for example classification of items (examinees, profiles, cognitive domains), estimations of cut-off scores (probability for success, pass/fail rates), etc.

**Column 7** (***Feedback***) gives information about whether (**yes/no/?**) providing feedback to judges is considered as an essential part of the judgment process. The feedback can have different formats and can be provided on different stages of the judgment process. In this column feedback is considered as providing judges with information about their own rating behavior. The question mark (**?**), in this and the next columns, indicates that the main source of reference does not provide information on this point.

**Column 8** (***Data***) indicates whether (**yes/no/?**) the judges are provided with empirical data during the judgment process.

**Column 9** (***Rounds***) specifies the number of rounds in the judgment process. For different methods this number can vary between 1 and 4.

**Column 10** (***Decision making***) concretizes how the decisions were made (individually or on the basis of group consensus) and whether the revision of the first decisions is allowed.

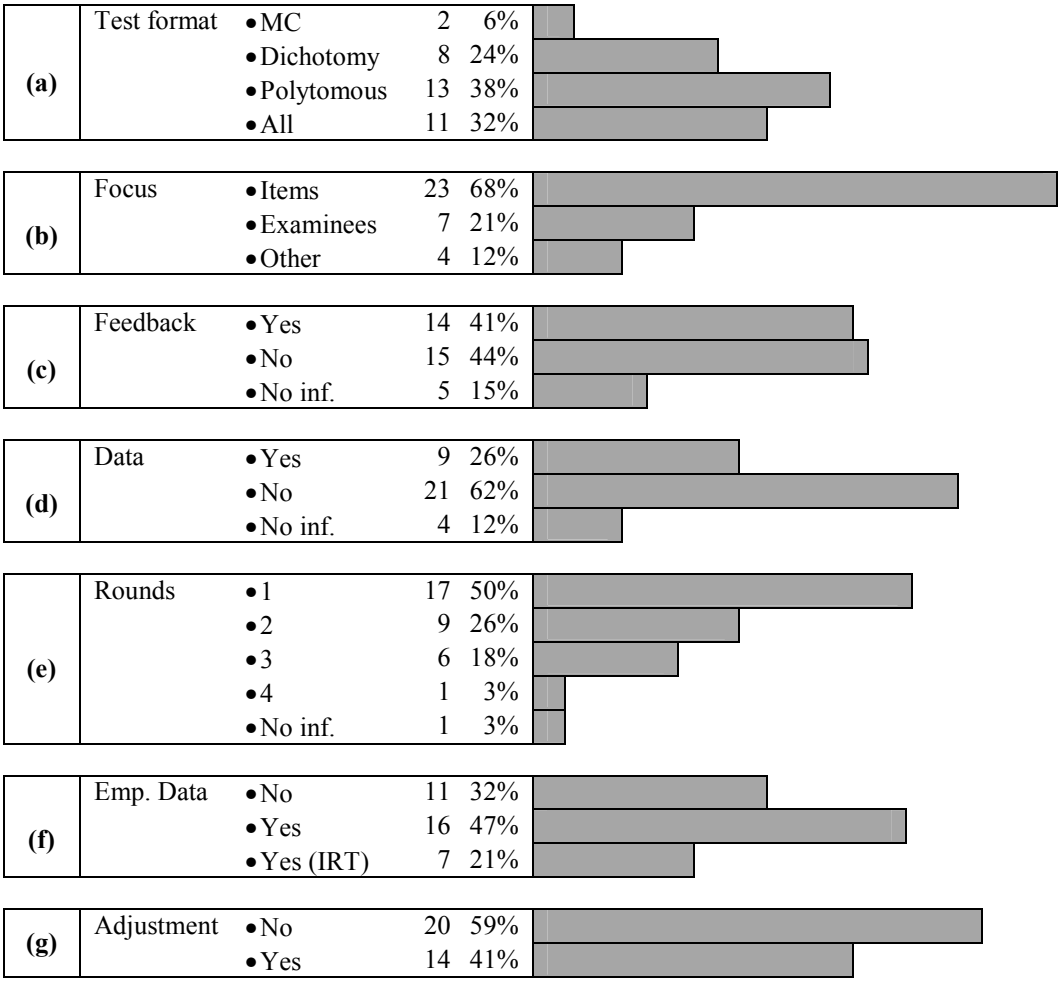
**Column 11** (***Decision rule***) briefly describes the decision rule applied for cut-off score establishment. It should be mentioned that in many cases different decision rules can be applied to the same set of judgments and most likely different approaches will yield different cut-off scores. The adequacy of the resulting cut-off score can be judged only on the basis of sufficient validity evidence.

**Column 12** (***Emp.data***) indicates whether (yes/no) empirical information is used on the stage of cut-off score establishment. The difference between this column and column 8 is the stage at which the empirical information is used. Column 8 indicates whether judges are provided with empirical data while column 12 indicates whether the empirical data is used on the stage of cut-off score establishment. Roughly speaking, the 'yes' in column 8 means that the method can be classified as judgmental-empirical in Berk's classification scheme, while the 'yes' in column 12 means that the method can be classified as empirical-judgmental. Some of the methods using empirical data on the stage of cut-off score establishment require Item Response Modeling to be applied to test items and sometimes also to judgments and if it is the case then the abbreviation (IRT) is added in Column 12.

**Column 13** (***Adjustment***) indicates whether (**yes/no/?**) some kind of adjustment between judgments and empirical data was applied in the stage of cut-off score establishment. The adjustment can take different forms and this will be discussed in more detail later in this chapter.

Fig. 2 summarizes the main characteristics of the methods listed in the table in the Appendix and in the next sections the main results will be briefly discussed.

**Fig. 2: Main Characteristics of the 34 Most Prominent Standard Setting Methods**




3.2.1. *Test Format*

The first chart in Fig.2 reflects one of the major changes in the standard setting methodology – the development of new methods suitable for performance assessment. While most of the old test-centered methods are appropriate mainly for multiple-choice dichotomously scored items, the majority of the methods (70%) presented in the Appendix are suitable either for all test formats or at least for polytomously scored items.

3.2.2. *Focus of the Judgment Task*

As far as it concerns the focus (Fig. 2b) of the judgment task most of the methods (68%) are **test-centered**. One of the main advantages of the methods in this group is that they allow the same objects (items) to be judged by a large number of judges which increases the reliability of the resulting cut-off scores. Another plus is that most of these methods can be applied *a priori* when there is no empirical data available yet. An additional important advantage in terms of practicality is that the implementation of test-centered methods as a whole is easier than the implementation of the other methods. If we sum up these three main advantages of test-centered methods it becomes clear why they are the most preferred standard setting methods.

On the other hand, all test-centered standard setting methods require judges to estimate item difficulty either by estimating the probability of correct answer for a certain target group of examinees or by classifying items into a number of proficiency levels. The ability of judges to estimate item difficulty

has been an object of a number of studies (Smith & Smith, 1988; Livingston, 1991; DeMauro & Powers,1993; Impara & Plake, 1998; Goodwin, 1999; Chang, 1999; Plake & Impara, 2001) and "…the most salient conclusion … is that the use of a judgmental standard setting procedures that requires judges to estimate proportion correct values, such as that proposed by Angoff (1971), may be questionable" (Impara & Plake, 1998). In the light of this important conclusion, the fact that the prevalent standard setting methods are test-centered and require judges to provide estimations of item difficulty makes questionable the validity of the established cut-off scores, based on these methods. There are a few possible approaches to deal with this issue:

- When a test-centered method is applied for standard setting, extensive appropriate training should be provided in order to improve the correlation between empirical and estimated item difficulty. The training should be accompanied by a validity check and some adjustment to empirical data should be made too. From this point of view test-centered methods which provide empirical data to judges (column 8) or incorporate them during the final stage of cut-off score establishment (column 12) or apply some kind of adjustment (column 13) are more preferable then the other test-centered standard setting methods.
- Taking into account the above mentioned potential flaw of test-centered methods it might be wise to use these methods in combination with methods from the other two groups, or following Jaeger (1989, p. 500) "… it might be prudent to use several methods in any given study and then consider all of the results, together with extrastatistical factors, when determining the final cutoff score".

As far as it concerns **examinee-centered** methods the main trend in recent development is narrowing the focus of the judgment task. In the examinee-centered methods like the border-group method (No 23) and the contrasting-groups method (No 24), developed in the era of awaking (1960-1980), the judgments about each examinee are based on the examinee's behavior during the whole instructional period while in the more recently developed methods (Body of work method – No 25, Generalized examinee-centered method – No 26, etc.) the judgments about each examinee are based only on his/her overall performance on the test under consideration. Narrowing the focus of the judgment task in such a way allows overcoming the main disadvantage of earlier examinee-centered methods – the limited number of judges able to provide estimation of the proficiency level of a given examinee.

The main advantage of all examinee-centered methods is that the judges are much more familiar with the task to assess examinees' performance than to assess item difficulty. The growing interest in examinee-centered methods in the last years can be explained with the fact that these methods are particularly appropriate for performance assessment in contrast to test-centered methods. That is why four out of the six examinee-centered methods presented in the Appendix were developed in the last 5-6 years together with a number of new modifications of the two well-known old methods – the border-group method (No 23) and the contrasting-groups method (No 24).

The limited number (4 or 5) of methods in the third category **(Other methods)** explains why this category still does not have a proper name. What all methods in this category (No 30 – No 34) have in common is that their focus is on the score distribution or score profiles. Most of them are applicable to all test formats and the cut-off score establishment based on both - empirical data as well as on judgments. In other words, the methods in the third categories might be described as empirical-judgmental in terms of Berk's classification scheme (Berk, 1986, p.139)

3.2.3. *Judgment Process*

The provision of **feedback** about rating behavior, **empirical data** about item difficulty and score distributions, as well as **group discussion** are considered among the most influential factors in standard setting (Fitzpatrick, 1989; Norchini, et al., 1988; Plake, et al., 1991; Maurer & Alexander, 1992; Hansche, 1998; Hambleton, et al. 2000; Buckendahl, 2000; Hambleton, 2001; Norcini, 2003).

There is also considerable evidence that the impact of these three components (feedback, normative data, and group discussion) strongly depends on their format and timing. Most of the authors support the idea that each of these components is important and should take place in the standard setting procedure, but there is also a common agreement that more research is needed in this area to ascertain which type and format of feedback and normative data are the most effective and what is the best time during the judgment process when this information should be given to the judges.

What is also needed is better documentation on the training and the judgment process as a whole. According to Reckase (2000, p. 46) "training seems to be an underappreciated part of the standard-setting process. Most reports of standard-setting procedures provide little detail about training". The summary results about the **feedback** (Fig. 2c) support to some extent Reckase's conclusion. According to these results feedback to judges is provided only in 41% of the methods. Taking into account that during the training stage some kind of feedback about rater behavior is usually provided irrespective of the standard setting method applied, the percentage mentioned above seems rather low. A possible explanation of this fact would be the lack of detailed information about the training stage, which coincides with the observation made by Reckase that in general the training process is not well documented and reported.

As far as it concerns the **normative data** (Fig. 2d), the fact that for most of the methods (62%) such data is not provided to the judges has a logical explanation. In most of the methods (68%) empirical data are used, but on later stage – during the process of cut-off score establishment (Fig.2 – f). There are at least three main reasons for this preference:

a. It is rather hard to monitor how and to what degree the judges use the empirical information they were provided with to adjust their rating. On the other hand, accommodating empirical data with the judgments on the stage of cut-off score establishment can be controlled and well documented.
b. In terms of practicality, it is easier to accommodate the empirical data on the last stage than to provide judges with it.
c. From the point of view of number of rounds, and consequently time required to provide judges with normative data, usually entails more than one round.

The last point (c) explains also why at least half of the methods require no more than one round (Fig. 2e) and only 21% of the methods require more than two rounds. Standard setting is a complex process with many participants involved and although it requires a lot of time, usually it is conducted under time pressure. That is why the KIS principle "Keep It Simple!" in terms at least of number of rounds plays an important role in the development and the application of standard setting methods.

3.2.4. *Cut-off Score Establishment*

The decision rules applied for establishing the cut-off scores are usually based on an aggregation function of the judgments. The choice of this aggregation function depends mainly on the focus of the judgment task and the characteristics of the responses to it. The analysis of the decision rules reveals also that although standard setting is considered as decision making there are still only a limited number of methods which are based on the decision theory approach (No 14, No 15, and No 30) while the nature of standard setting as such presupposes much broader usage of such methods. In fact, as Rudner (2001, p. 2) mentions only "isolated elements of decision theory have appeared sporadically in the measurement literature" and goes on suggesting that "… key articles in the mastery testing literature of the 1970s employed decision theory … and should be re-examined in light of today's measurement problems".

As far as it concerns the need of **empirical data**, the majority of methods (68%) require such data at least on the stage of cut-off score establishment. Besides, in almost one third (7 out of 23, see Fig. 2f) of the methods using the empirical data at that stage, **IRT** modeling is applied.

The IRT approach has many advantages: sample free estimation of item parameters; test-free estimation of person parameters; prior information about the standard error of measurement at each point of the ability scale. These advantages together with the availability of a variety of user-friendly software products designed for this kind of analysis makes IRT modeling a preferred approach to test development and analysis in all fields of educational measurement. For that reason it is not surprising that there is growing interest also in applying IRT modeling in standard setting. This approach, however, has its accompanying issues which have to be resolved before its broader application.

The main problem with all standard setting methods applying IRT modeling is that due to the probabilistic nature of IRT models they require an additional arbitrary decision to be made about so called 'item mastery level'. Item difficulty in most of the IRT models (at least one and two parameter models) is defined as that point of the proficiency scale where the chance of a person at this level to answer the item correctly is 50%. Although this definition of item difficulty is in harmony with item response theory, from the point of view of mastery testing many authors regard it as too low and suggest higher degrees of mastery to be considered. The satisfactory high probability of correct answer is usually called 'a mastery level', but nobody is able to say definitively what 'satisfactory high probability' means. That is for different methods and even for the same method, but in its different applications the mastery level varies in a very broad range – between 50% and 80%. Even within the same examination system, for example in the National Assessment of Educational Progress (NAEP) in the USA, the mastery level for items during the last 20 year has been changed from 80% in the early 1980s to 65% at the late 1980s, and then more recently went back to 50% giving up the 'mastery approach' and turning back to IRT model-based approach (Kolstad & Wiley, 2001).

Different standard methods deal in different ways with the problem of mastery level. For some of the methods the mastery level is defined *a priori* by the author. For instance, in the Bookmark method (No 17), it was set to be 66% (Reckase, 2000) and for the Item Domain method (No 20) the mastery level is predefined to be 80% (Schulz, et al., 1999). In some other standard setting methods, judges are those who have to define the item mastery level as is the case with the Combined Judgment-empirical method (No 19), but this approach also causes some additional, unexpected problems (Livingston, 1991). In the few applications of Item Mastery method (No 15) another approach was adopted – the mastery level was defined a posteriori on the basis of the analysis of the loss function and the efficiency of judges at different mastery levels (Kaftandjieva & Verhelst, 2000).

There are some other promising suggestions how to deal with the problem of item mastery level (Huynh, 1998; Haertel & Lorié, 2000; Kolstad & Wiley, 2001), but still a substantial amount of research will be needed before the problem will be properly resolved. And since "… arbitrary decisions often entail substantial risks of disruption and dislocation" before the problem is properly resolved it would be better to remember the warning Glass (1978, p. 258) gave 25 years ago: "Less arbitrariness is safer!"

Another limitation of the IRT approach is that getting a stable estimation of item and person parameters requires rather large samples of examinees as well as large item pools, which makes the approach inapplicable in case of small-scale examinations.

The basic flaw of many applications of IRT modeling in language testing especially is that there is not enough evidence provided about the model-data fit, which makes the findings of these studies more or less questionable. The model-data fit evidence (not only statistical) gains even more importance, when IRT modeling is applied in standard setting, because the established standards cannot be defensible if they were built on a doubtful basis.

As far as it concerns the **adjustment** between **judgments** and **empirical data** on the stage of cut-off score establishment, it is regrettable that the majority of standard setting methods (59%) do not apply it, because since "… there is no gold standard" (Kane, 1994, p. 448) the comparison between the empirical data and the judgments is the only reality check we have at our disposal.

Of course, the adjustment can be done in different ways and in different stages of the standard setting procedure. Cizek (1996, pp.16-17), for example, discuss three other forms of adjustment:

(a) adjustment to participants,
(b) adjustment to data provided by participants, and
(c) adjustment to the final standard (passing score).

According to Cizek (1996), an **adjustment to participants** means to give different weights to the judgments of different judges depending on their consistency with the empirical data or in the extreme case to eliminate the judges who deviate significantly from the established criteria.

There is no common agreement on this topic, mainly because the elimination of some of the judges is seen as 'politically incorrect', but at the same time a lot of indices of so called 'intra-judge consistency' have been suggested and applied in a number of studies (van der Linden, 1982; Kane, 1987; Maurer & Alexander, 1992; Taube, 1997; Chang, 1999). Going back to the issue of 'political incorrectness', the most important, from the psychometric point of view, is the validity of established cut-off scores. If the rating of some of the judges differs substantially from the empirical data this is an indicator of misunderstanding the judgment task and therefore the judgments of this judge cannot be trusted. If this is discovered during the training stage and the judge becomes aware of his/her deviance, he/she might adjust his/her rating behavior in an appropriate way. That is why providing feedback to the judges during the training is very important. If, however, the aberrant pattern was discovered only on the stage of cut-off score establishment the best way to deal with the problem is to assign different weights to the judges according to their intra-judge consistency. It may not be politically correct to the judges, but it is fair to the examinees and if we consider the standard setting as a due process we can refer to the possibility of ruling out some of the juror due to some of his/her personal characteristics which might lead to biased judgment.

An **adjustment to data provided by participants,** on the other hand, aims to reduce the variability among judges and is closely connected with inter-judge consistency. It can be done through appropriate training and/or guided group discussion. Reaching high inter-judge consistency will reduce the standard error, and increase the reliability of standard setting, but it should not be at the expense of taking into account that different parties involved in the judgment process might differ in their value systems and expectations.

If an **adjustment to the final standard** takes place, it is usually done after the establishing of the cut-off scores, and typically the decision for such an adjustment is made by another panel of judges, who weighs the proposed cut-off scores along with other considerations such as test reliability and standard error of measurement, classification error and passing rates (Mills & Melican, 1988). Two kinds of wrong decisions due to the error of measurement are possible when examinees are assigned to different levels of proficiency based on their test scores:

(a) to assign an examinee to a lower level, when he/she actually belongs to the higher level (*false negative error*), or

(b) to assign an examinee to a higher level, when he/she actually belongs to the lower level (*false positive error*).

More commonly the adjustment is done by lowering the final cut-off score by one, two or three standard errors in order to decrease false negative errors. The argument for such an adjustment is to give the examinee "the benefit of the doubt" (Cizek, 1996, p. 17). This procedure is very often applied and it is even recommended due to some legal considerations (Biddle, 1993). If such adjustment to the cut-off score is to be made, however, it should be taken into account that the decrease of one type of error automatically leads to the increase of the other type of error. Therefore, in case the adjustment is made, some additional evidence in support of this decision should be provided.

**In summary**, there is a large variety of standard setting methods and, as a rule, different methods usually yield different cut-off scores. To make the things even more complicated, it should be mentioned that the best standard setting method as such does not exist. Each of the methods has its own pros and cons and the choice of the method should depend mainly on:

- Test format
- Number of items
- Sample size
- Availability of normative data
- Stakes (high or low) of the examination
- Adverse impact of standard setting
- Perceptions and/or evidence about the validity of different standard setting methods
- Available resources in terms of time, staff, funding, equipment, degree of expertise, software available, etc.

And since there is no best method and different methods more often than not produce different cut-off scores, the best advice is to follow Jaeger's recommendation (Jaeger, 1989) to use several methods (2 or 3, if possible), preferably with different focuses of the judgment tasks and then, based on all results as well as the avalable other sources of information and external factors which have to be taken into account, to establish the final cut-off scores.

## 4. Validity Evidence

Standard setting is a complex endeavor, but to validate the standards is even more difficult (Kane, 2001, p. 54). That is why, although Chapter 6 in this Manual already covers to some extent the issue of empirical validation, some of the main aspects of building an interpretive argument with respect to standard setting validation are briefly discussed here as well.

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p. 9) *validity* refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests". In the context of standard setting, since there are no 'gold standards' and 'true cut-off scores', to validate established cut-off scores means to provide evidence in support of the plausibility and appropriateness of the proposed cut-off scores interpretations, their credibility and defensibility (Kane, et al., 1999).

As the cut-off scores are operational versions of performance standards, represented by points on the scale in which test results are presented, the validation of the cut-off scores cannot be done in isolation. The validity of interpretations of cut-off scores is confined within the validity of test scores as a whole and the validity of the applied performance standards. In other words, test validity and the validity of performance standards are necessary but not sufficient conditions for valid cut-off scores interpretations.

For example, as far as it concerns the CEF scales of language proficiency there is evidence of their validity as performance standards (North, 2002; Kaftandjieva & Takala, 2002). This fact, however, does not guarantee valid interpretations of the CEF scales in any particular case of their application. Therefore, the validation effort in every linkage between language examinations and the Common European Framework for Languages (CEF) should provide enough evidence not only for the plausibility of proposed cut-off scores interpretations, but also for the validity of CEF scale interpretations as well as for the validity of test score interpretations as a whole.

After highlighting the two main prerequisites for valid cut-off score interpretations (test validity and the validity of the performance standards adopted) let us focus on the validity issues concerning only

the standard setting. Two main types of validity evidence will be considered: procedural and generalizability evidence.

### 4.1. PROCEDURAL EVIDENCE

The main concern of procedural evidence is the suitability and the proper implementation of the chosen standard setting procedures with regard to the concrete circumstances. Although procedural evidence cannot guarantee the validity of cut-off scores interpretations, the lack of such evidence can affect negatively the credibility of the established cut-off scores.

Procedural evidence is important especially from the point of view of standard setting as a psychometric *due process*, since it reflects the procedural nature of the due process (Cizek, 1993, p. 100). On the other hand, standard setting is based on value judgments and therefore it is some kind of policy decision, and as such its credibility can be evaluated mainly on the basis of procedural evidence. In other words, "… we can have some confidence in standards if they have been set in a reasonable way …, by persons who are knowledgeable about the purpose for which the standards are being set, who understand the process they are using, who are considered unbiased, and so forth" (Kane, 1994, p. 437). In other words, "… the defensibility of standards is linked to the extent to which they can survive logical and judicial scrutiny and interpretation" (Cizek, 1993, p. 102).

The importance of procedural evidence becomes even greater if we take into consideration the fact that due to the nature of standard setting only a limited number of reality checks are available.

The role of careful documentation of the standard setting process is essential in providing sound procedural validity evidence and that is why one of the 20 criteria for evaluating standard setting research, suggested by Hambleton (2001, p. 113) is: "*Was the full standard-setting process documented (from the early discussions of the composition of the panel to the compilation of validity evidence to support the performance standards)? (... Attachments might include copies of the agenda, training materials, rating forms, evaluation forms, etc)*".

Two of the four recommended guidelines for standard setting provided by Cizek (1996, p. 14) also concern procedural evidence and proper documentation.

The provided procedural evidence should include (Kane, 1994; Cizek, 1996; Haertel & Lorié, 2000; Hambleton, 2001):

- Definition of the purpose of standard setting, and corresponding constructs
- Definition of performance standards applied
- A description of the standard setting method applied and the rationale for its choice
- Selection of the judges
- Training of judges
- Feedback from judges about their understanding of the purpose of standard setting, and judgment task as well as about their level of satisfaction with the process as such and with the final cut-off scores.
- Description of data collection procedures
- Description of procedures applied for cut-off score establishment
- Description of adjustment procedures, if such procedures were implemented.

## 4.2. GENERALIZABILITY EVIDENCE

Generalizability is one of the six aspects of Messick's unitary concept of construct validity (Messick, 1989). According to Messick (1995, p. 475) the generalizability aspect "… examines the extent to which score properties and interpretations generalize to and across population groups, settings and tasks, including validity generalizations of test criterion relationships" and focuses mainly on the consistency and replicability of the results.

Due to the subjective nature of standard setting, the consistency and replicability of the results do not guarantee the validity of the proposed cut-off score interpretations, but the lack of consistency can seriously jeopardize the cut-off score credibility. That is why "… the search for (a) *comparability* (i.e., convergence) between different methodologies and (b) *consistency* within methodologies" are defined by Cizek (1993, p. 96) as the implicit goals of any standard setting research and considered as means to verify that the arbitrariness of standard setting does not mean capricious standard setting (van der Linden, 1982, p. 295).

Most of the validity studies are focused on the generalizability across judges, examination tasks (Miller & Linn, 2000) and standard setting methods, but the other facets such as occasions or examinees deserve attention too, especially when examinee-centered standard setting methods are applied. And, as usual, the more sources are used for providing generalizability evidence, the more solid is the evidence and hence provides stronger support for the validity of the proposed cut-off scores interpretations. Some of these different sources of generalizability evidence will be discussed briefly in the following sections.

### 4.2.1. *Precision of cut-off score estimations*

The standard error of cut-off score estimations indicates how close to the established cut-off point would be a new cut-off point resulting from a replication of the standard setting, and according to Kane (1994, p. 445) this is one of internal validity checks.

A small standard error of cut-off score estimation is considered as one of the basic evaluation criteria for assessing the quality of a standard setting, but unfortunately, studies reporting the standard error of cut-off estimation are still rare according to Reckase (2000, p. 52).

Different approaches can be applied for the estimation of standard error – replicating the standard setting with different groups of judges or using different sets of items, or different samples of examinees, or applying different standard setting methods. The problem with all these approaches is that even conducting a single standard setting study is quite laborious and therefore the replications are very rare.

Another way to estimate the standard error is to apply generalizability theory (see Chapter 6 in the Manual, and Supplement E in this document for more details) to a single occasion estimating variance components for judges and items. Based on these estimates the standard error of measurement can be estimated too.

Hambleton (2001, p. 109) suggests even a simpler way – to split randomly the judges into two or more groups and to use the resulting cut-off scores from different groups as a basis for the estimation of the standard error. The formula which can be applied in this case is rather simple: $SE_C = \frac{SD_C}{\sqrt{n}}$, where $SE_C$ is the standard error of the mean cut-off point C, $SD_C$ is the standard deviation of the cut-off points, resulting from different groups of judges, and $n$ is the number of groups of judges.

When standard setting is based on independent judgments instead of dividing judges into two or more groups each judge can be considered as a group consisting of one element. For example, the following

table (Table 2) represents the cut-off points resulting from standard setting on the same test, but based on the independent judgments of 15 judges.

**Table 2: Cut-off scores based on 15 independent judgments**

| Judges | J1 | J2 | J3 | J4 | J5 | J6 | J7 | J8 | J9 | J10 | J11 | J12 | J13 | J14 | J15 | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cut-off | 96 | 80 | 96 | 95 | 94 | 96 | 84 | 89 | 81 | 89 | 82 | 89 | 89 | 89 | 86 | **89** | **5.6** |

Replacing in the above formula $\boldsymbol{SD_C}$ with **5.6** and $\boldsymbol{n}$ with **15** (the number of independent groups) the standard error of the **mean** cut-off point (**89**) will be equal to 1.44 ($SE_C = \frac{SD_C}{\sqrt{n}} = \frac{5.6}{\sqrt{15}} = \frac{5.6}{3.9} = 1.44$).

Whatever method for the estimation of the standard error is applied it should not be forgotten that in addition to the error of cut-off point estimation there is another source of error due to the measurement instrument (test). The standard error of the test can be used as a criterion for evaluating the magnitude of the standard error of cut-off score estimation. According to Cohen et al. (1999, p. 364) a standard error in the cut-off score that is less than one half of the standard error in the test (*SEM*) adds relatively little to the overall error and therefore would have little impact on the misclassification rates.

For the example above the ***SEM*** for that test is **8.7**, which means that the standard error of the cut-off score (**$\boldsymbol{SD_C}$ = 1.44**) is much less than one half of *SEM* (1.44/8.7 = 0.17) and therefore it can be considered as relatively small and acceptable.

It should be mentioned, however, that the above criterion is not absolute. In other words, if the standard error of the test is too large (the test has low reliability) then the fact that the $\boldsymbol{SE_C}$ is less than **½ *SEM*** does not provide very much support for the validity of the cut-off scores, since the total error of measurement will be too large for reliable ability estimation of the examinees and consequently for their reliable classification into different levels of proficiency.

It deserves to be mentioned that test reliability affects strongly the reliability of the classification decisions based on the established cut-off scores (Wright & Masters, 1982, pp. 105 – 106; Fisher, 1992; Wright, 1996; Schumacker, 2003). The so called **Index of Separation** ($I_{SEP} = \sqrt{\frac{Rel}{1-Rel}}$), which is based on the test reliability (*Rel*), can be used to estimate "… the number of statistically different performance strata that the test can identify in the sample" (Wright, 1996). The following table (Table 3) is based on this index and presents what should be the required level of test reliability in order to ensure a reliable separation into the desired number of proficiency levels.

**Table 3: Number of Proficiency Levels & Test Reliability**

| Number of Levels | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Number of Cut-off Points | 1 | 2 | 3 | 4 | 5 |
| Test Reliability | ≥ 0.61 | ≥ 0.80 | > 0.88 | > 0.92 | ≥ 0.95 |

The results in the above table demonstrate clearly the importance of test reliability for trustworthy classification decisions based on the proposed cut-off scores interpretations. That is why it is highly recommended that, instead of applying standard setting to an existing test, to specify in advance the number of proficiency levels and then to develop a test, matching as much as possible these levels, with more items whose difficulty is supposed to be at the same levels where the cut-off points are expected to be (Kane, 1994, p. 430). This approach is appropriate especially in the case of an existing Item Bank developed on the basis of IRT modelling.

Another good advice is, instead of using one long test in order to classify examinees in a larger number of proficiency levels (all 6 CEF levels, for example), to apply more than one shorter test, classifying

examinees in a more limited number of levels (2 or 3 preferably), applying a classification scheme like for example: *below B2, B2, above B2*. This approach can be considered as some kind of adaptive testing on test level and to ensure to some extent lower classification error.

And the last, but not the least important, advice is that there is a very simple way of increasing the precision of cut-off score estimates simply by increasing the number of judges and/or items and/or occasions used in the standard setting (Kane, 1994, p. 439). One of the most often put questions concerning standard setting is: *How many judges are enough?* Unfortunately, this question does not have a simple answer. Livingston & Zieky (1982) suggest the number of judges to be not less than 5. Maurer, et al. (1991) found that at least 9 to 11 judges are needed to produce adequately reliable rating at least when the Angoff standard setting is applied. Based on the court cases in the USA, Biddle (1993) recommends from 7 to 10 Subject Matter Experts to be used in the Judgement Session. As a general rule Hurtz & Hertz (1999, p. 896) recommend 10 to 15 raters to be sampled, preferably representing "… as many constituent groups as possible, including individuals who practice and hold expertise in different specializations within their professions". Although the Hurtz & Hertz (1999) advice concerns only the application of Angoff standard setting method, bearing in mind that most of the test-centered standard setting methods can be considered as modifications of Angoff method at least in terms of the format, focus and the outcomes of the judgment task, this general rule can be extended.

Another advice concerning the number of judges is given by Jaeger (1991, p. 10) who recommends the size of sample of judges to be such that the standard error of the mean of the cut-off points suggested by individual judges ($SE_C$) "… is small, compared to the standard error of measurement of the test for which a standard is sought".

4.2.2. *Inter-judge consistency*

Inter-judge consistency is another kind of internal validity check, which is closely related to the precision of the cut-score estimations and again it should be mentioned that high level of inter-judge consistency does not guarantee, but only support, the validity of cut-off score interpretations.

Inter-judge consistency refers to the degree of uniformity of judgments of different experts on the same objects (level descriptors, items, examinees or examinees' performances). There are many different factors which can affect the inter-judge consistency and although many studies were focused on this topic, still a lot of work has to be done. Irrespective of the factors having impact on the inter-judge consistency, there are three main sources of inconsistency:

- the inconsistency due to a different conception of mastery;
- the inconsistency due to different interpretations of performance standards (levels of language proficiency);
- the inconsistency due to different value systems.

That is why the first two stages of the Standardisation Process – Familiarisation and Training (see Chapter 5 in this Manual) – are of great importance, since their main goal is to reduce the inconsistency due the different interpretations of performance standards and different conceptions of mastery.

There are different ways of analysis of inter-judge consistency. Analysis of the correlation between ratings or calculating Cronbach $\alpha$ are among the most often applied methods although in the framework of standard setting they are hardly the most appropriate, since it is possible to have a perfect correlation of +1.00 between two judges with zero-agreement between them about the levels to which descriptors, items, examinees or their performances belong, as can be seen in the following hypothetical example (Table 4) – three judges rate 7 objects on a 6-point scale and although the correlation between Rater 1

and Rater 2 is equal to +1.00, the percentage of agreement between them is equal to 0% due to the fact that they use different ranges of the scale.

**Table 4: Relation between Correlation and Agreement**

| | Objects | | | | | | | Correlation / Agreement | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **Rater 1** | **Rater 2** | **Rater 3** |
| **Rater 1** | 5 | 6 | 4 | 4 | 5 | 5 | 6 | X | **+1.00** | +0.82 |
| **Rater 2** | 2 | 3 | 1 | 1 | 2 | 2 | 3 | **0%** | X | +0.82 |
| **Rater 3** | 6 | 6 | 4 | 4 | 4 | 5 | 6 | 71% | 0% | X |

A simple, but still quite appropriate, index for inter-judge consistency is the **percentage of exact agreement** between each two raters, or the average agreement with the corresponding range (min/max). The main disadvantage of this index is that it does not take into account the possibility of agreement by chance. For example in case of pass/fail decisions two raters can reach 50% agreement even if they guess randomly, while if the 6-point CEF scales are used the agreement by chance will be only 17%. That is why the interpretations of the percentage of exact agreement should always take into account the number of rating categories. The lower the number of these categories is the higher will be the percentage of chance agreement.

In contrast to the percentage of exact agreement **Cohen's coefficient** $\kappa$ takes into account the probability of agreement by chance. Kappa ($\kappa$) is based on the absolute percentage of agreement and might be interpreted as a percentage of agreement corrected for chance agreement and that is why it is lower then the percentage of exact agreement (except in the case of 100% agreement, when $\kappa = 1$).

**Table 5: Inter-judge Consistency**

| JudgeA / JudgeA$_1$ | A1 | A2 | B1 | B2 | C1 | C2 | TOTAL |
|---|---|---|---|---|---|---|---|
| **A1** | **3** | 1 | 0 | 0 | 0 | 0 | **4** |
| **A2** | 0 | **3** | 1 | 0 | 0 | 0 | **4** |
| **B1** | 0 | 0 | **2** | 1 | 1 | 0 | **4** |
| **B2** | 0 | 1 | 0 | **2** | 0 | 0 | **3** |
| **C1** | 0 | 0 | 0 | 1 | **2** | 0 | **3** |
| **C2** | 0 | 0 | 0 | 0 | 0 | **2** | **2** |
| **TOTAL** | **3** | **5** | **3** | **4** | **3** | **2** | **20** |
| % of exact agreement = **70**% | | | | | | | |
| Cohen's $\kappa$ = **0.637** ($p$ = .000) | | | | | | | |

| JudgeB / JudgeB$_1$ | PASS | FAIL | TOTAL |
|---|---|---|---|
| **PASS** | **5** | 4 | **9** |
| **FAIL** | 2 | **9** | **11** |
| **TOTAL** | **7** | **13** | **20** |
| % of exact agreement = **70**% | | | |
| Cohen's $\kappa$ = **0.381** ($p$ = .081) | | | |

Since the chance agreement depends on the number of categories it is possible for the same percent of exact agreement to correspond to different kappa's values as it is demonstrated in Table 5. This table summarizes the results of inter-judge consistency analysis in two cases when different scales with different number of categories (six and two). As can be seen from the table in both cases the two judges agreed in 14 out of 20

cases which means that the percentage of exact agreement is the same: 70% ($=\frac{14}{20}*100$). Cohen's kappa however is much higher in the first case than in the second. Even more, in the first case $\kappa$ differs significantly from the chance agreement ($p < .05$) while in the second case $\kappa$ indicates that the agreement between the two judges might be due to chance only ($p > .05$).

The example provided in Table 5 demonstrates that the same percentage of exact agreement might be interpreted in different ways (as high or low) depending on circumstances.A large number of other, more sophisticated methods for the analysis of inter-judge consistency exist, some of them, like intra-class correlation, based on the analysis of variance, others based on latent-variable modeling approach (Abedi & Baker, 1995) or IRT modeling (Engelhard & Stone, 1998). They all have advantages and limitations, but their main shortcoming is that in comparison with the simpler indexes, like the percentage of agreement, they require more time and expertise. If providing feedback to the judges is an essential part of the judgment process then the time factor becomes very important and the percent of agreement should be preferred.

4.2.3. *Intra-judge consistency*

The term '*intra-judge consistency*' might be interpreted in two different ways. The first possible interpretation is in terms of replicability (stability) of the ratings of a single judge over time periods and occasions. In other words, the degree to which a judge tends to make the same judgments about the same objects on different occasions. Although the degree of intra-judge consistency can be used as supporting validity evidence (another kind of internal validity check), especially to support the claim that irrespective of its arbitrariness standard setting is not capricious, the analysis of this kind of intra-judge consistency is very rarely conducted in the field of standard setting.

In 1982 van der Linden (1982) gave another interpretation of this term and suggested a latent trait method for its analysis. According to his definition, "intrajudge consistency arises when judges specify probability of success on the items which are incompatible with each other and, consequently, imply different standards" (van der Linden, 1892, p. 296). Since then this phenomenon (intra-judge consistency) has been extensively analyzed. The main reason for this constant interest is that the test-centered methods are still the prevalent standard setting methods, and almost all of them, in one way or another, require judges to make estimations of item difficulty. That is why the analysis of intra-judge consistency as almost the only 'reality check' of the established cut-off scores becomes one of the main sources for providing validity evidence at least for the test-centered standard setting methods.

The results of the analysis of intra-judge consistency and the effect of different factors on it lead to a better understanding of the judgment process. As a result, a number of new standard setting methods and/or different modifications of the existing standard setting methods were developed and implemented in order to decrease the degree of intra-judge inconsistency.

When the judgment task requires judges to estimate the probability of a correct answer for every item, then one of the most often used index of intra-judge consistency is the correlation between judgments and the empirical item difficulty. Two other indices suggested by Maurer, et al., (1991) and Chang (1999) are also appropriate when the judgment task is to estimate the probability of a correct answer.

When the outcomes of the judgment task are dichotomous or polytomous classifications of items then the above mentioned indices of intra-judge consistency are not very appropriate. In this case, some kind of scaling (calibration) of judgments should be applied first and then the correlation between these calibrations and item difficulty might be computed and used as an index of intra-judge consistency.

IRT modeling is one of the most promising approaches to the analysis of intra-judge consistency (van der Linden, 1982; Kane, 1987; Taube, 1997; Engelhard & Stone, 1998; Kaftandjieva & Takala, 2000),

but it has its own limitations too. The major limitation is that there is no guarantee that the data (either from test administrations or from judges) will fit the chosen IRT model. An additional limitation is that with a small number of items (judges) the stability of estimations will be questionable.

4.2.4. *Decision consistency and accuracy*

The aim of any standard setting procedure is to establish cut-off scores on the basis of which examinees are classified in a limited number of proficiency levels. *Decision consistency* refers to the agreement between the classifications of the same examinees on two different examinations with the same test (or with parallel forms of the test). Two statistics can be used as indices of decision consistency – the percentage of agreement between the two classifications and Cohen's $\kappa$. The main problem with establishing the decision consistency, however, is not in the computing of the indices, but in the fact that the above-mentioned indices both require two administrations of the test to the same examinees, which in practice is rather hard to implement. To overcome this problem a few methods for determining decision consistency, based on a single administration, were developed. Some of them can be applied only to tests with dichotomous-scored items (Huynh method, Subkoviak method, Marshal-Haertel method – Subkoviak, 1984), while a more recent one, developed by Livingston and Lewis (1995) and gaining more and more popularity can be applied to "… any test score for which a reliability coefficient can be estimated" (Livingston & Lewis, 1995, p. 179). Another advantage of the Livingstone and Lewis method is that it allows on the basis of a single administration to estimate decision consistency as well as decision accuracy. According to Livingston and Lewis (1995, p. 180), *decision accuracy* refers to "… the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known". The only drawback of this method is its technical sophistication (Hambleton & Slater, 1997), which might limit its application.

There are different factors which might influence the degree of decision accuracy. Based on a simulation study, Ercikan and Julian (2002) found that the degree of decision accuracy decreases when the number of proficiency levels increases. It confirms the already made recommendation to classify examinees on the basis of a single examination in a limited number of proficiency levels (2 or 3 preferably). The same study provides additional evidence that the decision accuracy depends strongly on test reliability, but the impact of the error of measurement (*SEM*) at the cut-off points is even stronger. According to their findings (Ercikan & Julian, 2002, pp. 290-291) to classify accurately at least 80% of the examinees in more than 3 proficiency levels, the reliability of the test should be not lower than 0.95. If the test reliability is below 0,95 the same level of accuracy (80%) can be obtained only if the number of classification categories (proficiency levels) is less than four.

As far as it concerns decision consistency, if two standard setting methods were applied, then the consistency of the decisions based on the two sets of established cut-off scores could be analyzed. This kind of analysis can be viewed as an 'external validity check' and a high degree of agreement would provide a strong validity evidence for the plausibility of the proposed cut-off scores.

Instead of applying another standard setting method, another external criterion (teacher's rating, self-assessment, another test, etc.) can be used to classify the same examinees and then to analyze the decision consistency of the two classifications. In line with Messick's unified view of validity (Messick, 1989, 1995) it can be considered not only as a generalizability evidence, but also as a kind of evidential validity evidence.

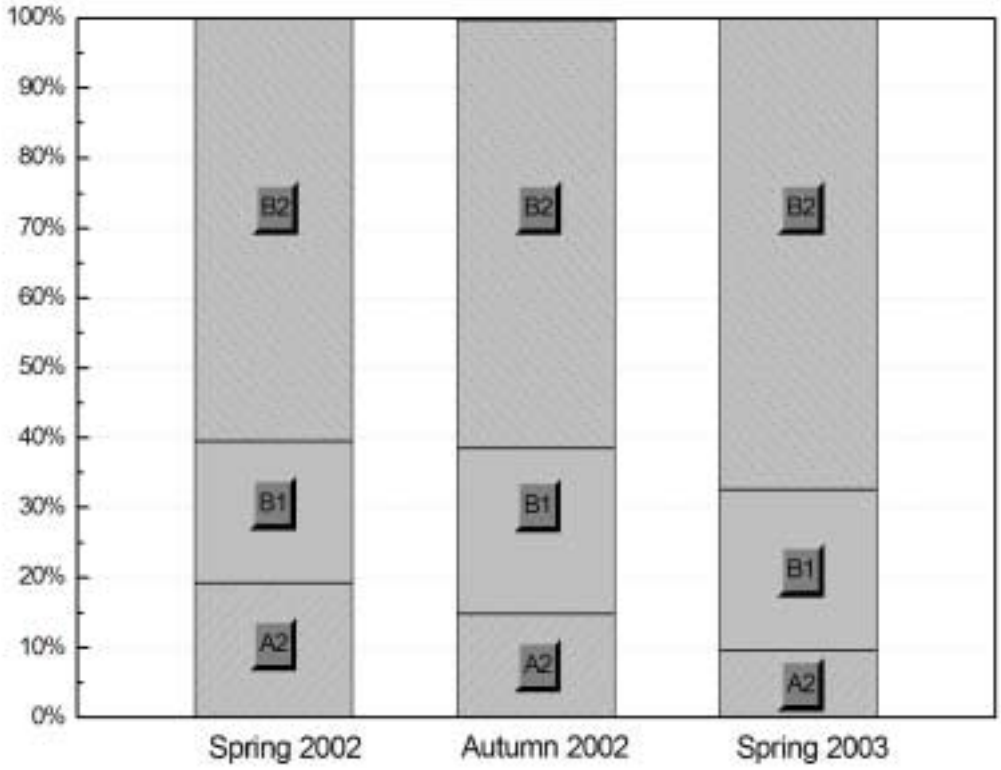
4.2.5. *Pass rate*

The analysis of the pass rate or the percentage of examinees assigned to each level is another way to support the validity of proposed cut-off score interpretations. It is especially valuable when the fairness of cut-off scores interpretations has to be demonstrated. The stability of the pass rate over years, examinations or samples drawn from the same population is a strong support for the consequential validity of cut-off

interpretations. And since "the chief determiner of performance standards is not truth; it is consequences" (Popham, 1997), the analysis of pass rates has great importance.

Fig. 3 gives an example of such kind of analysis. The graph presents the results of three consecutive test administrations of the Finnish National Language Certificate Tests (YKI) for English – Intermediate level (B1-B2), Reading comprehension. Different test versions were used for each administration, but the items included in these three different tests belong to an Item Bank built on the basis of IRT modeling, and hence the results of all tests are presented in the same scale and cut-off scores were established once (when the Item Bank was built) and applied for classification decisions in all subsequent test administrations.

**Fig. 3. Pass rate: English – Intermediate – Reading**




The number of examinees per session varied between 483 and 626, but as can be seen from Fig. 3 the pass rate over the sessions with different examinees and different tests is quite stable with a tendency of decrease of the percentage of examinees below B1 and increase of the percentage of candidates on level B2 and above.

The analysis of the pass rates can be used also as an external validity check if the pass rate, based on the newly established cut-off scores, is compared with the pass rates based on the implementation of another test. The comparability of the two pass rates will support the credibility of the newly established cut-off scores. On the other hand, if there is a big discrepancy between the pass rates from two different tests the only logical conclusion is that the interpretations of test scores of at least one of the two tests are inappropriate. Unfortunately, it is impossible to infer only from the inconsistency of the pass rates which one of the two test score interpretations is the more credible one.

## 5. Main steps in the standard setting process and some basic recommendations

### 5.1. SELECTION OF METHOD

It was already mentioned that many factors should be considered when the decision about which standard setting method to apply has to be made. Since there are more than 30 different standard setting methods, the choice of the method for the concrete situation should be based on a thorough review of the existing standard setting methods and their pros and cons in the light of the concrete testing situations. Different authors suggest different selection criteria (Cizek, 1996; Reckase, 2000; Hambleton, 2001), but the most important criteria are:

**(a)** The appropriateness of the method for the concrete situation;

**(b)** The feasibility of the method implementation under the current circumstances;

**(c)** The existing validity evidence for the quality of the selected method.

Of course, the last criterion does not guarantee automatically the validity of the cut-off score interpretations in every new implementation of the selected method, but the credibility of the established cut-off scores would increase if there is enough prior evidence of the quality of the method. That is why, if for one reason or another, a less widespread standard setting method is preferred, then a detailed methodological description of the method should be provided together with sound and compelling arguments for its development and implementation as well as strong enough validity evidence for its quality (Cizek, 1996).

Another issue to be considered when the standard method is selected is its complexity. Rightly or not, "… standard-setting methods that require effort are likely to be viewed as more credible than those that do not" (Norcini & Shea, 1997, p. 44), but although this should be taken into account it cannot be the main selection criterion, not only because "the intent is to demonstrate due diligence, not endurance" (Norcini & Shea, 1997, p. 44), but also, because of merely practical limitations, which in the most real world situations are of great importance.

### 5.2. SELECTION OF JUDGES

Since standard setting is a judgment process the role of judges in it is well recognized by virtually everybody who works in the field of standard setting. A number of recommendations have been made (Jaeger, 1991; Maurer & Alexander, 1992; Berk, 1996; Cizek, 1996; Norcini & Shea, 1997; Reckase, 2000; Hambleton, 2001; Raymond & Reid, 2001), sometimes contradicting each other. For example, according to Raymond & Reid (2001, p. 130) "… participants for standard setting panels should: (a) be subject matter experts; (b) have knowledge of the range of individual differences in the examinee population and be able to conceptualize varying levels of proficiency; (c) be able to estimate item difficulty; (d) have knowledge of instruction to which examinees are exposed; (e) appreciate the consequences of the standards; (f) collectively represent all relevant stakeholders.

It seems rather hard to fulfill all these requirements for all judges involved. It concerns especially requirements (a) and (f), because if we involved representatives of diverse groups like parents, administrators, managers, etc. more probably they will not be subject matter experts and will not possess many of the other characteristics, either.

On the other hand, the last requirement is important since, if it is taken into consideration, it definitely will increase the credibility of the established cut-off scores. That is why the recommendation given by Berk (1996, p. 222) makes a lot of sense. He recommends, instead of choosing two samples of judges, to choose one sample, representing, as well as possible, all relevant stakeholders and another sample, consisting of subject matter experts fulfilling as much as possible the requirements (b), (c) and (d). Only the second sample will be involved in the standard setting procedure, making judgments about items

(examinees or performances) while the first sample might be involved in the beginning and the end of standard setting process. In the beginning, to provide information about the expectations of the representatives of different groups about the possible consequences of standard setting, and at the end, to get feedback about the plausibility of the established cut-off scores and discuss and possibly apply some cut-off score adjustment.

Taking into account how important and at the same time how difficult it is to select the most appropriate judges Jaeger (1991, p.4-5) suggests the identification of judges with sufficient expertise to be done through post hoc analysis of judges' recommendations. In fact, what he suggests indirectly is to disqualify judges with high degree of intra-judge inconsistency or at least to apply different weights to the judgments of different judges. And although there are some arguments against this idea, it deserves at least to be considered.

As far as it concerns the number of judges, the general advice would be: as many as possible, but not less than 10 for the second group of judges, who will participate in the actual judgment process. As far as it concerns the first group of judges, representing different groups of stakeholders – the more diversity it represents the better.

### 5.3. TRAINING

Irrespective of the selected standard method, the crucial part in every standard setting procedure is the training of judges. At the same time, in practice, the training process is usually underestimated and poorly documented (Reckase, 2000; Raymond & Reid, 2001).

In the standard setting literature the stage of familiarization as it is presented in chapter 5 of this Manual is usually considered as an initial step in the training process and therefore the aim of the training process as a whole is threefold:

(a) to ensure a unified interpretation of proficiency levels by all judges;

(b) to guarantee that every judge understands completely the judgment task

(c) to get information about rating behavior and the degree of competence of every rater.

Raymond and Reid (2001, p. 148) mentioned three major criteria for effective training: (1) stability over occasions; (2) consistency with assumptions underlying the standard-setting method; and (3) reflective of realistic expectations.

There are a few important things which should be taken into account when the training is planned, organized and conducted:

1. Plan and give opportunity to judges **to take the test** under standard or near standard conditions.
2. Provide judges with the **scoring key** or the detailed scoring scheme for every test item.
3. Design easy to use **rating forms.**
4. Provide judges with as much as possible **feedback** about their rating behavior, and the degree of their inter- and intra-judge consistency.
5. Provide judges with **empirical data**. (If the judgment process is taking place before the examination, use old empirical data).
6. Give the judges an opportunity **to discuss** their ratings.
7. Continue the training until the satisfactory level of inter- and intra-judge consistency has been reached.

8. Get **feedback from judges** about their satisfaction with the training process and their confidence in their ability to complete the judgment task. (A good example of such an evaluation form is provided by Hambleton (2001, pp. 105-108)[1].
9. Do not forget to document well the entire training process.

### 5.4. JUDGMENT PROCESS

In contrast with the training process, there are no specific recommendations except probably one – follow as strictly as possible the prescribed procedures and document the process. If due to the circumstances some modifications have to be made – provide the rationale. And again as with the training – ask judges to fill in an evaluation form about the judgment process, the standard setting method applied and about their satisfaction with the resulting cut-off scores.

### 5.5. CUT-OFF SCORE ESTABLISHMENT

Irrespective of the quality of the method chosen, the choice of judges and the quality of the training, and how proper the implementation of the standard method is, it still might happen that the resulting cut-off scores are not very plausible.

Instead of defending them at any price, the wiser policy is to collect as much additional information as possible from different sources – past examinations, the expectations of different groups of stakeholders, the feedback from judges, and of course, whenever possible to apply an additional standard setting method. Taking into account all this information, adjust the already established cut-off scores in a way which will increase their plausibility and credibility.

This recommendation is in the line with Popham's view (Popham, 1997, p. 110) on standard setting as "fundamentally a consider-the-consequences enterprise".

Someone might say that standard setting is complicated enough even without the last recommendation to collect additional information, including the implementation of another standard setting procedure, and he or she will be right. On the other hand, nobody has ever claimed that standard setting is 'a piece of cake'. To set the passing scores is a great responsibility and everybody involved in this business should be aware of it.

A Bulgarian proverb says "Measure seven times before making a cut!" When the decisions based on the established cut-off scores will affect in one way or another a number of examinees, then collecting information from as many sources as possible does not seem such a burden, bearing in mind the consequences.

### 5.6. VALIDATION AND DOCUMENTATION

Providing strong validity evidence and documenting all steps in the standard setting endeavor might look as an additional burden, especially if this is considered only as a means to convince the other interested parties of the plausibility and credibility of the proposed cut-off scores. If, however, we look at it as a way to decrease our own uncertainty about the credibility of the established cut-off points and in this way to reduce the burden of the huge responsibility in taking decisions about the other human beings, then validation and documentation make a lot of sense and deserve the effort.

[1] The same form can be found in Hansche (1998, pp.107-111), which is available online.

**Conclusion**

There is a long list of references in this chapter and it is a sign of the amount of work done in the field of standard setting. My favorite book 'The Little Prince', however, is not in that list. But one of the characters in that book, the fox, used to say something which can be applied to everything concerning standard setting and it is: '*Nothing is perfect*!'

To summarize – there is no 'gold standard', there is no 'true' cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence. In three words – nothing is perfect. Cicero says that '*There are many degrees of excellence*', but when making decisions concerning the other human beings I would prefer the other saying made by Lucan: '*Don't consider that anything has been done if anything is left to be done*'. Whether it sounds pessimistic or optimistic depends on the point of view, but it is the same with all value judgments, including standard setting.

## REFERENCES


**Abedi, J. & Baker, E.** (1995). A Latent-Variable Modeling Approach to Assessing Interrater Reliability, Topic Generalizablity, and Validity of Content Assessment Scoring Rubrics. *Educational & Psychological Measurement*, 55, (5), 701-716.

**American Educational Research Association, American Psychological Association, and National Council on Measurement in Education**. (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

**American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.** (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

**Angoff, W. H.** (1971) Scales, norms and equivalent scores. In: *Educational Measurement.* Ed. by R. L. Thorndike, (Second Edition), Washington, D.C.: American Council on Education, 508-600.

**Berk, R.** (1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced tests. *Review of Educational Research*, 56, (1), 137-172.

**Berk, R.** (1996). Standard Setting: The next generation (Where few Psychometricians Have Gone Before!) Applied Measurement in Education, 9, (3), 215-235.

**Biddle, R.** (1993). How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing., *Public Personnel Management*, 22, (1), 63-70.

**Brandon, P.** (2002). Two versions of Contrasting-Groups Standard-Setting Method: A Review. *Measurement and Evaluation in Counseling and Development*, 35, 167-181.

**Buckendahl, C., Impara, J., Giraud, G., Irwin, P.** (2000). *The Consequences of Judges Making Advanced Estimates of Impact On a Cut Score.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

**Carson, J. D.** (2001). Legal Issues in Standard Setting for Licensure and Certification. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, andperspectives.* Hillsdale NJ: Erlbaum, 427-444.

**Cascio, W., Alexander, R., & Barret, G.** (1988). Setting Cutoff Scores: legal, Psychometric, and Professional Issues and Guidelines. *Personnel Psychology*, 41, 1-24.

**Case, S. & Swanson, D.** (1998). *Constructing Written Test Questions for the Basic and Clinical Sciences.* Philadelphia: National Board of Medical Examiners.

**Chang, L.** (1999). Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting methods. *Applied Measurement in Education,* 12 (2): 151–165.

**Cizek, Gr. J. (**1993**).** Reconsidering Standards and Criteria. *Journal of Educational measurement*, 30, (2), 93-106.

**Cizek, Gr. J. (**1996**).** Standard Setting Guidelines. *Educational Measurement: issues and Practice,* 15, 13-21.

**Cizek, Gr. J. (**2001**).** Conjectures on the Rise and Call of Standard Setting: An Introduction to Context and Practice. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum, 3-18.

**Clauser, B. & Nungester, R.** (1997). Setting Standards on Performance assessment of Physicians' Clinical Skills Using Contrasting Groups and receiver Operating Characteristic Curves. *Evaluation & the Health Professions*, 20, (2): 215-238.

**Clauser, B., Subhiyah, R., et al.** (1995). Scoring Performance Assessment by Modeling the Judgment of Experts. *Journal of Educational Measurement*, 32, (4), 397-415.

**Cohen, A., Kane, M. and Crooks, T.** (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education,* 14: 343–366.

**CRESST Assessment Glossary.** (1999). Retrieved December 12, 2003 from CRESST – National Center for Research on Evaluation, Standards, and Student Testing Web site: http://www.cse.ucla.edu/CRESST/pages/glossary.htm

**DeMauro, G. & Powers, D.** (1993). *Logical Consistency of the Angoff Method of Standard setting.* RR-93-26, Princeton, Educational testing Service.

**Dylan, W.** (1996). Meaning and Consequences in Standard Setting. *Assessment in Education: Principles, Policy & Practice*, 3, (3), 287-308.

**Engelhard, G. & Stone, Gr.** (1998). Evaluating the Quality of Ratings, Obtained from Standard Setting Judges. *Educational & Psychological Measurement*, 58, (2), 179-196.

**Ercikan, K. & Julian, M.** (2002). Classification Accuracy of Assigning Student Performance to Proficiency Levels: Guidelines for Assessment Design. *Applied Measurement in Education,* 15, (3), 269-294.

**Fisher, W. Jr.** (1992). Reliability Statistics. *Rasch Measurement Transaction*, 6:3, p.238, Retrieved December 8, 1999 from: http://209.41.24.153/rmt/rmt63.htm

**Fitzpatrick, A.** (1989). Social Influences in Standard Setting: The Effects of Social Interaction on Group Judgment. *Review of Educational Research*, 59, (3), 315-328.

**Glass, G. V.** (1978). Standards and criteria. *Journal of Educational Measurement,* 15, (4), 237–261. Retrieved October 12, 1999 from http://glass.ed.asu.edu/gene/papers/standards

**Goodwin, L. D.** (1999). Relations between Observed Item Difficulty Levels and Angoff Minimum Passing Levels for a Group of Borderline Examinees. *Applied measurement in Education.* 12, (1), 13-28.

**Goldman, A. I.** (1999). *Knowledge in a Social World*. Oxford: Clarendon Press.

**Green, B. F.** (2000). Setting Performance Standards. Paper presented at MAPAC meeting. Retrieved August 16 from: http://www.ipmaac.org/mapac/meetings/2000/berrtgre.pdf

**Haladyna, Th. & Hess, R. (2000).** An Evaluation of Conjunctive and Compensatory Standard-Setting Strategies for test Decision. *Educational Assessment*, 6, (2), 129-153.

**Hambleton, R. K.** (1978). On the Use of Cut-off Scores with Criterion-Referenced Tests in Instructional Settings. *Journal of Educational Measurement,* 15, (4), 277–289.

**Hambleton, R. K.** (2001) Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In G. J. Cizek (Ed.) *Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, N.J.: Erlbaum, 89-116.

**Hambleton, R. Jaeger, R., Plake, B. & Mills, C.** (2000). Setting Performance Standards on Complex Educational Assessments. *Applied Psychological Measurement,* 24 (4), December 2000, 355–366.

**Hambleton, R. & Slater, Sh.** (1997). Reliability of Credentialing Examinations and the Impact of Scoring Models and Standard-Setting Policies. *Applied Measurement in Education,* 10, (1), 19-38.

**Hansche**, **L**. (1998). *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I*., Washington, DC: US Department of Education and the Council of Chief State School Officers, Retrieved October 23, 2003 from SCASS CAS Publications and Products Web site: http://www.ccsso.org/projects/SCASS/Projects/Comprehensive_Assessment_Systems_for_ESEA_Title_I/Publications_and_Products/

**Haertel, E. & Lorié, W.** (2000) Validating Standards-Based Test Score Interpretations. Retrieved [December 12, 2003] from http://www-stat.stanford.edu/~rag/ed351/Std-Setting.pdf

**Huff, C.** (2001). *Overcoming Unique Challenges to a Complex Performance Assessment: A Novel Approach to Standard Setting.* Paper presented at the Annual meeting of NCME.

**Huynh, H.** (1998). On Score Locations of Binary and Partial Credit Items and their Applications to Item Mapping and Criterion-Referenced Interpretation. *Journal of Educational and Behavioral Statistics*, 23, (1), 35 – 56.

**Impara, J. & Plake, B.** (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 34, (4), 353-366.

**Impara, J. C. & Plake, B. S.** (1998). Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement*, 35 (1), 69-81.

**Jaeger, R. M.** (1989). Certification of student competence. In: *Educational Measurement*, **(**Third Edition**),** Ed. by R. L. Linn, Washington, DC: American Council on Education, 485-511.

**Jaeger, R.** (1991). Selection of Judges for Standard Setting. *Educational measurement: Issues and Practice,* 10, (2), 3-10.

**Jaeger, R. M., & Mills, C. N.** (2001). An integrated judgment procedure for setting standards on complex large-scale assessments. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum, 313-338.

**Kaftandjieva, F. & Takala, S.** (2000). Intra-judge Inconsistency or What Makes an Item Difficult for Experts. Paper presented at EARLI Assessment SIG Conference, Maastricht, The Netherlands.

**Kaftandjieva, F. & Takala, S.** (2002). Council of Europe Scales of Language Proficiency: A Validation Study. In: *Common European Framework of References for Languages: Learning, Teaching, Assessment. Case Studies*. Strasburg: Council of Europe, 106-129.

**Kaftandjieva, F. & Takala, S.** (2002). *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales*. Paper presented at Helsinki Seminar on Linking Language Examinations to Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

**Kaftandjieva, F., Verhelst, N. & Takala, S.** (1999). *DIALANG*: *A Manual for Standard setting procedure.* (Unpublished).

**Kaftandjieva, F., Verhelst, N.** (2000). *A new standard setting method for multiple cut-off scores.* Paper presented at LTRC 2000, Vancouver.

**Kane, M.** (1987). On the Use of IRT Models with Judgmental Standard Setting procedures. *Journal of Educational Measurement,* 24, (4), 333-345.

**Kane, M.** (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, (3), 425-461.

**Kane, M.** (2001). So Much Remains the Same: Conception and Status on Validation in Setting Standards. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum, 53-88.

**Kane, M., Crooks, T. & Cohen, A.** (1999). Designing and Evaluating Standard-Setting Procedures for Licensure and Certification Tests. *Advances in Health Sciences Education*, 4, 195–207.

**Kingston, N., Kahl, S. R., Sweeney, K., & Bay, L.** (2001). Setting performance standards using the body of work method. In G. J. Cizek (ed.), *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum, 219-248.

**Kolstad, A. & Wiley, D.** (2001). *On the Proficiency Penalty Required by Arbitrary Values of the Response Probability Convention Used in Reporting Results from IRT-based Scales.* Paper prepared for presentation to the annual meetings of the American Educational Research Association, Seattle, Washington, Retrieved [September 29, 2003] from http://www.c-save.umd.edu/ResearchPublicationsAndReports.html

**Linn, R. L.** (2001). *The Design and Evaluation of Educational Assessment and Accountability Systems*. CSE Technical Report 539. CREST/University of Colorado at Boulder.

**Linn, R. L.** (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives,* 11, (31). Retrieved [September 29, 2003] from http://epaa.asu.edu/epaa/v11n31/

**Livingston, S.** (1991). *Translating Verbally Defined Proficiency Levels into Test Score Intervals.* Paper presented at the Annual meeting of NCME, Chicago.

**Livingston, S. & Lewis, Ch.** (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement,* 32, (2), 179-197.

**Livingston, S. & Zieky, M.** (1982) *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ: ETS.

**Loomis, S. C., & Bourque, M. L.** (2001). From tradition to innovation: Standard-setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Standard setting: Concepts,methods, and perspectives.* Mahwah NJ: Erlbaum, 175-218.

**Maurer, T. J., Alexander,R. A., Callahan, C. M., Bailey, J. J.,&Dambrot, F. H.** (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel Psychology*, *44*, 235-262.

**Maurer, T. & Alexander, R.** (1992). Methods for Improving Employment Test critical Scores Derived by Judging Test Content: A Review and Critique. *Personnel Psychology,* 45, 277-745.

**Messick, S.** (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education.

**Messick, S.** (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

**Miller, M. & Linn, R.** (2000). Validation of Performance- Based Assessments. *Applied Psychological Measurement,* 24, (4), 367–378.

**Mills, C. & Melican, G.** (1988). Estimating and Adjusting Cutoff Scores: Features of Selected Methods. *Applied Measurement in Education*, 1, (3), 261-275.

**Mitzel, H. D. et al.** (2001). The bookmark procedure: Cognitive perspectives on Standard-setting. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum, 249-282.

**Nedelsky, L.** (1954). Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement*, 14, (1), 3-19.

**Norcini, J.,** (2003). Setting Standards on Educational Tests. *Medical Education*, 37, 464–469.

**Norcini, J. & Shea, J.** (1997). The Credibility and Comparability of Standards. *Applied Measurement in education,* 10, (1), 39-59.

**Norcini, J., Shea, J. and Kanya, D.** (1988). The Effect of Various Factors on Standard Setting. *Journal of Educational Measurement,* 25, 7–65.

**North, B.** (2002). Developing Descriptor Scales of Language Proficiency for the CEF Common Reference Levels. In: *Common European Framework of References for Languages: Learning, Teaching, Assessment. Case Studies*. Strasburg: Council of Europe, 87-105.

**Philips, S. E.** (2001). Legal Issues in Standard Setting for K-12 programs. In: G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, andperspectives.* Hillsdale NJ: Erlbaum, 411-426.

**Plake, B. & Hambleton, R.** (2000). A Standard-Setting Method designed for Complex Performance Assessment: Categorical Assignment of Student Work. *Educational Assessment*, 6 (3), 197-215.

**Plake, B. S., & Hambleton, R. K.** (2001). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, andperspectives.* Hillsdale NJ: Erlbaum, 283-312.

**Plake, B., Hambleton, R. & Jaeger, R.** (1997). A New Standard-Setting Method for Performance assessments: The Dominant Profile Judgment method and Some Field-test results. *Educational and Psychological Measurement*, 57, (3), 400-411.

**Plake, B. & Impara, J.** (2001). Ability of Panelist to Estimate Item Performance for a Target Group of Candidates: An Issue in Judgmental Standard Setting. *Educational Assessment,* 7, (2), 87-97.

**Plake, B., Melican, G. & Mills, C.** (1991). Factors Influencing Intrajudge Consistency During Standard Setting. *Educational Measurement: Issues and Practice,* 10, (2), 15-26.

**Popham, W. J.** (1978). As Always, Provocative. *Journal of Educational Measurement,* 15, (4), 297–300.

**Popham, W. J.** (1997). The Criticality of Consequences in Standard Setting: Six lessons learned the hard Way by a Standard Setting Abettor. Section 7 in *Proceedings of Achievement Levels Workshop*, Boulder, National Assessment Governing Board, U.S. Department of Education, The Nation's Report Card, NAEP, Retrieved December 4, 2003 from: http://www.nagb.org/pubs/conf_proc.pdf

**Putnam, S., Pence, P. & Jaeger, R.** (1995). A Multi-Stage Dominant Profile Method for Setting Standards on Complex Performance Assessments. *Applied Measurement in Education*, 8, (1), 57-83.

**Reckase, M. D.** (2000). A Survey and Evaluation of Recently Developed Procedures for Setting Standards on Educational Tests. In: *Student performance Standards on the National Assessment of Educational progress: Affirmations and Improvement*. Ed. By M. L. Bourqey & Sh. Byrd, Washington: NAEP, pp. 41 – 70.

***Random House Webster's Electronic Dictionary and Thesaurus***, (1992), College Edition, Version 1.0, Reference Software International.

**Raymond, M. & Reid, J.** (2001). Who Made Thee a Judge? Selecting and Training Participants for Standard Setting. In: G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, andperspectives*. Hillsdale NJ: Erlbaum, 119-173.

**Rudner, L.** (2003). *The Classification Accuracy of Measurement Decision Theory***.** Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago. Retrieved December 25, 2003 from: http://edres.org/mdt/papers/ncme2003c.pdf

**Rudner, L.** (2001). *Measurement Decision Theory*. Retrieved December 25, 2003 from: http://edres.org/mdt/

**Schulz, E. M., Kolen, M. J. & Nicewander, W. A. (1999).** A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores. *Applied Psychological Measurement,* 23 (4), 347–362.

**Schumacker, R.** (2003). Reliability of Rasch Measurement: Avoiding the Rubber Ruler. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

**Sireci, S.** (2001). Standard Setting using Cluster Analysis. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, andperspectives.* Hillsdale NJ: Erlbaum, 339-354.

**Smith, R. & Smith, J.** (1988). Differential Use of Item Information by Judges Using Angoff and Nedelsky Procedures. *Journal of Educational Measurement,* 25 (4), 259-274.

**Stephenson, A., Elmore, P. & Evans, Jh.** (2000). Standard-Setting techniques: An Application for Counseling Programs. *Measurement and Evaluation in Counseling and Development*, 32, 229-243.

**Stone, Gr. E.** (2002). *The Emperor has No Clothes:What Makes a Criterion-Referenced Standard Valid?* Paper presented at the Fifth Annual International Objective Measurement Workshop, New Orleans, Louisiana.

**Subkoviak, M. J.** (1984). Estimating the reliability of mastery-nonmastery classifications. In: R. A. Berk (Ed.), *A guide to criterion-referenced test construction,* Baltimore: The Johns Hopkins University Press, 267–290.

**Taube, K. (1997).** The Incorporation of Empirical Item Difficulty Data into the Angoff Standard-Setting Procedure. *Evaluation & the Health Professions*, 20 (4), 479-498.

**van der Linden, W. J.,** (1982). A Latent Trait Method for Determining Intrajudge Inconsistency in the Angoff and Nedelsky Techniques of Standard Setting. *Journal of Educational Measurement*, 19, (4), 295 – 308.

**van der Schoot, F. C. J. A.** (2002). *IRT-based method for standard setting in a three-stage procedure.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

**Verhelst, N.D., and Kaftandjieva, F.** (1999). *A rational method to determine cutoff scores (Research Report 99–07).* Enschede, The Netherlands: University of Twente, Faculty of Educational Science and Technology, Department of Educational Measurement and Data Analysis.

**Winter, Ph.** (2001). *Combining Information from Multiple Measures of Student Achievement for School-Level Decision-Making: An Overview of Issues and Approaches*. Washington: Council of Chief State School Officers, Retrieved December 30, 2003 from Center for the Study of Assessment Validity and Evaluation (C-SAVE) Web site: http://www.c-save.umd.edu/rept1_final.pdf

**Wright, B.** (1996). Reliability and Separation. Rasch Measurement Transactions, 9:4, p.472, Retrieved December 8, 1999 from: http://209.41.24.153/rmt/rmt94.htm

**Wright, B. & Masters, N.** (1982). Rating Scale Analysis. Chicago: MESA Press.

**Wright, B. & Grosse, M.** (1993). How to Set Standards. *Rasch Measurement Transactions.* 7:3, 315-6. Retrieved December 17, 2003 from Institute for Objective Measurement Web site: http://www.rasch.org/rmt/rmt73e.htm

**Zieky, M. J**. (2001). So Much Has Changed: How the setting of Cutscores Has Evolved Since 1980. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum, 19-52.

**A P P E N D I X**

| No | Method | Source | Judgment Task | | | Judgment Process | | | | Cut-off score establishment | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Test forma | Focus | Outcome | Feed-back | Data | Rounds | Decision making | Decision rule | Emp. data | Adjustment |
| **1.** | Angoff | Angoff, 1971 | Dichoto-mous items | Items | Estimated pro-bability of correct answer | No | No | 1 | Individual | Sum of estimated pro-babilities | No | No |
| **2.** | Angoff (Derivatives) | Loomis & Bourque, 2001 | Polytomous items | Items | Estimations of:<br>• Percent of partially correct<br>• Typical score<br>• Mean scores<br>• Probability for each score | ? | No | 1 | Individual | Sum of averages | No | No |
| **3.** | Angoff (adjusted) | Taube, 1997 | Dichoto-mous items | Items | Estimated proba-bility of correct ans-wer | No | No | 1 | Individual | Sum of estimated pro-babilities | Yes (IRT) | Yes |
| **4.** | Angoff 'Yes/No' | Angoff, 1971 | Dichoto-mous items | Items | Item classification | No | No | 1 | Individual | Sum of items correctly answered by a borderline person | No | No |
| **5.** | Angoff 'Yes/No' (modified) | Impara & Plake, 1997 | Dichoto-mous items | Items | Item classification | Yes | Yes | 2 | Individual + Revision | Sum of items correctly answered by a borderline person | No | No |
| **6.** | Ebel | Livingston & Zieky, 1982 | MC items<br>OE items | Items | • Item classification in two-way table (relevance-diffi-culty)<br>• Percentage of items in each cell to be answered correctly | No | No | 2 | Individual | Weighted sum of per-centages | No | No |

| No | Method | Source | Judgment Task | | | Judgment Process | | | | Cut-off score establishment | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Test forma | Focus | Outcome | Feed-back | Data | Round | Decision making | Decision rule | Emp. data | Adjust ment |
| **7.** | Nedelsky | Livingston & Zieky, 1982 | MC items | Items | Eliminated alternatives | No | No | 1 | Individual | Sum of estimated probability of correct answer | No | No |
| **8.** | Nedelsky (Modified) | Reckase, 2000 | MC items | Items | Probability of eliminating each distractor | No | No | 1 | Individual | $P=\Sigma(pi+1)/n$ | No | No |
| **9.** | Jaeger | Jaeger, 1989 | MC items OE items | Items | Item classification | Yes | Yes | 3 | Individual + Revision | Sum of items correctly answered by a person on a specific level | Yes | Yes |
| **10.** | Item Score Distribution | Reckase, 2000 | Polytomous items | Items | Probability distribution of item scores at the borderline | No | No | 1 | Individual | Average | No | No |
| **11.** | Compound cumulative | Kaftandjieva & Takala, 2002 | MC items OE items | Items | Item classification | Yes | No | 1 | Individual | Sum of items in the lower category (averaged) | Yes | Yes |
| **12.** | Item score string estimation | Loomis & Bourque, 2001 | Polytomous items | Items | Estimated item scores for a borderline person | Yes | Yes | 2 | Individual+ Revision | Sum of averages | No | No |
| **13.** | Cluster | Sireci, 2001 | All | Items | Domain classification | No | No | 1 | Group Consensus | K-means cluster analysis | Yes | No |
| **14.** | IRT modeling of judgments | Kane, 1987 | Dichotomous items | Items | Estimated probability of correct answer | No | No | 1 | Individual | Minimizing Loss function | Yes (IRT) | Yes |
| **15.** | Item Mastery | Verhelst & Kaftandjieva, 1999 | Dichotomous items | Items | Item classification | Yes | No | 1 | Individual | Minimizing Loss function | Yes (IRT) | Yes |

| No | Method | Source | Judgment Task | | | Judgment Process | | | | Cut-off score establishment | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Test forma | Focus | Outcome | Feed-back | Data | Round | Decision making | Decision rule | Emp. data | Adjust ment |
| **16.** | Objective stan-dard setting | Wright & Grosse, 1993 | Dichoto-mous items | Items | Item classification | ? | Yes | 2 | Individual+ Revision | Direct establishment | Yes (IRT) | Yes |
| **17.** | Bookmark (Item mapping) | Mitzel et al., 2001 | MC items OE items | Item map | Cut-off scores | Yes | Yes | 3 | Individual + Revision | Median cut-off score | Yes (IRT) | Yes |
| **18.** | Multistage IRT | van der Schoot, 2002 | MC items OE items | Item map | Cut-off scores | Yes | Yes | 3 | Individual + Revision | Direct establishment | Yes (IRT) | Yes |
| **19.** | Combined judgment-empirical | Livingston, 1991 | Dichoto-mous items | • Items<br>• Mastery level | • Item classification<br>• Level specific pro-bability of success | Yes | Yes | 2 | • Individual + Revision<br>• Group Consensus | Median $\theta$ value for the group of items at the specified probability of success level | Yes (IRT) | Yes |
| **20.** | Item Domain | Schulz et al., 1999 | Dichoto-mous items | • Items<br>• Mastery level | • Item domain cla-ssification<br>• Probability of suc-cess | No | No | 1 | ? | $\theta$, corresponding to the established probability of success | Yes (IRT) | No |
| **21.** | Cognitive Components | Reckase, 2000 | All | • Items<br>• Cognitive componen ts | • Item decomposi-tion in cognitive components<br>• Cognitive compo-nents probability of success | No | No | 2 | Individual | Aggregated product of probabilities | No | No |

| No | Method | Source | Judgment Task | | | Judgment Process | | | | Cut-off score establishment | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Test forma | Focus | Outcome | Feed-back | Data | Rounds | Decision making | Decision rule | Emp. data | Adjustment |
| **22.** | Multistage Aggregation | Reckase, 2000 | All | • Items<br>• Profiles<br>• Examinee performance | Item classification<br>Profile classification<br>Cut-off score | ? | ? | 4 | Individual | Logistic regression | Yes | No |
| **23.** | Border Group | Livingston & Zieky, 1982 | All | Examinees | Examinee classification | No | No | 1 | Individual | Median of the score distribution | Yes | No |
| **24.** | Contrasting Groups | Reckase, 2000<br>Brandon, 2002<br>Clauser & Nungester, 1997 | All | Examinees | Examinee classification | No | No | 1 | Individual | Intersection point of the score distributions | Yes | Yes |
| **25.** | Body of work | Kingston et al., 2001 | All | Examinee overall performance | Examinee classification | Yes | No | 3 | Individual + Revision | Logistic regression | Yes | Yes |
| **26.** | Generalized Examinee-Centered | Cohen, Kane & Crooks, 1999 | All | Examinee overall performance | Examinee classification | Yes | No | 1 | Individual | Curve-fitting between ratings and test-scores | Yes | No |
| **27.** | Analytical Judgment (Anchor-Based) | Plake & Hambleton, 2001 | All | Examinee performances | Examinee rating | Yes | No | 2 | Individual + Revision | Average of borderline scores | Yes | No |
| **28.** | Examinee Paper Selection | Hambleton et al., 2000<br>Hansche, 1998 | Polytomous items | Examinee performances | Borderline performance | No | No | 3 | Individual + Revision | Sum of averages | Yes | No |

| No | Method | Source | Judgment Task | | | Judgment Process | | | | Cut-off score establishment | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Test forma | Focus | Outcome | Feed-back | Data | Round | Decision making | Decision rule | Emp. data | Adjust ment |
| **29.** | Integrated Judgment (holistic; booklet classifi-cation) | Jaeger & Mills, 2001 | All | Examinee booklets | Examinee classifica-tion | Yes | Yes | 2 | Individual + Revision | Average<br>Linear regression | Yes | Yes |
| **30.** | Measuremen t Decision Theory | Rudner, 2003 | Dichoto-mous items | • Population<br>• Items | • Proportion at each level<br>• Level specific item difficulty | No | No | 1 | Individual | Maximum a posteriori decision criterion | Yes | No |
| **31.** | Hofstee | Case & Swan-son, 1998<br>Huff, 2001 | All | Score distribution | • Min & max failing rates<br>• Min & max cut-off points | ? | ? | 1 or 2 | Individual | Intersection between the cumulative score distribution curve and the diagonal of the min-max square | Yes | Yes |
| **32.** | Judmental Policy Capturing | Hambleton et al., 2000<br>Hansche, 1998 | Performance assessment | Score profiles | Profile classification | Yes | Yes | 2 | Individual + Revision | Multiple regression analysis | Yes | Yes |
| **33.** | Direct Judgment | Hambleton et al., 2000 | All | Score profiles | • Task weights<br>• Overall cut-off score | ? | ? | ? | Individual | Average | Yes | No |
| **34.** | Dominant Profile Judgment | Putnam et al., 1995 | Complex performance assessment | Standard setting strategies | Standard setting policies | Yes | ? | 3 | Consensus building strategy | Prevailing standard setting strategy | No | No |

# Section C

# Classical Test Theory

**N.D. Verhelst**
**National Institute for Educational Measurement (Cito)**
**Arnhem, The Netherlands**

In this section an overview is given of the main concepts and theoretical results of Classical Test Theory (CTT). The text has been written to be as accessible as possible for the non-technical reader. The first two sections (Basic Concepts and Procedures) do not contain any formulae. They are meant to be read as a whole, since concepts introduced at the start are used in later parts. As CTT is a statistical theory, it is not possible to present and discuss it in great depth without having recourse to the exact and compact way of expression provided by mathematical formulae. Where it is felt that some deeper understanding of the theory might be wished, reference is made to a more technical section. These technical sections are stand-alone sections, and are added to the main text in the order they are referred to.

Classical Test Theory has been used for more than fifty years as a guide for test constructors to understand the statistical properties of test scores, and to use these properties to optimise the test under construction in a number of ways. The main purpose of this appendix is to review the main issues of Classical Test Theory, and to emphasise what can be expected from Classical Test Theory and what not. We will first present some basic concepts and then go on to procedures which are used in the framework of Classical Test Theory.

## C.1. Basic Concepts

**Items**. In many cases a test is composed of a number of elementary parts, for example, twenty questions. A generic name for such a part is: 'item'. There is, however, no stringent rule of identifying items with questions. Suppose a reading test consists of five text passages, and four questions are to be answered about each passage. One might conceive the twenty questions as twenty items, but one might also consider the four questions associated with each text as a single item. In the latter case, one sometimes refers to those composite items as super items, testlets or item bundles.

**Observed score.** When a test is administered, the result is summarized by a **number** (for example, the number of correct item responses). This number is called the (observed) **test score.** Usually the test score is the sum of the **item scores.** In all analyses to be carried out in CTT, the item scores are usually the basic quantities that enter such analyses. But it should be kept in mind that these scores are not given as such; they come about through a decision by the test constructor, and CTT does not provide any rules for taking such a decision. It is customary to grant one point for the 'correct answer' in a multiple choice item, and zero points for any other choice. In some cases, however, it might be more informative to grant 2 points for a particular choice, 1 point for another (not optimal) choice and zero points for the remaining choices. The actual choice the test taker makes is the basic observation; the granting of points is a decision to be taken a priori, sometimes on intuitive grounds, sometimes on the basis of extended qualitative studies and quantitative analyses of the set of observations. Therefore it is wise to keep as detailed records of the observations as possible: for multiple choice questions, the option actually chosen; for open ended questions, it is advisable to develop a quite detailed categorizing system, and to keep records (in a data base) of as much detail as possible. To the data stored in this manner, different scoring rules may then applied, yielding in each case a file with (numerical) item scores which may then be submitted to quantitative analyses.

**True score.** The basic assumption of CTT is that in a second administration of the same test to the same person under similar circumstances as the first time, we will probably not observe the same score as the first time. This reasoning can be generalized to an arbitrary number of similar test admini-

strations, giving rise to the idea of a **distribution of (possible) test scores.** This distribution is associated with a single person, and hence could be characterized as his or her 'private' distribution. In CTT the average of this private distribution is called the person's true score. True score is a statistical concept, and has nothing to do with conceptions like 'ideal score' or 'the score a person really deserves'. The observed score actually obtained is conceived as a sample (of size 1) from the 'private' distribution. True scores are not observed. The true score of a person is symbolized (in this appendix) with the Greek letter tau ($\tau$). Notice that it is a number.

**Measurement error**. In CTT the measurement error is defined as the difference between the observed score and the true score. If the observed score is greater than the true score, we say that the measurement error is positive; if it is smaller, the measurement error is negative. Since the true score of a person is not known, the measurement error (in a particular case) is not known either. It is possible, however, to say something more concrete of measurement errors in a population. The symbol used for the measurement error is E.

**Variability: standard deviation and variance.** Phenomena showing no variability are not very informative. If everybody (from a certain population) gets the maximum score on a test, all one can say is that the test is apparently too easy for this population. Things are becoming interesting if they show variability, as test scores in a calibration sample usually do. In statistics one needs a **measure of variability**. A well known measure is the standard deviation. The variance is the square of the standard deviation. Although the standard deviation is usually easier to interpret, the variance is a more useful concept in statistics (e.g., in such techniques as analysis of variance.)

**Sources of variance.** Suppose John's observed score is 18 and Mary's is 20. One could ask why these observed scores differ. CTT distinguishes two sources of variability: the scores may differ because John's and Mary's true scores differ or because the two measurement errors differ; or both. These two sources cannot be disentangled at the individual level, i.e., we cannot know the answer in the concrete case of John and Mary; but they can be distinguished at the level of the population. In the population the true score is not a number, but a variable (which can assume different values for different persons). To indicate the true score as a variable we use the symbol T. The important result is that (in the population) the variance of the observed scores is the sum of the variance of the true scores and the variance of the measurement errors. (Notice that this decomposition rule does not hold for standard deviations.) Shorthand names are sometimes used: observed variance for the variance of observed scores, true and error variance for variance of true scores and measurement errors, respectively.

**Reliability of test scores**. The reliability of test scores is defined as the ratio of the true variance to the observed variance. Multiplied by 100, it can be interpreted as a percentage: it is the percentage of the observed variance which is true variance. The minimum value of the reliability is zero, meaning that all variation in the observed scores is due to measurement error. The maximum value is one, meaning that there is no measurement error. A reliability coefficient of 0.8 means that 80% of the observed score variance is due to variation in the true scores and 20% to measurement error. Reliability is a key concept in CTT, but from the definition it is not clear how it can be determined. Further down, this problem will be discussed, together with some examples of the importance of the concept. The expression 'reliability of a test' is often used, but it is not correct; it should be understood as 'reliability of test scores'.

## C.2 Procedures

### *P*-values.

In the process of constructing test items it is important to have a rather precise idea of the target population. Administration of items that are too easy or too hard is not adequate for several reasons. It may lead to boredom or frustration, which in turn will almost invariably cause loss of motivation for the test taker. Moreover, in this case, the item responses will give very little information about the proficiency level of the test takers. Therefore, it is important to have a rather precise idea about the degree

of difficulty of the items; decisions on inclusion or exclusion of items are often based on information about their degree of difficulty, usually called *p*-values. (The '*p*' refers to proportion or probability.) For binary (scored 0/1) items, the *p*-value of an item is the proportion of correct responses in the population. Usually, a *p*-value is considered as a property of an item, which is correct, as long as one realises that this property is valid only with respect to a certain population. A common way of expressing this relativity is to say that *p*-values are **population dependent**. This can easily be understood with the following simple example. Suppose an item is developed for a test to be applied in the fourth year of English learning. With respect to this population, let us assume that the item is rather easy, and has a *p*-value of 0.8. It will easily be understood that such an item may be much harder in the population of second year students, yielding a *p*-value of 0.25 or even lower in this population. Thus speaking of the *p*-value of an item has no meaning; implicitly or explicitly there is always a reference to some population. This population dependency has immediate implications when one tries to establish the psychometric properties of a test from the sample observations. The sample must be representative for the population.

Note 1. *P*-values are values which pertain to items in some population, but they are computed on a sample. Representativeness of the sample does not mean that the value computed will be equal to the value in the population. If we compute the *p*-value of an item in two independent samples, we will usually find two different values. The *p*-value found in the sample is to be considered as an **estimate** of the *p*-value in the population. The accuracy of the estimate depends mainly on the sample size. Details and examples are given in Section C.3

Note 2. Items where one can get 0, 1 or 2 points, or 0, 1, 2 or 3 points, etc., are called partial credit items. *P*-values of partial credit items are defined as the average relative score. See Section C.4 for details.

Note 3. It is common to interpret *p*-values as measures of difficulty, but notice that the higher the *p*-value, the easier the item is. Some authors use $1 - p$ as the measure of difficulty. Both measures are acceptable, as long as it is clearly indicated which one is used.

**Item discrimination**

Simply stated, the discriminating power of an item is to extent to which it is possible to separate high proficiency levels from low levels on the basis of the responses to the item. Or, stated otherwise: what is the psychometric quality of a test which consists of this particular item? Suppose that a quite difficult binary item is used as a test. We will say that the item discriminates well if the very best students have the item correct, and the others not, but since a binary item has only two categories (right or wrong), if the item separates the very best from the others, it cannot separate the students of medium proficiency from the weak ones. That is, discrimination is a local property, and it is fairly difficult to catch (and describe) the discriminating power of an item in a single number. Yet, there exist several indices of discrimination which are used within CTT. We list some of them:

- the correlation between item score and test score (item-test correlation);
- the correlation between item score and the score on the test with that item excluded (item-rest correlation);
- in particular for multiple choice items: the correlation between test score and each of the distractors.

Item-test and item-rest correlations should be positive; correlations between the test score and the distractors should be negative. (See Section C5 for the exact meaning of this notion) Rules of thumb for a minimum value of item-test or item-rest correlations may be misleading, because the correlation is strongly influenced by the *p*-value of the item.

**Graphical Item Analysis**

The usual output from software for item analysis consists of a number of tables containing *p*-values, discrimination indices like item-test and item-rest correlations, and other indices usually interpreted

also as indices of discrimination. There exists, however, a simple and powerful tool to judge the quality of the items. Each item is represented by one or more curves as exemplified in Figure C.1

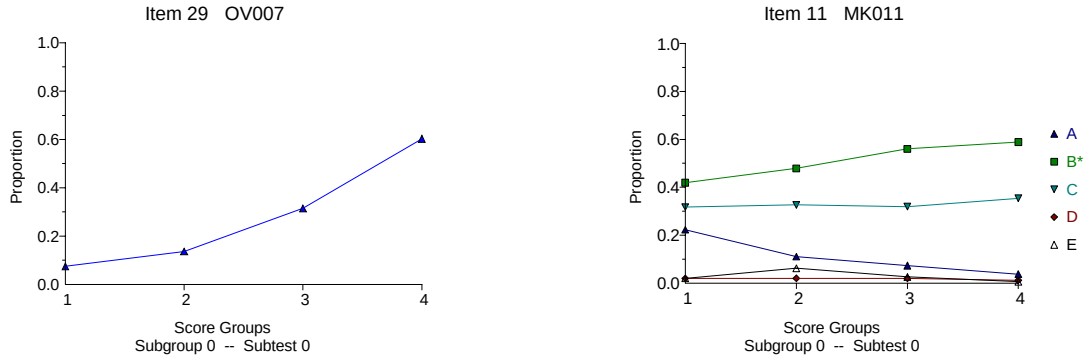



Figure C.1. Graphical item analysis

The figures are constructed using the same principle: the total sample is split into a small number of homogeneous groups (four in the examples; '1' denotes the groups with the lowest scores, '4' the group with the highest scores, and '2' and '3' intermediate groups) on the basis of the test scores. In each group the proportion of correct responses is computed and plotted against the group number, as is shown in the left-hand panel of the figure (item 29). One sees immediately that the item is relatively difficult: even in the highest group (4) only about 60% of the test takers gave a correct answer. We also see that the item-test correlation will be positive: the higher the group number, the higher the proportion of correct responses. In the right-hand panel a similar figure is drawn for a multiple choice item with five alternatives (where B is the correct alternative). Here we see immediately that the item is not of a very high quality: the discriminating power is low (the curve for alternative B increases very slowly); the distractors D and E are almost never chosen (and so prove to be useless as distractors), and distractor C remains attractive at a constant and quite elevated level (more than 30%), which may suggest that this item is a catch item. In summary, the figure suggests clearly that the item deserves revision, and cannot function as a 'model item' to help train item writers. More examples are given and discussed in Section C.6.

The figures displayed above are standard output of the computer program TiaPlus. To obtain this program, a request should be sent to Ton.Heuvelmans@Citogroep.nl.

**Estimation of Reliability**

From the definition of reliability, it is clearly not possible to compute the reliability coefficient directly, because of the presence of a quantity which is not observable: the variance of the true scores. In order to compute the reliability, a new concept has to be introduced, the concept of a **parallel test**. Two tests are parallel if the following two conditions hold: the true scores on both tests are equal for all persons in the population, and the variance of the measurement error is equal in both tests.

An important and reassuring result of CTT is that the reliability of a test equals the correlation between the test and a parallel test. Two parallel tests have the same reliability.

There are two problems associated with this finding: (1) how do we know that two tests are parallel, and (2) in order to compute the correlation, we need test scores on the same sample of test takers on the two tests, i.e., two test administrations are required. We comment on both problems.

1 The construction of parallel tests
   a Two parallel tests have the same average observed score and the same observed variance. Moreover, their correlation with all other tests, whatever these measure, should be the same. But this holds in the population; we cannot expect that these equalities will also hold in a sample. In practice, significance testing can be used, but one should be careful: if the differ-

ence between two sample averages does not differ significantly from zero, this does not imply that the population differences do not differ. The risk that a real difference in the population is not detected by a significance test is larger when the sample size is small.

b Two methods are commonly used in applications of CTT, parallel form and retesting. In the retesting method, the same test is applied at two different points in time. The main threat to parallelism is the memory effect. Here we have to distinguish between two cases:
   i) In general memory effects are beneficial to the test performance, yielding a higher test score on the second administration than on the first one. If memory effects are uniform, i.e., if the increase from the first to the second administration (in true score) is the same for every person, the two series of test scores are not parallel, but their correlation nevertheless is the reliability of the test. If the increase is uniform, the two (population) means may differ, but the variances will not differ.
   ii) If memory effects are not the same for every person, the retesting will not yield a parallel form. This may occur when there are ceiling effects: low scores in the first administration may increase considerably by memory effects, but high scores may probably not increase by the same amount, because they are already close to the maximum score of the test. If this is the case, the correlation between the two series of test scores is not the reliability.

   The construction of parallel forms is not easy either. A necessary condition for parallelism is that the contents of both forms should be comparable, which may be hard to accomplish in cases where complex items are constructed (e.g., a text passage with four or five associated questions). There exists a rather simple method to use psychometric indices to aid in constructing parallel forms. This method is discussed in section C.7.

c Sometimes, only one test is available, but for the sake of estimating the reliability it is split into two halves which are meant to be parallel. Notice that the correlation between the two halves – if they are really parallel - is not the reliability of the test, but of the half tests. To obtain the reliability of the test, the Spearman-Brown formula has to be applied (see below).This method is known as the split-half method. If the two halves are not parallel, the resulting coefficient underestimates the reliability.

2 Reliability estimation from a single test administration

a In principle it is impossible to determine the reliability of a test from a single test administration. All that can be reached is a so called lower bound to the reliability; this is a number such that one can be certain that the reliability is not lower than that number. If for a given test this lower bound is 0.7, all one can be sure of is that the reliability is at least 0.7. If the lower bound is high (more than 0.95, for example) this will not be a big problem. If it is low, however, 0.30 say, it does not follow that the reliability is that low.

b The best known lower bound is Cronbach's coefficient alpha. It can be used for any mixture of binary and partial credit items.

c The KR20-coefficient is the same as Cronbach's alpha, but it is defined only for binary items.

d Cronbach's alpha is sometimes labelled as an index of internal consistency, i.e., an index that shows the extent to which all items in the test measure the same concept. If the test is really one-dimensional, the index will be close to the reliability; if the test is heterogeneous, alpha can be substantially lower than the reliability.

e There exist more lower bounds. In fact there exists a **greatest lower bound** (GLB). It is at least as large as all possible lower bounds. The computation of the GLB is not easy (there does not exist a closed formula), but it is available in published software; the program TiaPlus does compute it.

f Lower bounds such as Cronbach's alpha, the KR20 and the GLB are quantities which apply to the population. They are estimated from the calibration sample and contain an estimation error. The estimate of the GLB from small samples tends to be a serious overestimate of the population GLB. In the program TiaPlus, a correction to this bias is applied if the sample size is not too small.

**The Spearman-Brown formula.** Tests are administered to collect information on a person's proficiency. The information is conveyed through the scores obtained on the items, but we have to admit that these scores contain errors, some positive, others negative. By summing the item scores, positive

and negative errors will tend to cancel each other, the more so if the test gets longer. It follows that we can trust the result of a long test generally more than the result of a short test, or, what is the same, the reliability of a long test is higher than that of a short test. The Spearman-Brown formula expresses the relation between test length and reliability. It can be used in two ways, which we illustrate with an example:

1. A test consisting of 25 items has a reliability of 0.7. What will the reliability be if 10 items are added? (The answer is 0.766; see Section C.8)
2. A test consisting of 25 items has a reliability of 0.7. How many items must the test contain to have a reliability of 0.8? (The answer is 43; see Section C.8.)

The second example shows how the Spearman-Brown formula can be used to plan work on extra item writing. It should be noticed that it is more expensive (in terms of the number of items) to raise the reliability from 0.8 to 0.9 than from 0.7 to 0.8. The increase from 0.7 to 0.8 requires 43 – 25 = 18 extra items; to reach 0.9, another 54 items are required.

The Spearman-Brown formula must be used very cautiously: it only applies if the added items are of the same quality as the items already present. The standard expression is that the test must be lengthened homogeneously.

The formula can also be used in the reverse sense: if a planned test with a known reliability happens to be too long to be useful in practice, the formula can be used to compute the reliability of a shortened version of the test. Taking the example above: if the test with 43 items and a reliability of 0.8 is shortened (homogeneously) to 25 items, the shorter version will have a reliability of 0.7.

Finally, it can be used to compute the reliability in case of the split-half method. If the correlation between the two test halves is symbolized as $r$, the reliability of the full test is $2r/(1+r)$.

**The Standard Error of Measurement**. Although we can never know in a particular case what the measurement error is, we can have a quite precise idea of the magnitude of the measurement error 'on the average'. Recall the 'private' distributions of the observed scores. If in such a private distribution of possible observed scores all (or most) of the values are very near to the average (the true score), this distribution will have a small standard deviation; if on the contrary, many values are far away from the average, the standard deviation will become large. So the standard deviation of the private distribution gives an indication of a typical error. This standard deviation is called the standard error of measurement.
There is a strong relation between the standard error of measurement and the reliability of the test: the standard error of measurement is the standard deviation of the observed scores (in the population) multiplied by the square root of one minus the reliability.

The standard error of measurement can be used to define confidence intervals for the true score. It is instructive to look at examples of such confidence intervals to learn about the relative merits of testing. Even with a reliability as high as 0.96, the 90% confidence interval for the true score is larger than half a standard deviation. Details are discussed in Section C.9.

Decisions on individuals are sometimes based on a test score, for instance an examination score. One should realize that such decisions are of necessity based on observed test scores, which contain an unknown measurement error. This implies that able candidates may fail on an examination because of a negative measurement error, and weak candidates may succeed because of a positive error. This leads to wrong (unintended) classifications. The percentage of such erroneous classifications depends strongly on the reliability of the test. Even if it is as high as 0.9, the percentage of wrong classifications can be substantial.

**Kelley's formula**. Sometimes an estimate of the true score is needed. The best known estimate is computed using the famous formula by Kelley. The result of this formula is a compromise between the observed score and the population mean of the scores. A compromise means a weighted sum; the

weight of the observed score is the reliability of the test, the weight for the population mean is one minus the reliability. Suppose X = 112 and the population mean is 100; the reliability equals 0.88. Kelley's estimate of the true score is $112 \times 0.88 + 100 \times (1 - 0.88) = 110.56$. Notice that the estimate is closer to the population mean than is the observed score. This is known as 'shrinkage'. This estimate can be interpreted as follows: it is the average true score of all people in the population having an observed score of 112. If John's observed score happens to be 112, we cannot infer from this that his true score is exactly 110.56, i.e., the estimate also contains an error. This error is called the estimation error, and its standard deviation is called the standard error of estimation. It is smaller than the standard error of measurement.

**Theoretical results**.

There are three important results which are useful in the discussion of external validation. One can conceive test results as measurements that are polluted in some way by measurement error. It may be interesting to know as precisely as possible what the results would be if one could measure without measurement errors, i.e., the results in the ideal case where the observed scores are equal to the true scores. These are the results: (details can be found in Section C.10)

1. The correlation between observed scores and true scores is the square root of the reliability.
2. The correlation between the observed scores on two tests is 'attenuated' (lowered) by the unreliability of the two tests. The correlation between the true scores on both tests equals the correlation between the observed scores divided by the square root of the product of their reliabilities. The corresponding formula is called the correction for attenuation.
3. If two tests really measure the same concept, the correlation between the true scores of both tests should equal one. If this is the case, the tests are called **congeneric**. But the correlation between the observed scores will be attenuated by their unreliability. If two tests are congeneric, the correlation between the observed scores is equal to the square root of the product of their reliabilities.

**Population dependency**

In the discussion on the *p*-values, it was stressed that it is meaningless to speak about the *p*-value of an item, because there is always a reference (explicitly or implicitly) to a certain population. The same argument applies to all item- and test-indices that are used in Classical Test Theory. In particular it applies to the concept of reliability. The reliability of a test is a characteristic of the test scores in some population. The same test can have a high reliability in some population and a very low one in another population. Here is an example. Suppose a test is used as an entrance test to the university, and assume it has a reliability of .85 in the population of candidates. This very same test will have a lower reliability in the population of first year students at the university, because this population is more homogeneous with respect to true score than the population of candidates, i.e. the variance of the true scores at the university will be smaller than in the population of candidates. Or more generally, the more homogeneous the population (with respect to true score), the lower the reliability will be. But, of course, this is not the only reason why the reliability of a test can be low. Sloppy items with ambiguous scoring rules will usually lead to low reliability, and one cannot use the homogeneity of the population as an excuse for the bad quality of the test.

## C.3. The accuracy of *p*-values

A good method of getting an impression of a *p*-value of an item is to construct **confidence intervals**. A *p*-value is a theoretical quantity which applies to the population, and which one usually estimates by a corresponding quantity in the sample. If the *p*-value of an item in the population is 0.75, say, it is almost sure that one will not find a proportion correct of 0.75 in the sample. But in general, we do not know the population value, we only observe a proportion correct in the sample. The problem of **inferential statistics** is to make clear what one can say about the population value on the basis of a sample value. To this end, one usually constructs **confidence intervals**. In what follows, the theory of confidence intervals is summarized and a practical formula for constructing intervals is given.

We represent the unknown *p*-value in the population by the Greek letter $\pi$; the proportion one can observe in a sample is denoted as *p*. The observed proportion is called a **random variable**, because it can assume different values in different samples.

1. Assume we could draw a very great number of samples, all independent of each other, and all of the same size, *n*. In each sample we can compute the observed *p*-value, and we can construct a histogram with these *p*-values. From theoretical statistics we can tell interesting things about this histogram:
    a. Its average equals the unknown value $\pi$;
    b. Its standard deviation equals $\sqrt{\pi(1-\pi)/n}$ ; this standard deviation is called the standard error of the random variable *p*;
    c. The form of the histogram looks very much like the graph of the normal distribution, and the similarity is more striking for large *n* than for small *n*.
2. Of course, we do not draw many samples, we usually draw a single one, but from the theoretical results we can say that the *p*-value we will observe will, with a probability of 90% lie in an interval from the mean ($\pi$) minus 1.645 times the standard deviation to the mean plus 1.645 times the standard deviation. The value of 1.645 is to be found in published tables of the normal distribution. If we want a 95% interval, we have to replace 1.645 by 1.96, and for a 99% interval, we use 2.58.
3. We express the preceding paragraph by means of a formula:

$$P\left(\pi - 1.645\sqrt{\frac{\pi(1-\pi)}{n}} \le p \le \pi + 1.645\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 0.9 \qquad \text{(c1)}$$

4. The expression between parentheses in the preceding formula is called an **event** (*p* lies in some interval). The whole formula reads as: the probability of this event is 0.9 But we can replace this event by an equivalent event. We do this in two steps: the first step concentrates on the first inequality, where we move the term with the square root to the other side of the inequality sign:

$$\pi - 1.645\sqrt{\frac{\pi(1-\pi)}{n}} \le p \;\Leftrightarrow\; \pi \le p + 1.645\sqrt{\frac{\pi(1-\pi)}{n}}$$

and, in the second step (concentrating on the second inequality in formula (c1)) by a similar move we get:

$$p \le \pi + 1.645\sqrt{\frac{\pi(1-\pi)}{n}} \;\Leftrightarrow\; p - 1.645\sqrt{\frac{\pi(1-\pi)}{n}} \le \pi$$

and combining the two right-hand sides gives

$$p - 1.645\sqrt{\frac{\pi(1-\pi)}{n}} \le \pi \le p + 1.645\sqrt{\frac{\pi(1-\pi)}{n}}$$

and this event reads as: the population value $\pi$ is embraced by two values which will vary from sample to sample, because the observed *p*-value is a random variable. And since we work with equivalent events, we can say that

$$P\left(p - 1.645\sqrt{\frac{\pi(1-\pi)}{n}} \le \pi \le p + 1.645\sqrt{\frac{\pi(1-\pi)}{n}}\right) = 0.9 \qquad \text{(c2)}$$

5. It deserves some attention to understand well the equivalence of (c1) and (c2) and the difference in wording of the two statements. In (c1) we say that the event is that the value of a random variable

($p$) will lie between two fixed values; in (c2) we say (equivalently) that the fixed population value ($\pi$) will be embraced by two variable bounds.

6. There is, however, a further problem with formula (c2): the two bounds depend on the variable $p$, but also on the unknown value of $\pi$. In practice, then, one replaces $\pi$ by the observed value $p$ under the square root sign, giving a practical formula:

$$P\left( p - 1.645\sqrt{\frac{p(1-p)}{n}} \le \pi \le p + 1.645\sqrt{\frac{p(1-p)}{n}} \right) \approx 0.9 \qquad \text{(c3)}$$

7. Here is a simple example. Suppose $p = 0.51$ and $n = 100$. Then, $\sqrt{0.51(1-0.51)/100} = 0.04999$ ($\approx 0.05$), and using these values in (c3), we find that

$$\begin{aligned} &P(0.51 - 1.645 \times 0.05 \le \pi \le 0.51 + 1.645 \times 0.05) \\ &= P(0.428 \le \pi \le 0.592) = 0.9 \end{aligned}$$

8. Notice that the observed $p$-value (0.51) lies precisely in the middle of the defined interval, or, as one says, the confidence interval is symmetric around the observed $p$-value. If the observed $p$-value is around 0.5, this is reasonable. But now, suppose that the observed $p$-value is as high as 0.95, $n$=100 and we want a 99% confidence interval. The standard error of $p$ is now approximated by $\sqrt{0.95(1-0.95)/100} \approx 0.0218$ and $2.58 \times 0.0218 = 0.056$ so that we find

$$P(0.894 \le \pi \le 1.006) = 0.99$$

but the upper bound of the confidence interval is larger than 1, while we know that $\pi$ can not be larger than one. Moreover, with very high observed $p$-values, we would rather believe that the population value is smaller than that it is larger than the observed value. But this asks for an **asymmetric** interval, for which we need another formula. Here is one which looks complicated but which gives nice results in many cases (Hays, 1977, p. 379[1]):

$$\frac{n}{n+z^2}\left[ p + \frac{z^2}{2n} \pm z\sqrt{\frac{p(1-p)}{n} + \frac{z^2}{4n^2}} \right]$$

In the formula, $z$ stands for the value from the tables of the normal distribution: 1.645 for a 90% interval; 1.96 for a 95% and 2.58 for a 99% interval. The sign '±' must be replaced by a '+' to yield the upper bound and by a '-' to find the lower bound of the interval. We apply this to the preceding example ($2.58^2 = 6.656$), finding

$$\begin{aligned} &\frac{100}{100+6.656}\left[ 0.95 + \frac{6.656}{200} \pm 2.58\sqrt{\frac{0.95 \times 0.05}{100} + \frac{6.656}{4 \times 100^2}} \right] \\ &= \frac{100}{106.656}\left[ 0.983 \pm 2.58 \times 0.0253 \right] \end{aligned}$$

which gives 0.860 as lower bound and 0.983 as upper bound. Notice that the observed $p$-value of 0.95 is much closer to the upper bound than to the lower bound.

[1] Hays, W.L., *Statistics for the social sciences*. London: Holt, Rinehart and Winston, 1977[2].

## C.4. Partial credit items and *p*-values

A binary item is an item where the score can assume only two values: zero for an incorrect and one for a correct response. A partial credit item is an item where the score can range from zero to a certain maximum that is larger than one, and where all intermediate (whole numbered) scores can be obtained as 'partial credits'. The simplest form is where one gets two points for a perfect response, zero points for a totally wrong answer and one point for an answer that is neither totally wrong nor totally correct.

The (observed) *p*-value of a binary item is the proportion of test takers in the sample having the item correct. When one tries to generalize the definition of the *p*-value for binary items to partial credit items, one runs into trouble, because the notion of 'correct' becomes ambiguous in this case. There is, however, a convenient way to look at *p*-values which easily generalizes to partial credit items, namely, the notion of average relative (item) score. For binary items this is illustrated in Table C.1 with a numerical example and symbolically.

Table C.1 The observed *p*-value as average score

| | example | | symbolically | |
|---|---|---|---|---|
| score | frequency | proportion | frequency | proportion |
| 0 | 189 | 0.30 | $N_{i0}$ | $1-p_i$ |
| 1 | 441 | 0.70 | $N_{i1}$ | $p_i$ |
| total | 630 | 1 | $N_i$ | 1 |

The average score on this item is computed as

$$\frac{189\times 0+441\times 1}{630}=\frac{189}{630}\times 0+\frac{441}{630}\times 1=\frac{441}{630}=0.7=p_i$$

So, in the case of a binary item, we see that the proportion correct or the average score mean the same thing. Now we apply the same procedure to a partial credit item with a maximum score of 3. (See Table C.2.)

Table C.2 The average item score for a partial credit item

| | example | | symbolically | |
|---|---|---|---|---|
| score | frequency | proportion | frequency | proportion |
| 0 | 126 | 0.20 | $N_{i0}$ | $p_{i0}$ |
| 1 | 189 | 0.30 | $N_{i1}$ | $p_{i1}$ |
| 2 | 252 | 0.40 | $N_{i2}$ | $p_{i2}$ |
| 3 | 63 | 0.10 | $N_{i3}$ | $p_{i3}$ |
| total | 630 | 1 | $N_i$ | 1 |

It is easily checked that the average score in this case is

$$\frac{126\times 0+189\times 1+252\times 2+63\times 3}{630}=1.4$$

As an index of difficulty this average is not very useful, because we have to remember that the maximum score for this item is 3. Therefore, the average score is divided by the maximum score (yielding a relative average score) of 1.4/3 = 0.467, i.e. 46.7% of the maximum score. The relative average score is (by definition) a number between zero and one. Notice that with binary items, average score and

relative average score coincide, because the maximum score is one. If the term *p*-value is used with partial credit items, it refers to the average relative score.

### C.5. Correlations between distractors and test score

To compute a correlation, one needs two series of scores. To compute the item test correlation, for example, one score is the test score, and the other score is the item score. The latter equals one if the answer is correct and zero if the answer is incorrect. The correlation is computed using the usual formula for a product-moment correlation (Pearson correlation). The computation will only fail if the observed *p*-value of the item is either zero or one, because in these cases the variance of the item score is zero.

To compute the correlation between a distractor and the test score, one must **recode** the answers given by the test takers. Suppose the item under study is a multiple choice item with four alternatives (A, B, C and D), alternative B being the correct answer: this means that an item score of one is given to every test taker who chose B, and a zero to the others. To compute the correlation between test score and distractor A, one has to create a new binary variable, giving a 'score' of one to every test taker who chose A, and zero to the others. The correlation looked for is the correlation between this new variable and the test score. To compute the correlation between test score and distractors C and D, one should proceed in a similar way. When using multiple choice items, it is good practice to compute the correlations between distractors and test score. In well constructed items, these correlations should be negative.

This application also illustrates the need of storing in some way the original observations. If one stores only the item scores (zeros and ones), it is not possible to compute the correlation between distractor and item score, because it is impossible to know which one of the distractors has been chosen from the mere knowledge that the answer was not correct.

### C.6. More on graphical item analysis: DIF

The discussion on graphical item analysis is a good opportunity to introduce a concept that has received a lot of attention in the last two decennia, the so-called Differential Item Functioning (DIF). The ideal of fair testing requires that an item 'behaves similarly' in distinct populations, for example in the populations of boys and girls. It is, however, not so easy to state what is meant or should be meant by 'similar behaviour'. One could claim, for example, that an item should be equally difficult in the populations of boys and girls, but using such a definition will cause serious trouble. It is a well established fact that at the age of 12, girls tend to be less proficient in arithmetic than boys. If the difficulty of the item is operationalised by its *p*-value, it is to be expected that the *p*-value of a typical arithmetic item will be lower in the girls' population than in the boys' population. This illustrates nicely the population dependence of the *p*-value. Usually this will hold for most or all items in an arithmetic test. But if we stick to the requirement that to be fair each item should be equally difficult in both populations, (and suppose an admissible test is required to have this property, and that only items with this property are included in the test), then by necessity we will find that on a 'fair' test, the average score of boys and girls is the same. But this approach implies that all differences are unfair, because it can be applied to any pair of populations, including the populations consisting of myself and my neighbour respectively.

So we need a more qualified definition of DIF, one that leaves room for differences between populations. Such a definition is formulated as a conditional statement. We apply it to the example of boys and girls. An item shows no DIF if in the (conceptual) population of boys with an arbitrary but fixed level of proficiency and the (conceptual) population of girls with the same level of proficiency, the *p*-values of the item are identical. Notice that this identity of the two *p*-values must hold at each level of proficiency. Stated more simply: absence of DIF means that the item should be equally difficult for boys and girls with the same level of proficiency.

In practice of course, we do not know the exact proficiency level of any test taker, but we can use the test score as a proxy. If, as before, test takers are grouped in a number of groups (of reasonable size), we can plot the observed *p*-values in each group for boys and girls separately. In Figure C.2, two examples are given from a mathematics examination. The legend refers to girls (Sg = 1; Sg stands for subgroup) and boys (Sg = 2).

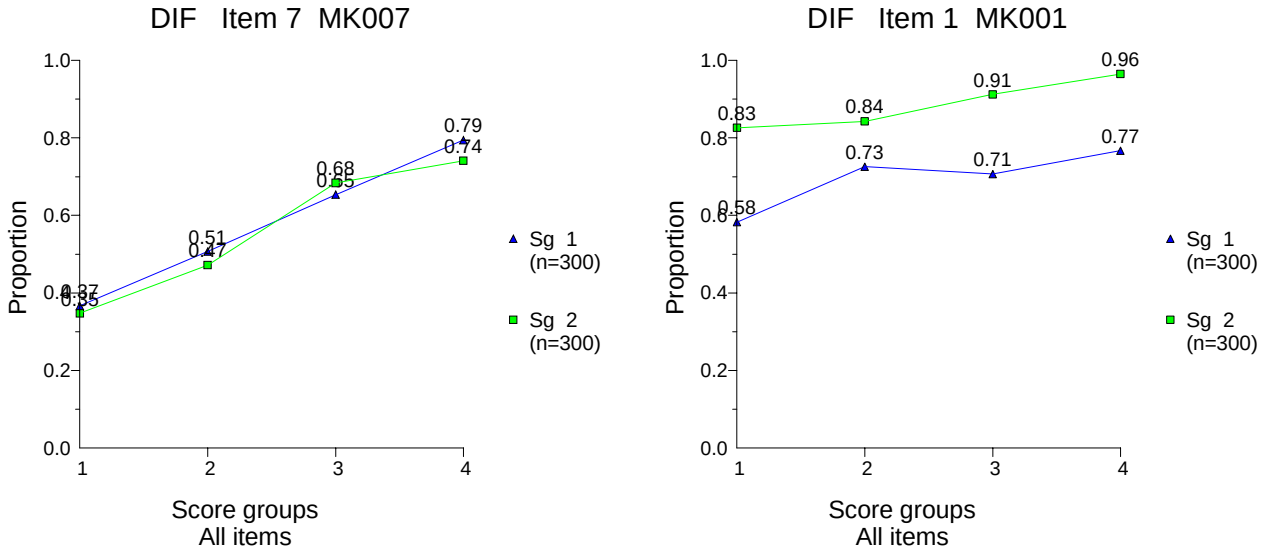



Figure C.2. Examples of DIF analysis

For item 7, there is no evidence of DIF: the *p*-values for boys and girls are very similar in each group (remember that these *p*-values contain an estimation error; so we cannot expect them to be identical in a sample). For item 1, on the other hand, there is clear evidence of DIF: the item is substantially harder in each girls' group than in the corresponding boys' group. Although there exist techniques for testing these differences statistically, in a clear-cut case as this, a plot is convincing enough. Scanning similar plots for all items in the test will reveal immediately important DIF as with item 1.

Although gender is commonly used as an example to explain and illustrate DIF, it is by no means the only variable where DIF can be investigated. In the United States of America cultural fairness of tests is often a strong requirement, and ethnical and racial background is often used as the contrasting variable in DIF-studies. In the general domain of achievement tests, an important variable to be used in DIF studies is the method of instruction used: it may be the case that some items turn out to be easier when the content matter of the test has been taught by method A, say, rather than by method B. A detailed DIF analysis may be revealing in such a context. Another highly relevant example is the use of mother tongue as the DIF-variable in case a test is administered to groups with different linguistic backgrounds, like the TOEFL.

### C.7. A graphical aid in constructing parallel forms

The construction of parallel forms can occur in different situations:

- A parallel form for an existing (and already used) test has to be constructed;
- Two (or even more) parallel forms are to be constructed from scratch;
- An existing test has to be split in two halves which are parallel (to use the split half method for estimating the reliability).

In all these cases a simple method can be used to construct the parallel forms in a graphical way. The idea is to construct two test forms which are approximately **strictly parallel**. This means that each item in one form has a twin in the other form with (approximately) the same psychometric qualities. In the framework of CTT one tries to have a match on two qualities: the difficulty and the discrimination, which are usually operationalised by the *p*-value and the item-test (or item-rest) correlation.

The starting point of the method is to construct a scatter diagram where each item is represented by a point in the plane. The *x*-coordinate is the *p*-value of the item, the *y*-coordinate the item-test correlation. The position of the item is symbolized by a (short) item label, such that items can easily be identified. An example is given in Figure C.3. Two items with graphical representation near each other have approximately the same *p*-value and the same discrimination. In Figure C.3 pairs are represented by lines connecting two item points. Pairs are formed such that the distance between the two item points in each pair is a small as possible.

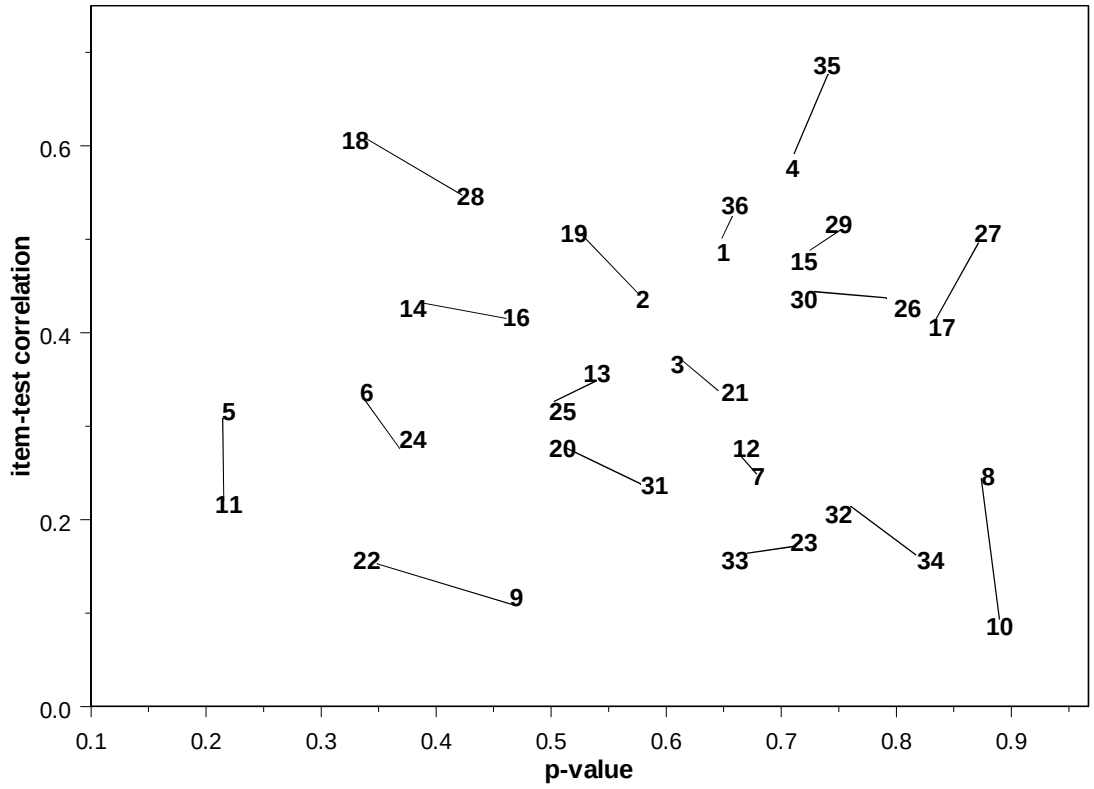



Figure C.3 Graphical construction of parallel forms

To construct the approximate parallel forms, the two items belonging to a pair should be assigned to the forms at random. There are a number of remarks to be made at this point:

1 If data are available on all the items from the same sample (and this will be the case when splitting an existing test in parallel halves, or in the construction of two parallel forms from scratch), it is always wise to check the extent to which the formation of parallel forms has been successful. In the two parallel forms the *p*-values of the items will not be different from their values when considering all items as belonging to a single test, but usually the item-test correlations will change.

2 If data are collected on two different samples (which may be the case if a new parallel form to an existing test has to be constructed), one should be very careful in using statistically equivalent samples. Both samples should be representative for the same target population.

3 If a parallel form for an existing test has to be constructed, it is wise to have more items to select from than what is strictly needed in the test. If the existing test consists of 35 items, it is advisable to have at least 50 items for the new test, such that 35 pairs can be formed, leaving 15 or more items unused. If one does not have such a provision, it may appear that it is not possible to construct a parallel form, because, for example, the new items are on average easier than the old ones.

4 The construction of the two parallel forms, as exemplified in Figure C.3 is done ‘by hand’, and it is not guaranteed that the proposed solution in the figure is the best possible. This is not a big problem, however: the aim is to construct two forms which are reasonably in balance with respect to the two psychometric qualities of the items. But it may appear that by proceeding in this way the two test forms show a quite strong unbalance in other respects, for example, with respect to

content. It is **not** the case that psychometric balance has priority to content. The ultimate decision is in the hands of the test constructor, and the method exemplified in Figure C.3 is only meant as a convenient tool in the construction of the parallel forms. One can extend control by very simple means, just as using a different colour of the item labels to distinguish between open ended and multiple choice items, or underlining and italicising to distinguish different content categories, and try to form pairs where content category, item format, *p*-value and discrimination are as similar as possible.

### C.8 The Spearman-Brown formula

There exists a powerful formula to control the test reliability, known as the Spearman-Brown formula. It says how the reliability changes as the test is lengthened (or shortened). Suppose a prototype of a test has been constructed which contains twenty items; this number of items is in some way considered as a standard length. So, we could say that it has the length of 1. The reliability of this test will be denoted by $\rho(1)$ for short. The Spearman-Brown formula can tell us what the reliability of the test would be if it contained forty items, that is, if it had the length of 2. And more generally, it tells us what the relation is between the reliabilities of a test of length 1 and a test of length $k$, where $k$ is an arbitrary positive number. Here is the formula:

$$\rho(k) = \frac{k\rho(1)}{1+(k-1)\rho(1)}$$

and here is an example. Suppose the test of 20 items has a reliability of 0.63, but the possibility exists to extend the test to 30 items, i.e. to make the test 1.5 times as long is it actually is. So, we have to apply the formula with $k = 1.5$ and $\rho(1) = 0.63$, yielding

$$\rho(1.5) = \frac{1.5\times 0.63}{1+(1.5-1)\times 0.63} = 0.719$$

The formula can be applied also to see the effect of shortening the test. Suppose we can apply only a test of 10 items instead of 20, then $k = 10/20 = 0.5$ and applying the formula gives

$$\rho(0.5) = \frac{0.5\times 0.63}{1+(0.5-1)\times 0.63} = 0.460$$

Some users do not understand fully the meaning of '*k*' in formula (10). It definitely does not denote the number of items; it denotes the ratio of a new number of items to some reference number, usually the number of items in an existing test. This latter number is then considered as the standard length (length of 1). The effect of test lengthening (or shortening) can be displayed graphically by a number of curves, as in Figure C.4.

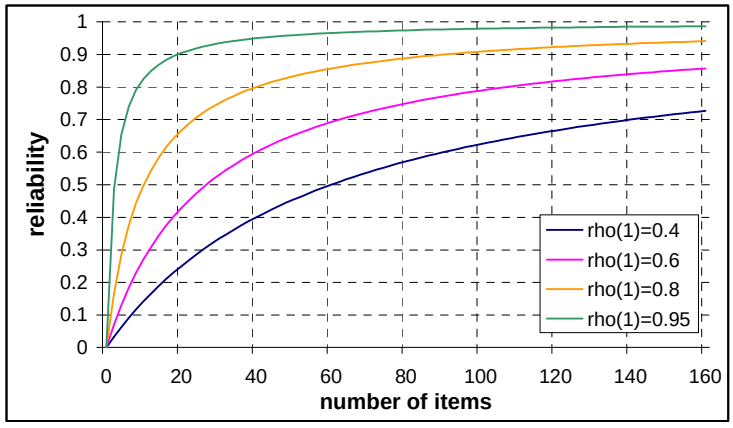



Figure C.4 Graphs of the Spearman-Brown formula

These graphs display a number of interesting characteristics:

1. All curves will eventually go to 1 if the number of items is large enough.
2. Of course, many more curves can be produced. The curves in Figure C.4 are just a few examples, and were produced with 40 items as standard length.
3. All curves have the same feature: starting with a small number of items, and then adding progressively more items, makes the curves grow rapidly at the start and more and more slowly as the number of items increases. A nice example is offered by the second curve from below. With 20 items, the reliability is (about) 0.40; adding 20 items causes an increase to 0.60, but adding another 20 items is not sufficient to reach a reliability of 0.70. Or, in short, adding items leads to a modest gain but removing items causes a great loss in reliability.

The Spearman-Brown formula is the most important practical tool to control the reliability of a test under construction. Sometimes a certain reliability is set as a minimum requirement for a test (in a certain population). One starts with the construction of the test, and the first analysis reveals that the target is not reached. Then one can use the Spearman-Brown formula to estimate the number of items that must be added to reach the target. Here is an example. Assume that the target reliability of a test is 0.85. Assume that a first analysis is done with a provisional test of 25 items, which yields an (estimated) reliability of 0.77. A very practical question then is to know how many items should be added to reach the target. If we take 25 items as the standard length, then it must hold (by applying the Spearman-Brown formula) that

$$0.85 = \frac{k \times 0.77}{1 + (k-1) \times 0.77}$$

and this equation (with *k* unknown) can be solved to find *k*:

$$k = \frac{0.85 \times (1 - 0.77)}{0.77 \times (1 - 0.85)} = 1.693$$

meaning that the test should have 1.693 times its present length, that is, contain $25 \times 1.693 = 42.3$ items. As fractions of items do not exist, this means that we will need at least 43 items to reach the target (42 will not be enough.). The preceding calculation leads to a very useful and practical formula:

$$k = \frac{\rho_{\text{target}}(1 - \rho_{\text{obs}})}{\rho_{\text{obs}}(1 - \rho_{\text{target}})}$$

where $\rho_{\text{obs}}$ is the reliability one actually has reached, and $\rho_{\text{target}}$ is the target reliability. (But again, remember that the result *k* of the formula is not the number of items, but the factor with which the actual number has to be multiplied.)

We will end this section with an example of the popular saying: the sting is in the tail. There is a big risk in applying the Spearman-Brown formula purely mechanically. The Spearman-Brown formula is only valid under quite strict conditions (which can not be discussed in detail in this appendix). Suppose one has to double the actual test length to reach the target reliability. If the provisional test contains 25 items that are constructed in a careful and professional way, one cannot hope to reach the target by adding 25 sloppy items, constructed in a hurry on a Sunday afternoon. More generally, one can express the requirement for the validity of the formula by saying that the test should be lengthened homogeneously. This means the added items should be very comparable (as a whole) to the items already present in many respects: the content coverage should be the same, the general level of difficulty and discrimination, perhaps also the format (a test consisting of 25 essay questions is not doubled homogeneously by adding 25 multiple choice questions.) All this of course cannot be controlled in full detail, and that is why the Spearman-Brown formula, beautiful as it is, will only yield approximations in practice.

## C.9 Confidence intervals for the true score

We need some mathematical notation to express the relation between the standard error of measurement and the reliability. The symbol $X$ will be used to represent the **observed** test score, and the reliability of $X$ will be symbolized as $\mathrm{Rel}(X)$. The standard deviation of the observed test scores is denoted as $\mathrm{SD}(X)$, and the standard error of measurement as $\mathrm{SE}(X)$. The relation between the standard error of measurement and reliability is given by the following formula:

$$\mathrm{SE}(X) = \mathrm{SD}(X)\sqrt{1-\mathrm{Rel}(X)}$$

The important fact about this formula is that we can compute the standard error of measurement from observable quantities: the standard deviation of the observed scores and the reliability. We use a well-known case as an example. In the use of intelligence tests, the scores (IQ) are expressed on a scale such that (in a well defined population) the mean IQ is 100 and the standard deviation is 15. Notice, that these quantities refer to observed scores, not to true scores, and that the reliability of many intelligence tests is well above 0.9, but certainly not equal to one. In Table C.3, the standard error of measurement is given for a number of cases.

Table C.3. Standard error of measurement with $\mathrm{SD}(X) = 15$

| Reliability | $\mathrm{SE}(X)$ |
|---|---|
| 0.85 | 5.81 |
| 0.88 | 5.20 |
| 0.91 | 4.50 |
| 0.94 | 3.67 |
| 0.97 | 2.60 |

These figures may come as a surprise, yet they are the result of a simple calculation. The table is important, as it should dissuade us from statements like “the reliability is as high as 0.97, which is virtually one” and then proceed as if it is really equal to one. Let us see what we can say about John’s IQ, if we have found that his observed IQ equals 112, and the reliability of the IQ-test is indeed as high as 0.97.

Since our measurement is not perfect, but contains a measurement error, the best we can hope is to define an interval that contains John’s real IQ (to be understood as his true score). But here a new problem crops up: Classical Test Theory does not say anything about the shape of John’s private error distribution. We cannot say that it is symmetric, and a fortiori we cannot be sure that it has the form of a normal distribution. Although it is possible in statistics to define confidence intervals without any additional assumption about the shape of the distribution, these intervals are usually disappointingly large. We can narrow these, but at the price of extra assumptions. Commonly, it is assumed that the error distribution is normal. If we buy this assumption, we can define a confidence interval in the usual way (see Section C.3), which as a mathematical expression looks like this:

$$\mathrm{Prob}(X_{\mathrm{John}} - 1.645\times\mathrm{SE}(X) \le \tau_{\mathrm{John}} \le X_{\mathrm{John}} + 1.645\times\mathrm{SE}(X)) = 0.90$$

or, in words, there is a probability of 90% that the constructed symmetric interval true score will contain the true score; the lower bound of the interval is the observed score minus 1.645 times the standard error of measurement and the upper bound is the observed score plus 1.645 times SD(E). Replacing the symbols by the numbers we know, we find

$$\mathrm{Prob}(112 - 1.645\times 2.6 \le \tau_{\mathrm{John}} \le 112 + 1.645\times 2.6) =$$
$$\mathrm{Prob}(107.7 \le \tau_{\mathrm{John}} \le 116.3) = 0.90$$

This means that the 90% confidence interval is 116.3 – 107.7 = 8.6 IQ points, which is more than half a standard deviation of the observed scores. Of course, we can apply a similar procedure not only to John but to an arbitrary member of the population. But if we do so, we have to remember that in 10% of the cases, the true score will lie outside the thus defined interval. So we see clearly that we cannot treat a reliability of 0.97 as being 'virtually one'.

**C.10 Important theoretical results**

The theoretical definition of reliability (see Section C.1) is the ratio of true score and observed score variance. This ratio cannot be computed in practice, because the true score variance is not known. If, we have a test which is parallel to a certain X (and which is commonly denoted as X'), then the reliability can be computed because it is theoretically shown that the correlation between two parallel tests equals the reliability of the test (and of its parallel form as well). There is, however, another important theoretical concept which is closely related to the reliability, namely, the correlation (in the target population) between observed and true scores. This relation is presented together with the earlier results in the following composite equation:

$$\mathrm{Rel}(X) = \frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} = \rho(X, X') = \rho^2(X, T)$$

Notice that the reliability is the squared correlation between observed and true score, and it follows immediately that

$$\rho(X, T) = \sqrt{\mathrm{Rel}(X)} \qquad \text{(C.1)}$$

This is an important theoretical result. One might wish to be able to measure without measurement error, but in language testing, as in many other areas, this is practically not possible, and all one can obtain is fallible results: the observed outcomes of a measurement procedure are in error. The above formula expresses directly the correlation between observed values and the theoretical construct of interest.
Since the reliability of a test is a number between zero and one, the correlation between observed and true score is larger than the reliability (it is equal only in case the reliability is zero or one). In Table C.4, some examples are displayed.

Table C.4. The relation between reliability and $\rho(X, T)$

| $\mathrm{Rel}(X)$ | $\rho(X, T)$ |
|---|---|
| 0.2 | 0.45 |
| 0.4 | 0.63 |
| 0.6 | 0.77 |
| 0.8 | 0.89 |
| 0.9 | 0.95 |

This relation has important implications for the discussion on validity. An important aspect of validity concerns the relation between the test scores and some other variable, which in many cases is also a test score. But both test scores are in error, and these measurement errors will tend to attenuate (i.e., lower) the correlation. Ideally one would like to know the correlation between the true scores on both tests. There exists a famous formula for this correlation, but we need some extension of the notation to write it down compactly. The two observed test scores will be denoted by X and Y and their corresponding true scores are denoted by $T_X$ and $T_Y$ respectively. The formula is:

$$\rho(T_X, T_Y) = \frac{\rho(X,Y)}{\sqrt{\text{Rel}(X)\ \text{Rel}(Y)}} \qquad \text{(C.2)}$$

or in words, the correlation between the true scores is the correlation between the observed scores divided by the square root of the product of the reliabilities. Since reliabilities are generally smaller than one, the denominator of the fraction will also be smaller than one, whence it follows that the correlation between true scores is larger than the correlations between observed values, or, as one usually says, the correlation between observed scores is attenuated by measurement error. The formula is also called 'the correction for attenuation'. (Notice that the formula does not apply when one or both reliabilities are zero, but in such a case the correlation between the true scores is also zero.)

This formula plays an important role in discussions about the construct validity of a test. If two tests measure the same concept, one usually finds that they correlate less than one, and this can be explained by the attenuation formula: the correlation is lowered by the fact that both test scores contain measurement error. But if X and Y really measure the same concept, then the correlation between their true scores should be equal to one, i.e., they should be **congeneric**. Replacing the left hand side of the attenuation formula by 1, we find immediately that

$$X \text{ and } Y \text{ are congeneric } \Leftrightarrow \rho(X,Y) = \sqrt{\text{Rel}(X)\text{Rel}(Y)}$$

i.e., if X and Y are congeneric then their correlation should be equal to the square root of the product of their reliabilities.

In practice, one cannot use formula (C.2) as its stands, because this formula refers to population values, and in practical situations one has to use sample estimates for the correlation and the two reliabilities, and because of the fraction in the formula, the result can be a number that is larger than one, which of course cannot be a correlation. The most notorious pitfall, however, with this formula is when one uses a lower bound to the reliabities, such as Cronbach's alpha. If tests are heterogeneous, this coefficient can be substantially lower than the reliability, and using these as estimates of the reliability in the formula, will make its denominator too small, and as a result the result of the fraction too high, giving in some cases results far exceeding one, or results near one, even if the two tests are not congeneric at all.

**SECTION D**

# QUALITATIVE ANALYSIS METHODS

**Jayanti Banerjee**
**Lancaster University**

Chapter 6 of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF) (henceforth referred to as 'the Manual'), explains that 'internal validation is a pre-requisite for acceptable linking to the CEF' (Council of Europe, 2003: 100). This chapter focuses on how intrinsic test quality might be established by answering questions such as:

i. Are the items really the level(s) they are supposed to be?
ii. Are the results awarded by different raters comparable?
iii. Are subtests supposedly testing different things providing different information?
iv. Are learners focussing on what is being tested or are they focussing on something quite different?
v. Do the interviewers elicit a good performance effectively?

[extracted from Council of Europe, 2003: 100 – 101]

This section of the Reference Supplement is intended to demonstrate how questions about test quality can be answered using qualitative analysis methods. Its content is as follows:

i. An overview of qualitative methods
ii. Verbal reports
iii. Diary Studies
iv. Discourse/conversation analysis
v. Analysis of test language
vi. Data collection frameworks
vii. Task characteristic frameworks
viii. Questionnaires
ix. Checklists
x. Interviews

Sub-sections ii – x have been grouped according to the nature of the data gathered. They will each follow a standard pattern: description of the qualitative method; examples of research using that method; and advice on how to use the method. Where possible, a key reference will be suggested for each method. A full list of references can be found at the end of the section.

Despite the focus of this section upon issues of test quality, I would like to suggest that many of the methods described here could also be used as part of standard-setting procedures. I will return to this in sub-section 6. However, it is important first to understand what each qualitative method entails and how it has been used already in language testing research.

## 1. Qualitative analysis methods

Qualitative approaches to test validation enable test developers and test users to look more closely at how a test is working by focussing on individuals or small groups. They can be distinguished from quantitative approaches in a number of ways. First, as has already been intimated, qualitative approaches focus on individuals or small groups rather than large test populations. Their aim is to gather detailed information about the specific experiences of these individuals or groups. If a quantitative method, such as a large-

scale survey, has revealed a trend, then a qualitative method can be used to explore that trend at the level of the individual – perhaps in order to explain it.

Second, qualitative approaches have been termed 'interactive and humanistic' (see Cresswell, 2003: 181). The involvement with the research participant is closer. This demands a great deal of sensitivity on the part of the researcher. In many cases, the research participants also contribute to the direction of the research.

Third, qualitative research is interpretive and tends to be cyclical and emergent. For instance, if a researcher wanted to explore test administration procedures (in order to check how test secrecy is maintained) they might design a questionnaire to be completed by everyone involved in the administration of the test (teachers, examiners, office staff). They might then decide to interview a selection of respondents in order to explore the answers to certain questionnaire items. Since, the researcher already had answers to the questionnaire, they might go into the interview with a very clear idea of the issues they wished to explore. However, during the interview, the researcher will need to respond to what the respondent says, interpret meaning and judge whether to (and how) to explore unexpected lines of enquiry.

Despite these distinguishing characteristics, however, it is important to view qualitative and quantitative (such as those described in the other sections of this reference supplement) analysis methods as complementary. Each will give you different information about the test that you are validating and will offer an illuminating perspective. Indeed, studies that use qualitative and quantitative methods in this way are increasingly common.

One recent example is a study by Brown (2003) that explored the effect of the interviewer on a test-taker's speaking proficiency. This research developed on an earlier study by Brown & Hill (1998), which used multifaceted Rasch analysis to derive measures of interviewer difficulty. Brown (2003) identified the easiest and most harsh interviewers from this study and selected a candidate that had been interviewed by both these interviewers. Brown & Hill (1998) had established that raters perceived this candidate to be more proficient when she was interviewed by the easy interviewer than when she was interviewed by the more difficult interviewer. Brown (2003) analysed the transcripts of both interviews using conversation analysis (see 3.1, below) in order to understand better the effect of the interviewer on the test-taker's speaking performance. As a result of this analysis, Brown concluded that the 'easy' interviewer provided more support to the test-taker during the speaking test. For instance, she was explicit about what she expected of the test-taker. She also provided feedback that indicated understanding and interest.

Brown (2003) also wished to explore whether the raters' views of this test-taker were affected by the interviewers' behaviour. Therefore, she gathered retrospective verbal reports (see 2.1, below) from 4 of the raters for each of the interviews. Her analysis of the verbal reports shows that the raters paid attention to whether or not the test-taker had produced extended discourse. They consistently judged that the test-taker produced extended discourse more readily with the 'easy' interviewer than with the 'difficult' one.

This combination of quantitative (multifaceted Rasch analysis) and qualitative (conversation analysis and verbal reports) methodology has established that interviewer style/behaviour can affect the speaking score a test-taker receives. It has also explored the features of interviewer style that are particularly influential on candidate performance. This study is useful in demonstrating the complementarity of qualitative and quantitative methodology. The remainder of this section will discuss various qualitative analysis methods, beginning with those that employ the technique of reflection.

## 2. Reflection

Qualitative analysis methods that employ the technique of reflection ask their informants to write or talk about their thought processes and/or actions when preparing for a test, taking test items, reading a test performance, or using a rating scale. Researchers can choose whether or not to be present during the reflection. If the researcher decided that it was not necessary to be present, then it would be more likely that a diary study (see 2.2, below) would be used. Even if they decided to be present, researchers could also decide how much they wish to probe (through interruption at various points or by a post-reflection interview) the informants' reflections. This sub-section will discuss two ways of gathering reflections on test-preparation, test-taking, and assessment processes: verbal reports and diary studies.

### 2.1 Verbal reports

Verbal reports are also referred to as 'verbal protocols'. They are data collected from test takers and/or examiners in which they talk about their thought processes while they take a test or assess a test performance. Verbal reports have been defined in many different ways but the most helpful is probably Green (1998). She defines verbal reports along three parameters:

i. The type of data collected – informants could be asked to speak only their thoughts aloud (a talk aloud) or to also provide other information that is not already in verbal form such as physical movement (a think aloud)
ii. The time lag between the thought or action and the verbalisation – concurrent verbal report or retrospective
iii. The nature of the intervention (if any) – the researcher might ask for explanations of utterances or prompt for more information (mediated) or may remain silent, allowing the informant to report unprompted (non-mediated).

Verbal reports are very useful sources of data about test takers' and/or examiners' processes when taking or assessing tests. However, they are very demanding for informants to provide because you have to perform the test-taking or assessment task and simultaneously talk about what you are doing and thinking. This presents a tremendous cognitive load. It is important, therefore, to train informants in giving verbal reports. The training should be a two-stage process and should be conducted separately with each informant:

**Stage One**

1. Explain what a verbal report is and what is involved.
2. Demonstrate a verbal report. Show the informant an example of a verbal report either by doing one yourself or by playing a video-recording of someone doing a verbal report.

**Stage Two**

Give the informant an opportunity to practice providing verbal reports. Two tasks should be provided, both similar to the tasks that the informant will have to perform for the real data collection. For instance, if you wish to collect verbal report data about a reading test, select two or three items from an equivalent version of the test to use as practice material.

1. Give your informant the first item to complete as a verbal report. Give your informant detailed guidance about the verbal report that is required. If necessary, interrupt during this task to prompt your informant for more information and to make explicit what you would like them to report on.
2. After they have completed the first verbal report practice task, give the informant further feedback (e.g. explain where you would have liked more detail).
3. Then give them the second task. Allow the informant to perform this verbal report under the conditions that you will use for your study.
4. Give the informant more feedback. It is good to tell your informant what you particularly liked about the verbal report they provided. Also explain where you would have liked more detail.

It is important to note that (despite training) some informants are better at giving detailed verbal reports than others. Alderson (1990) investigated the reading comprehension skills used by test-takers when

completing a 10-item academic English reading comprehension test. He conducted verbal reports with two test-takers. Each session lasted approximately one hour and was recorded for transcription and analysis. Alderson (1990) found that one test-taker had considerable difficulty expressing his thoughts while the other seemed to be much more able. He concluded that it is important to identify good informants. I would recommend that you use the training procedure to identify informants who will be comfortable providing a verbal report and who will give you useful data. As Alderson points out, "in qualitative research of this kind, it is more important to identify good informants than to find representative informants" (1990: 468).

It is also important to consider what language the verbal report should be given in. The choice of language is not necessarily straightforward. You and the test-takers might be first language (L1) speakers of Language A but the test might be in Language B. Should you ask the test-takers to give their verbal reports in your shared L1 (Language A) or in the language of the test (Language B)? In some circumstances, you might find that your test-takers speak Language A but you are an L1 speaker of the language of the test, Language B. In this case, should you request that the test-takers use Language B even though it is not their L1? You might like to consider the following issues:

1. Will informants be able to express their thoughts more fully and accurately if they provide verbal reports in their L1 (regardless of the language of the test)?
2. Will it add to the cognitive load experienced by informants if they are taking the test items in one language and providing verbal reports in another language?
3. What would the informants prefer? The test-takers in the Alderson (1990) study both used the language of the test (English) for their verbal reports. When Alderson observed that one test-taker was having great difficulty expressing his thoughts, he encouraged the informant to use his L1. The informant refused because he wished to improve his English (1990: 467).

Once you have identified skilled informants (who can provide verbal reports) and have decided what language you would like to collect the data in, you will need to decide whether or not you would like to collect your data concurrently or retrospectively. Concurrent data has the advantage that you capture the thoughts as they occur, in so far as it is possible to capture instantaneous information about thoughts. However, it is not always easy to collect concurrent data. This can be because of the nature of the task. For instance, it would be very difficult to ask a test-taker to provide a verbal report while they were taking a speaking test. It would prove very difficult to distinguish between the test performance and the verbal report.

The context in which the data is being collected is also important. For instance, it would be difficult to collect concurrent verbal reports during the live administration of a test. The verbal report process might influence the test-taker's performance and this would be unfair if his/her performance were to contribute to an official score.

However, a retrospective report has the disadvantage that the informants' memory of their thoughts during the test-taking or assessment process might be incomplete or inaccurate. Even if the verbal report were collected immediately after the test or rating, informants might forget details of their behaviour. In such circumstances it might be useful to employ 'stimulated recall methodology' (Gass & Mackey, 2000). This is a variation on more traditional retrospective reports because it provides some support for the informant during the recall. This support can take the form of an audio-tape or video recording of the test-taker (taken while they were performing the task) or it can be a copy of their test performance e.g. the written product of an essay task. Gass & Mackey explain that concrete reminders like this will prompt informants to remember the mental processes that occurred during the original activity (2000: 17).

One possible way of using stimulated recall methodology would be through the following two-stage process:

Stage One

The informants view the recording/read their written performance and report on their thoughts at the time that they were taking the test. They should be allowed to stop and/or rewind the tape if they wish.

Stage Two

In the case of an audio or a video-recording, the researcher can play the recording, stopping the tape at various points to probe for further details about the thoughts of the informant at that point in the test. In the case of a written performance, the researcher might want to draw the informant's attention to specific aspects of the text (perhaps certain lexical choices) and probe for further details about how/why the informant made those choices.

It is important to note that stimulated recall methodology need not necessarily be used in conjunction with verbal reports. Gass & Mackey (2000) describe how it might be used in the form of a questionnaire or in a diary study (see sections 5.1 and 2.2 respectively). The key thing to remember is that stimulated recall methodology can be used to support informants when you ask them to provide you with details of their behaviour during tests, their reactions to tests and/or test performances, and their behaviour during the assessment process.

Verbal reports (whether or not in conjunction with stimulated recall methodology) have been used primarily in the areas of reading and writing (both test-taking and assessment). Cohen (1984) used verbal reports to explore the match between the test-taking processes of examinees taking a reading test and the predictions of the test designers. Cohen reported a number of different studies with different groups of students taking different tests. The number of students in each study varied between 22 and 57 and the tests varied in length and composition. Some tests comprised 10 multiple-choice items (based on a single reading passage) while others combined more than one task type (e.g. multiple-choice, short answer questions, cloze passage). The verbal reports in the different studies revealed interesting information about the students' test-taking strategies as well as their test-taking processes. For instance, Cohen reported that students taking the cloze test tended to ignore the test's instruction to read the entire passage before completing any of the blanks (1984: 74).

Alderson's (1990) study also examined test-taker processes in a reading test but had a slightly different aim. He was responding to arguments that reading skills were separable and could be ranked as higher order or lower order. He gathered verbal reports from 2 students. With one student, Alderson gathered a concurrent verbal report. The student voiced his thought processes while he was taking the test. The second student completed the test first. Alderson conducted a retrospective verbal report with this student. Alderson (1990) found that the students did not necessarily use the micro-skills predicted by experts when responding to particular items. His analysis also revealed that it was possible for different test-takers to get an item correct but to arrive at that correct response by different processes. He further found that it was difficult to identify a body of low-order and high-order skills. As a result of this investigation, he questioned whether test developers could state with any confidence what an item in a test was testing.

In the area of writing assessment, verbal report methodology has primarily been used to investigate assessment processes though it has also, as in the case of Cohen (1994), been used to explore how test-takers perform a particular writing task. Cohen's (1994) study encompassed both phases of testing – the test-taking process and the assessment process because he was interested in how summarising tasks work as a testing format. So, he explored the strategies that test-takers use when they have to write a summary and as well as the strategies that assessors use when rating such tasks. His respondents were 5 students (who completed the summary task) and 2 assessors.

Cohen's study was conducted as follows (1994: 177 – 178):

**Test-taker verbal reports**

1. The test-takers were given a two-part test to complete. They were asked to provide verbal reports of their thoughts and their actions while they were taking the test. They were also asked to comment on the input texts they were reading and to describe any difficulties they had in performing the tasks.
2. A researcher observed the test-takers during the test-taking process. She took notes of what the test-takers did while completing the test (all observable strategies) and also intervened when she felt that the test-taker had not reported on an action or had been silent for some time.
3. When they had completed the test, the test-takers were given a questionnaire. This asked them to comment on whether their English course had helped them to perform the summary tasks, their opinion of this test format, their reactions to the presence (and interventions) of the researcher, as well as whether any difficulties they experienced with the summary tasks were due to reading problems or writing problems.
4. All these stages in the study were conducted in the test-takers' L1 (Portuguese).

**Assessor verbal reports**

1. The assessors were asked to provide verbal reports of their thoughts and their actions during the rating process. They were asked to comment on: the way they determined the topic of the input texts, the stages in their rating process, and also to give their views on how well the test-takers had understood the input texts.
2. A researcher was present during the rating process and noted any observable strategies that the raters used.
3. When they had completed the assessment exercise, the raters were given a questionnaire. This asked them to comment on the summary tasks in relation to previous tests of summarising that they had encountered. It also asked the assessors to point out if they had found any aspect of the test difficult to rate and to comment on the test format, the input texts and the scoring procedures.
4. All these stages in the study were conducted in the assessors' L1 (English)

Cohen's (1994) analysis of the resulting data revealed that assessors varied in the criteria that they applied to the summary tasks as well as in the rating procedures they adopted. Cohen concluded that improvements could be made to the reliability of the marking by establishing clear marking procedures and by developing a scoring key (content) for each task. Cohen also found that the test-takers would benefit from training in this task type. Nevertheless, he concluded (1994: 202) that the summary task type was very useful for 'reactivating what [the students] had learnt in their EAP courses'.

Weigle's (1994) research looked at the effect of rater-training on rating processes. Her respondents were 16 raters working on an English as a second language (ESL) placement test of which half were experienced (having been assessors for this test in previous years) and half were inexperienced/new raters. Weigle's study had three main stages (1994: 203 – 204):

**'PRE'**

1. The raters provided background information during an initial interview.
2. They were then given the placement test marking criteria and asked to rate 13 scripts.
3. Following this rating task, the raters were trained in giving verbal reports.
4. The raters practised the verbal report methodology with four scripts for which the scores were known (taken from a previous administration of the test).
5. Finally, the raters were given 13 more scripts (these differed in topic from the scripts assessed as part of step 2) to rate silently.

**'NORM'**

1. Each rater received a 'norming packet' before this stage of the study. The packet contained 10 representative sample compositions that had previously been rated. Each sample had an official score for each subscale on the marking criteria.
2. The raters were required to mark the compositions before attending the norming session and to compare their marks to the officially assigned marks.
3. During the norming session, the raters discussed their scores in order to understand the rationale behind the official score.
4. Each rater was interviewed immediately after the norming session. They were asked for their reactions to the norming session and to comment on what they had learned. They were also asked to discuss the compositions where their judgements had diverged from the official scores.

**'POST'**

1. After the norming session the raters participated in live rating of the placement test. Two weeks after the end of the live rating the raters attended a second interview. At this interview they were first re-trained in verbal reports.
2. They were then given six scripts to mark while practising the verbal report methodology. Four of these scripts were the same scripts they had marked during the 'PRE' stage (step 4, above).
3. After they had completed the verbal reports, the raters were asked to indicate whether or not they had read each essay before. Where they recognised an essay, they were asked if they could remember the scores they had given previously.

All the data-collection sessions (including the norming sessions) were video-recorded. The transcripts of the verbal reports took special note of pauses, false starts and repetitions. Weigle analysed the verbal reports for the four inexperienced raters whose ratings varied the most between the 'PRE' and 'POST' ratings. She found that the rater-training had had two important effects on the ratings that these raters gave. First, they understood the rating criteria better as a result of training. Second, they became more realistic in their expectations of the student performances at each level of ability.

Finally, Lumley (2002) investigated how assessors negotiate their understanding of the rating scale and the test script to arrive at a judgement of the test performance they are rating. The test in question was a high-stakes test that (at the time of data collection) was used as part of the Australian immigration process. Lumley (2002) focussed on 4 experienced assessors, all of whom were accredited raters for this test. His study followed a five-step process (2002: 253):

1. Re-orientation to the rating process (using four practice scripts)
2. Simple rating (no verbal report) (12 scripts of two tasks each)
3. Practice verbal report rating (one practice script)
4. Data collection phase of rating plus concurrent verbal reports (12 scripts of two tasks each)
5. Post-rating interview

You can see from this structure that Lumley employed a well-developed training framework for his assessors, both to re-orient them to the rating process and to familiarise them with the verbal report methodology. Lumley's analysis of the resulting verbal reports revealed the complex relationship between the rater, the writing performance and the rating scale. He was able to identify criteria that the raters used in their judgements but which were not reflected in the rating scale (in this case a criteria relating to the content of the writing – the quantity of ideas) (2002: 263 – 265). He was also able to illustrate how raters negotiate the effect of a test-taker's writing with the criteria in the rating scale, some of which might not be stated explicitly (2002: 265 – 266).

It is more rare to find examples of verbal report methodology in the areas of speaking and listening. This does not mean, however, that such research is not possible. One example is Buck (1994), who used verbal reports for a listening test. At the time that Buck conducted this study he had been unable to find any

published studies using verbal reports with listening comprehension (1994: 153). He therefore conducted a number of pilot sessions in order to explore how best to use verbal report methodology in this context. He conducted the main study with 6 students, all speakers of Japanese. His procedure was as follows:

1. The test-takers took a 54-item test based on a single listening text. The text was divided into 13 short sections that were played to the test-takers one at a time. The items were all short-answer questions. These were divided between the 13 sections. All the questions were in the students' L1 (Japanese) but the students were free to write their responses in either Japanese or English (the language of the test).
2. Each test-taker attended a post-test interview. During this interview, they took the test items again (using the same procedure as adopted during the first administration). But, before they proceeded to each subsequent section, Buck (1994: 154) asked them a number of questions to check how well they had understood the input text and the questions as well as to explore the test-takers' listening and test-taking strategies.

The interviews were conducted in the students' L1 (Japanese) and each lasted approximately two hours. As a result of his analysis of these interviews, Buck concluded that "top-down processes are crucial in listening comprehension" (1994: 163). He also found that listening comprehension was affected by non-linguistic factors such as interest in the subject matter. Listeners make predictions and inferences while listening based on what they have already understood and their background knowledge. Finally, he identified a number of factors that interact to affect student's performance on individual test items.

It is clear from the preceding discussion that verbal reports can offer insights into test quality in a number of ways. These include:

1. The match between test-designers' predictions and the actual skills and processes test-takers use during the test.
2. The role of test-taking strategies in the successful completion of certain task types.
3. The distribution of micro-skills across a test (in order to establish test coverage).
4. An examination of aspects of a particular task-type in order to establish its usefulness in achieving the aims of the assessment.
5. An exploration of what assessors pay attention to and why in order to better understand the effect of these variables on the score that the test-takers receive.
6. The effect of training on what assessors pay attention to and its consequences for inter and intra-rater reliability
7. The effect of the rating scale and rater expertise on what assessors pay attention to.

Though not discussed here, verbal reports could also be used to explore whether and how students' writing processes differ in test and non-test conditions or in different test conditions (such as between paper-based and computer-based tests).

The studies reported here also illustrate some keys points:

1. There is no optimum sample size in a verbal report study. Some studies have involved as few as two respondents while others have involved 50 or more. You will need to judge how many respondents you need in order to be confident that you have captured a healthy range of possible behaviours. However, it is common to have sample sizes of 10 or less.
2. Verbal reports can be gathered for a variety of task-types but you need to bear in mind the length of the data collection session. With the exception of Buck (1994), the sessions reported have been up to 1 hour long. Beyond this you might find that exhaustion sets in and the quality of the verbal report diminishes. If you find that you need to take more time, you might wish to consider breaking the verbal report process into two parts so that you can give your informants a rest period.

3. It is not possible to predict all the directions that the verbal report will take. However, you can increase your preparedness by piloting your methodology.
4. It is usually helpful to combine verbal reports with another type of data collection methodology such as a questionnaire or observation. This will help you to triangulate the information that you gather (i.e. complement it with a view of the same events from another perspective). As a result you may be able to explain more easily what respondents report and/or you might more easily follow up on gaps in the verbal reports.

Despite the potential of this methodology, researchers will inevitably encounter a number of challenges. The first is choosing the context in which to gather the verbal report data. Cohen (1984: 78) argues that you are more likely to capture actual test-taking processes if you gather verbal report data in circumstances when the test result will be official. However, he notes that this places you in something of a 'Catch-22' situation because the students might not be willing to be completely honest. They might worry that a true report of their test-taking processes could adversely affect their mark. Also, as I pointed out earlier, the verbal report process might interfere with the test-taking process and this could also negatively influence the test-takers performance.

The second challenge is ensuring that the verbal reports are sufficiently detailed for profitable analysis. Cohen (1984: 78) points out that verbal reports cannot necessarily capture the level of detail that you might wish for. He gives the example of a multiple-choice item. He explains that, in order to understand fully how one option was selected, you might want the examinee to explain how they eliminated/rejected the alternatives. Yet, despite this attention to detail, Cohen argues that it might not be possible to capture all the processes that occurred in the selection of the answer. Part of this problem, as Alderson (personal communication) suggests, is due to the fact that some processes are simply not accessible to verbal report, perhaps because they occur so quickly and are so automatic that the informant is not aware of them.

Alderson (1990: 477 – 478) also explains that the interviewer might not be aware during the interview of all the areas in the test-taking process that should be probed. As a result, he/she might fail to adequately probe in certain areas at the time and would only realise the gaps during the analysis. He believes that this is due to the reactive nature of the methodology. It is not possible to predict in advance (and therefore be fully prepared for) what will emerge during the verbal report. He suggests that researchers should plan to go back to their informants as soon as possible with follow-up questions and requests for clarification and/or confirmation of interpretations.

One final challenge is that of making sense of the data collected. Buck (1994: 155) points out that the information is often scattered through a number of hours of recordings and it is difficult to decide how best to summarize and present the data in a meaningful form. His solution was to organise his discussion around his initial hypotheses. Cohen adopted a taxonomy developed by Sarig (1987, cited in Cohen, 1994: 179). Unfortunately, there is no single solution to this problem. The approach adopted by one researcher might not be applicable to data gathered in a different context and for a different purpose. As a result each researcher has to find his/her own 'path' through the data collected. Since this conundrum applies to virtually all the methods described in this section of the reference supplement, I will return to it in section 7.5, where I offer some approaches to analysing rich verbal data.

## 2.2 Diary studies

In general, diary studies offer a way of collecting data relatively unobtrusively but regularly. Diary keeping is a familiar activity, even for people who do not keep diaries of their personal lives. It allows researchers to capture people's thoughts and experiences before they can be forgotten or lose their immediacy and significance. However, diaries can vary widely in format. The most familiar format is unstructured, a blank page on which the informant is asked to write everything relating to the area being researched. For instance, a study of how learners prepare for a test and what they focus on might simply

give informants the instruction to write about their daily test-preparation activities. The simplicity of the instruction can result in very interesting and widely varying responses. However, the drawback of providing such an open-ended task is that informants will self-select the information they believe interesting and important. They might provide less data. Alternatively, you might find that the data is extremely varied with the result that if you use an unstructured format with large numbers of respondents, you could find the resulting data very difficult to analyse. It will not have a pre-determined structure and you will have to establish this structure post-hoc.

Symon (1998) argues that most diary studies give their informants more guidance. Some studies can be very structured. They provide informants with diary forms to complete with a combination of closed and open-ended questions (see 5.1 for more discussion of these terms). Respondents have a very clear idea of what they need to include in their diaries and little or no space for including information that has not been explicitly asked for. Taking, once again, the example of a study of how test-takers prepare for a test, a very structured diary entry might list different test preparation activities as a pro-forma. Respondents might then be asked to complete this pro-forma at regular intervals, each time simply ticking the activities they engaged in during the period covered by the pro-forma.

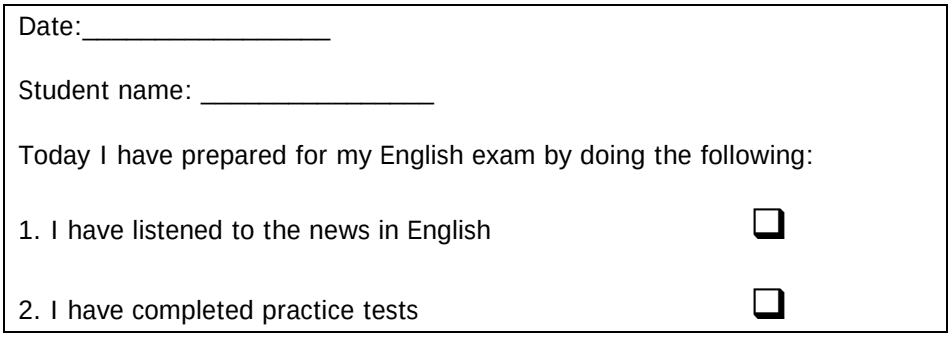



**Figure 1: Excerpt from a structured diary pro-forma**

This approach to diary studies makes analysis very easy because the pro-forma is so structured. It is, therefore, a very good way of using diaries with large numbers of respondents. The problem with providing such strict guidance, however, is that you will only get the information that you ask for. Unless you have been able to successfully predict what your respondents will tell you, a very structured diary form could result in your missing interesting information.

One solution to this is to adopt the middle ground between no guidance and very strict guidance. For instance, if the diary study is of learner strategies when preparing for an examination, it might be possible to give your respondents some examples of test preparation activities that learners might engage in. You could then ask your informants to indicate whether or not they engaged in any of those activities that day or week. You would ask your informants to describe anything else they have done in order to prepare for the examination. You might also ask them to reflect upon how useful they found each of the activities they engaged in. In fact, in order to ensure that your respondents are prompted to provide this additional information (indeed, to check that they are taking the diary study seriously), you should not include the most common test preparation strategies on your initial list. You would expect many of the respondents to add these strategies into their diaries.

As the discussion so far has shown, when deciding on how structured your guidance should be, it is important to think carefully about the purpose of the diary study, the number of respondents you wish to

include in your study and the use that will be made of the data. It is also important to consider a number of other questions:

i. Is the diary study the best way to gather the data? Diary studies provide in-depth, longitudinal data and it is important to decide whether this is appropriate for the research question.
ii. Who is going to complete the diary? Some informants might need more guidance than others – depending on their age and/or their educational level.
iii. What language will the diary be completed in? As in the case of verbal reports, the answer to this question is not always obvious. One consideration might be whether you would like the diary to perform two functions, a research tool for you but also a pedagogic (language learning) tool for your respondents. If you do decide that your respondents should use the diary as a language learning tool, you might wish them to complete it in the target language rather than in their L1.
iv. How often should the diary be completed and for how long? It is particularly important to judge the best time to collect the diaries and this is most successful if the researcher stays in good contact with the informants.
v. How often should you monitor the progress of the diary? Symon (1998: 101) reports that informants are most likely to abandon their diary during the first week of diary-keeping. It is important, therefore, to have frequent contact during that week and then, perhaps, to taper off. However, it is important that contact should be regular.

Though diary studies have not been widely used in published language test validation research, the most common context of use is likely to be learner diaries. Test-takers can be asked to report on their language learning experiences and difficulties post-test. The data collected can be compared to the test score each test-taker was awarded and could provide information about the language abilities of test-takers at different score levels. Other contexts in which diary studies might be useful are examiner/assessor diaries. These could record how markers interpret rating scales and how they apply them to test performances. Diary studies could also be used to explore the behaviour of interlocutors in speaking tests.

## 3. Analysis of samples

Reflections such as verbal reports and diary studies are data that are gathered either after or during test-taking or rating. The next type of qualitative analysis method does not generally involve gathering additional data from test-takers or assessors. Instead, the language of the test becomes the focus of the analysis. In the case of discourse analysis and conversation analysis (see 3.1, below), the test discourse is scrutinised for its social and interactional features. Alternatively, the language of the test can be analysed for features such as grammatical complexity or lexical density (see 3.2, below) perhaps in order to explore whether different tasks tap into different aspects of a test-taker's language resources.

### 3.1 Discourse/conversation analysis

Discourse analysis and Conversation analysis are distinguished from one another in two ways:

1. Discourse analysis is concerned with issues such as power relations and gender inequalities whereas Conversation analysis is more concerned with the extent to which interactions conform to expected patterns.
2. Discourse analysis can be performed on transcripts of conversation or on interviews. It could even be applied to documents (such as test manuals or specifications, perhaps). As Silverman (2001: 178) comments, Discourse analysis is far more 'catholic' about the data it admits. Conversation analysis, however, focuses on transcripts of spoken interaction ('talk').

I will deal with each separately, beginning with Conversation analysis.

Conversation analysis (henceforth CA) is primarily used in the analysis of data from speaking tests. It has three basic assumptions (Heritage, 1984: 241 – 244):

i. Talk has a stable and predictable pattern. The structure of talk can be treated as a 'social fact'.
ii. Each speaker's contribution can only be understood in relation to the context i.e. the preceding sequence of talk. In other words, each utterance inevitably builds on previous utterances and cannot be analysed in isolation from them.
iii. Transcripts must be extremely detailed in order to capture every relevant aspect of speaker meaning because all inference/claims must be grounded in evidence from the data.

CA, therefore, is essentially the analysis of talk in interaction. Hutchby & Wooffitt (1998) provide an excellent introduction to the method. Other good resources are ten Have (1999), Silverman (2001) and Lazaraton (2002). The latter is particularly interesting because it focuses on the use of CA in the validation of speaking tests.

Transcription is a key feature of CA because the transcript must capture as accurately as possible the interaction between the speakers. Hutchby & Wooffitt (1998: 86 – 87) demonstrate the importance of the transcript by presenting two transcriptions of the same conversation. In the first the script simply records what was said, in the order it was said by the two speakers. In the second script, the researcher has indicated where turns overlap and the length of pauses. He/she has also noted other features such as intonation, in-breath, out-breath and emphasis. This transcript shows much more clearly the interaction between the two speakers. It is this transcript that is more helpful in CA. Indeed, because transcripts must be a vivid record of the original interaction, the field has a well-developed glossary of transcription symbols. These can be found in full in Hutchby & Wooffitt (1998: vi – vii). Some of the symbols used are demonstrated in the following example:

```
R: well .hhh let's start with the (0.5) well the MBAs=
I: =yes that sounds fine
R: (1) .hhh Emmanuel=
I: =Emmanuel↑=
R: =yes (.) did the four week course with you:: (0.5)
I: (.) I mean he [was]
R:               [yes] (1) came with first class degree from M ((erased for
confidentiality))=
I: =first class?
R: (1) yes (.) with some experiential learning before that ((reading from
student file)) with business experience before that. (.) this is somebody
who the MBA office asked to do an essay because the experience wasn't so
great (.) they often make sure that the student is understanding .hh is
going to understand what the course is about (.) then they ask them to do
an essay (0.5) and apparently this was a very (3) um (3) this was o.k.::
((laughs))=
I: =right (2) so it wasn't outstanding↑
```

**Figure 2: Example of CA transcription symbols**

| | |
|---|---|
| **(0.5)** | The number in brackets indicates a time gap in tenths of a second |
| **(.)** | A dot enclosed in a bracket indicates a pause in the talk of less than two-tenths of a second |
| = | The 'equals' sign indicates 'latching' between utterances i.e. one utterance follows immediately after the previous one with no break/pause |
| **[ ]** | Square brackets between adjacent lines of speech indicate the onset and end of a spate of overlapping talk |
| **.hh** | A dot before an 'h' indicates speaker in-breath. The more h's the longer the breath |
| **(( ))** | A description enclosed in a double-bracket indicates a non-verbal activity. Alternatively, double brackets may enclose the transcriber's comments |
| **:** | Colons indicate that the speaker has stretched the preceding sound or letter. The more colons the greater the extent of the stretching. |
| **?** | Indicates a rising inflection. It does not necessarily indicate a question. |
| **↓↑** | Pointed arrows indicate a marked falling or rising intonational shift. They are placed immediately before the onset of the shift. |
| Under | Underlined fragments indicate speaker emphasis |

(all taken from Hutchby & Wooffitt, 1998: vi – vii)

The unit of analysis typically is the 'adjacency pair'. An adjacency pair consists of two utterances occurring together that are spoken by two different speakers and function as complementary parts of an exchange. For example:

```
R: well .hhh let's start with the (0.5) well the MBAs=
I: =yes that sounds fine
```

Some common adjacency pairs are:

question – answer
greeting – greeting
invitation – acceptance (refusal)
compliment – acceptance
request – compliance
offer – acceptance (refusal)
complaint – apology

You can see, that the example (above) shows an 'offer-acceptance' adjacency pair. It is important to note, however, that the two parts of an adjacency pair may not be found immediately next to one another. For example:

```
I: =and then what do you do with that book?=
S: =you mean the notebook[?
I:                        [((murmurs agreement))
S:                                  [whenever I have time I just
   take it out and read them.
```

In this example the two sentences in bold are an adjacency pair that is separated by what is called an insertion sequence (another adjacency pair).

CA assumes that these paired (and adjacent utterances) follow certain patterns and rules of interaction. The focus of the analysis is usually on:

1. The structure of the adjacency pair - does the data follow expected patterns such as the ones listed above. How do speakers negotiate breakdowns in the adjacency pairs?)
2. Turn-taking - how speakers negotiate when and for how long they will each speak. This too is believed to be rule governed. In particular, if there is a breakdown in communication or a miscommunication, turn-taking can be inspected and explanations sought.
3. Topic organisation and repair - Test data can be analysed to see who introduces and controls topics and initiates repair as well as the nature of the topic organisation and repair.

I: =which is a fail?=
R: =no (.) actually (.) you're ok .hhh as long as you get over 40% for each module and an average of 50% overall (.) you're ok! he didn't actually fail anything (0.5) I don't remember him doing any re-sits (2) I don't think (.) I didn't keep the breakdown any more than that (2) .hhh so I think he got through every thing in some way (2) but (.) just (.) just overall ((student name)) just seemed to (.) well was quite considerably higher (2) and particularly in the exams as well ((student name)) seemed [to
**I: [57%=**
R: =57% as compared to the 46%↑=
I: =yes (.) but **exams were clearly a problem for both** (1) **so do you think exams place a greater strain on the students' language ability?**=
R: =oh absolutely↑ I think so (.) I think anyone who's doing an exam in a second language (0.5) I mean it's bad enough doing it in your own language but (1) um (1) yes I think (.) you know (.) trying to sort of write under such pressure and such a short time scale and remember everything and be translating it in your head all the time (.) yes. I do.

In this example (above) the excerpts presented in bold type are initiations of topic change. Both topic changes were initiated by 'I', the interviewer. The remainder of the interview could be analysed to establish the extent to which the interviewer initiated the topics that were discussed as well at the extent to which this indicated that the interviewer was in overall control of the interview.

CA has been used by a number of researchers interested in analysing the language of speaking tests. Lazaraton (2002) discusses the use of conversation analysis to analyse test language in the Cambridge EFL examinations. This volume is part of the Studies in Language Testing series published by Cambridge University Press and the University of Cambridge Local Examinations Syndicate. It focuses primarily on CA and includes a number of chapters that explain this analytical approach in detail. In the final chapter, Lazaraton (2002) describes how CA can be used to analyse interviewer behaviour in a speaking test. She presents two studies that were part of the validation programme for the now unavailable Cambridge Assessment of Spoken English (CASE). The data comprised transcripts of test performances for 58 language school students (24 males and 34 females, all Japanese L1 speakers). The performances had been elicited by a pool of 10 examiners. The transcripts were a full record of the elicitations and the student responses. Lazaraton's (2002: 126 – 139) reports the results of these studies, showing how she analysed the transcripts for: the interlocutors' use of the interlocutor frame (which was intended to standardise the input each test-taker received) and also to examine specific aspects of interlocutor behaviour. Her analysis showed that the interlocutors varied widely in their use of the interlocutor frame, using the prompts 40% - 100% of the time. It was also important to note that the same interlocutor would use a different number of prompts in each interview. One interlocutor used between 54% and 77% of the prompts in 6 interviews.

The analysis of specific interlocutor behaviour showed that one interlocutor in particular provided test-takers with supportive behaviour such as:

1. supplying vocabulary
2. rephrasing questions
3. evaluating responses (e.g. 'sounds interesting')
4. repeating and/or correcting responses
5. stating questions that require only confirmation
6. drawing conclusions for candidates

Some interlocutors also used strategies such as 'topic priming' where they first asked a closed question such as 'Do you like to go dancing?' before developing on this with a more open question such as 'What sorts of dancing do you like?'. This too was considered supportive behaviour because it prepared the test-

taker for the upcoming interview question. Supportive behaviour of this kind had a significant effect on the test-takers' performances in one part of the test.

Brown (2003) has also examined the influence of the interviewer upon test-taker performance. She looked in detail at one candidate, who had been interviewed by two different interviewers (in an experimental design). She selected this candidate (Esther) because her scores for the two interviews were markedly different. Indeed, for one interview she was judged as far less able than for the other. Brown (2003) analysed the transcripts of both interviews. She found that one interviewer (Pam) developed on Esther's responses and indicated an interest in what she said, prompting her to elaborate her answers. Pam also used topic primers such as those identified by Lazaraton (2002). Brown (2003) also notes that Pam would close topics consistently i.e. signalling clearly to the test-taker (Esther) that she was about to change to another topic.

Brown's (2003) analysis of the other interviewer (Ian) however, revealed that his behaviour was qualitatively different. Esther had not performed as well when interviewed by Ian as she had when interviewed by Pam. Brown's (2003: 11 – 16) analysis revealed that Ian tended to ask closed questions to which Esther gave short, unelaborated responses. Ian's topic shifts were also more abrupt and did not display the topic priming found in Pam's elicitations. As a consequence, Esther's performance was far less assured. She spoke very little and tended to speak only in short sentences. Brown (2003) argues that the interviewers' behaviour had a clear but unpredictable effect on the test-taker's performance. She concludes that it is very important to examine interviewer behaviour for its possible threat to test validity.

It is clear from the research described above that CA can be used to analyse test language in order to:

1. check the extent to which the test is measuring the desired competences.
2. explore whether test-taker performance is being affected by construct irrelevant factors such as interviewer behaviour.

Like CA, Discourse analysis (henceforth DA) can focus on test performances and need not require the collection of additional data. However, while CA focuses on talk (and therefore is useful in the analysis of the language of speaking tests), DA can also be used to analyse other forms of verbal data such as post-test interviews and test documents e.g. test manuals/handbooks. The other key difference between these two approaches (as mentioned earlier) is the scope of analysis. Whereas CA is primarily concerned with how talk conforms to expected patterns of interaction, DA helps researchers to explore issues such as power relations and gender inequalities. It is defined as the analysis of "texts and talk as social practices" (Potter, 1997: 146) so the analysis focuses on how people use language to 'do' things such as to construct a particular identity or to have a particular effect on their listener. Good introductions to how DA might be performed are provided in Potter & Wetherall (1987), Potter (1996) and Potter (1997) but the use of DA in language testing is best illustrated by examples of research such as Brown & Lumley (1997) Kormos (1999) and O'Loughlin (2002)

Brown & Lumley (1997) studied test-taker performances on the Occupational English Test (OET), a test taken by medical professionals hoping to gain accreditation to practise in Australia. This test consists of two role-plays in which the interlocutor performs the role of a patient or relative of a patient. The test-taker plays their role as the medical professional. The purpose of these role-plays was to simulate, as far as possible, the real situations in which medical professionals need to communicate in order to assess how well the test-takers could cope with these situations. It was important, however, that each test-taker received a comparable level of challenge during the role-plays. Interlocutor variability in the role-plays could undermine the validity of the speaking test.

Consequently, Brown & Lumley's (1997) study explored the behaviour of the interlocutor and its effect on the test-taker's performance (and test score). They analysed test transcripts, paying particular attention

to what the interlocutor said (as part of their role) and the responses they received. The features of interviewer behaviour that appeared to make the test harder were: sarcasm, interruption, repetition (an unwillingness to accept the test-taker's answer to a question), and unco-operativeness. The features of interviewer behaviour that appeared to make the test easier were: the asking of factual questions, linguistic simplification (in the form of repetition of key information, reformulation of key information, slowing of speech etc), and allowing the candidate to initiate topics and to control the interaction.

Brown & Lumley (1997) contended that interlocutors varied in their behaviour depending on the identity they constructed for themselves. An interlocutor who identified with their role as a patient was more likely to produce challenging behaviour whereas an interlocutor who identified more with the test-taker was more likely to produce supportive behaviour. Test-takers who encountered an interlocutor who was more challenging because he/she used sarcasm or was unco-operative had a more difficult test than those who encountered an interlocutor who was generally more supportive. Brown & Lumley (1997) argued that all test-takers should encounter the same level of challenge. In saying this they reminded their readers that this did not preclude the inclusion of some challenging behaviour (for instance, sarcasm) if the construct of the test demanded it. But they contended that if the ability to cope with patient sarcasm should be included as part of the test construct then all the test-takers should receive that challenge.

Kormos (1999) used discourse analysis to examine the effect on the language of the test of different test tasks. She gathered speaking test performances from 30 candidates (10 male and 20 female, all Hungarian L1 speakers). The speaking tests were all conducted by four examiners. Each speaking test comprised three tasks: a general non-scripted interview, a guided role-play, and a picture-description task (1999: 168). Kormos focused particularly on the two interactive tasks – the interview and the role-play. She was interested in exploring the power and dominance relations between the test-taker and the interlocutor in each of these tasks. In order to do this Kormos looked particularly at topic control (topic initiation, ratification and closing) but also looked at how the participants in the speaking test gained the floor (perhaps through interruptions) and retained it. Her analysis revealed a strikingly different pattern of relations between the interview and the role-play. During the interview part of the test, the examiner was dominant. He/she largely had control over the topic (its initiation and closing). The test-taker rejected topics in only 1% of the cases. However, during the role-play task, the test-takers exercised far more control. They initiated 50% more topics than the examiners. During this part of the test, both parties (the test-takers and the examiners) ratified each others' topic initiations 97% of the time. On the basis of this analysis, Kormos (1999) argued that the role-play tasks were a better measure of test-takers' conversational competence because such tasks distributed power more evenly between the candidates and the examiners.

O'Loughlin (2002) was interested in the role of gender on the test-taker's performance and score. His study explored whether there was a gender effect during the interview (in terms of the nature of the interaction between the interlocutor and the candidate) and also during the rating process. He collected test performances from 16 test-takers (8 male and 8 female), each of whom took an International English Language Testing System (IELTS) test twice – once with a female interlocutor and once with a male interlocutor. In the IELTS test, the interlocutor is also the assessor. In addition to the ratings provided by the interlocutor-assessors, O'Loughlin (2002) gathered further ratings of all the test performances from four other assessors (2 male and 2 female). He performed a Rasch analysis of the test scores and a discourse analysis of the test performances. His DA of the test performances focused on three aspects of spoken interaction: overlaps, interruptions and minimal responses. These were chosen because previous research had indicated that these features of spoken interaction were "highly gendered" (O'Loughlin, 2002: 175). O'Loughlin found, however, that there was no clearly gendered pattern in the use of any of the three features he analysed. He conceded that he might have found patterns of gendered language use had he included other features of language in his analysis.

Looking back over these three examples, it is important to note that the research reported here exemplifies the use of DA to analyse speaking tests. This is perhaps the most common use of DA in investigations of test quality. Nevertheless, it is still possible to use DA to analyse other test products such as test manuals or the texts used for reading and listening input.

When using DA to analyse oral language, two important points should be noted. The first is that DA makes use of many of the analytical concepts of CA. For instance, the analysis often focuses on adjacency pairs, turn-taking and topic organisation and repair. Kormos (1999) looked at patterns of topic initiation and uptake while O'Loughlin (2002) looked particularly at how speakers took and held the floor (overlaps, interruptions and minimal responses). The difference, however, is in the perspective taken on the data. In both these cases, the researchers were interested in effect of an aspect of the context or the test-taker upon the patterns of interaction. So Kormos was interested in the effect of the task-type upon the distribution of power in the test discourse and O'Loughlin explored differences in speaker discourse by gender of the test-taker.

The second point to be noted is that, as with CA, DA analyses transcripts of spoken interaction. But, unlike CA transcripts, DA transcripts need not include precise notations of intakes of breath or of each non-verbal contribution (for instance, particles such as 'mm' and 'uh huh'). Instead, they are more likely to use a sub-set of the transcription annotations described above. Particular attention is paid to pauses, para-linguistic behaviour (such as hand movements or the shrugging of shoulders), overlapping speech and emphasis.

It is clear from all the examples provided in this section that Conversation analysis and Discourse analysis have typically been used to analyse spoken test discourse. They can offer insights into speaking test quality in the following ways:

1. The effect of interlocutor behaviour upon the test-taker's performance.
2. An exploration of the influence of test-taker characteristics (such as gender) upon test performance
3. The effect of task-type upon the test-taker's performance.
4. A comparison between test and non-test language in order to establish the extent to which the test has captured relevant aspects of the test-taker's language ability.

The size of the data sample in the research reported here has varied. Brown (2003) focused on just one test-taker and two interlocutors (selecting this from a larger pool of data). Kormos (1999) analysed the performances of 30 test-takers (and four interviewers) each performing two different tasks. O'Loughlin's (2002) dataset comprised 32 performances from 16 test-takers. You will need to judge how much data you will need in order to be confident about the claims you make but it appears that most researchers gather 30 – 60 performances, depending on the depth and focus of their analysis.

Since the language sample is central to CA and DA, the quality of that sample is important. Recording equipment must be in good working condition so that the recording is clear. The transcription stage is also crucial. A lot of useful detail can be lost if the transcription fails to capture it but you can also waste time and resources if you include more information in your transcript than you eventually use. In the case of CA, there is a well-defined transcription system. DA transcriptions can be more flexible (and less detailed) but, because it is not always possible to tell from the outset what aspects of the data will be salient, I would recommend that you perform one practice transcription and analysis in order to identify the precise level of detail that you need to go into in your transcription. Also, be prepared to modify this detail as your analysis proceeds. This means that you will always need to have the original data recordings at hand so that you can refer to them easily should you need to add detail to the transcript or perhaps simply confirm a particular interpretation of the transcript.

Finally, as the example of O'Loughlin (2002) demonstrates, though it is important to be guided by the literature when selecting features to analyse, it is also important to be data-driven i.e. to look for patterns in the data and seek to explain them.

### 3.2 Analysis of test language

The Conversational analysis and Discourse analysis approaches to analysing language samples focus on the social and interactional features of test language. It is also possible to analyse a test-taker's language output (spoken or written) and/or the test input (e.g. a reading text) for a range of linguistic features. This can be useful for a number of reasons. For instance, Kim's (2004: 31) analysis of cross-sectional data from a group of learners indicates that more proficient learners use more subordinate clauses and more phrases in their writing output. This indicates that better-performing students produce grammatically more complex writing and suggests that an analysis of test-taker output might help us to understand better the language features that distinguish one level of performance from another.

Turning to test input, Laufer & Sim (1985) interviewed students in their L1 about their comprehension of L2 academic reading texts. They found that the students needed vocabulary most in order to understand the texts they were reading. Kelly (1991) presents a similar finding in a study of listening comprehension. In this study, advanced language learners in Belgium were asked to transcribe and translate excerpts from British radio broadcasts. The resulting transcriptions and translations were analysed for their errors and Kelly reports that more than 60% of the errors were lexical in nature (i.e. where the meaning of the word had not been understood). These studies indicate that it might be useful to analyse the language of test input in order to better understand sources of test-taker difficulty and to perhaps better estimate the appropriacy of an input text for a particular level of ability – a measure of 'listenability' or 'readability'.

The range of linguistic features that might be investigated include:

1. lexical richness
2. rhetorical structure/functions
3. genre
4. discourse markers
5. grammatical complexity
6. register
7. accuracy

To do this you would first need to identify appropriate measures of the language feature you would like to analyse. This is more complex than it might at first seem. For instance, Read (2001) describes the different considerations involved in measuring lexical richness. It is important to understand how a 'word' is defined. The first key distinction is between 'function' or 'grammatical' words such as *and*, *a*, *to*, and *this* (articles, prepositions, pronouns, conjunctions, auxiliaries etc) and 'content' words such as nouns, verbs, adjectives and adverbs. Taking the age old example:

> The **quick brown fox jumped** over the **lazy dog**

The words highlighted in 'bold' are the content words. The remainder are the function/grammatical words. The other key distinction is that between 'types' and 'tokens'. In vocabulary research, a 'token' is, quite simply, a word used in a text. Therefore, the number of tokens in a text is equal to the number of words in that text. A 'type' however, is a more selective measure. It takes into account only the number of different word forms used in a text. In other words, if a word form is used more than once (e.g. 'the') it will only be counted the first time it is used. More selective again is the term 'lemma'. This is used only in relation to 'content' words and is a super-ordinate term used to describe a base word and all its inflections e.g. *play*, *plays*, *played*, *playing* or *test*, *tests*, *test's*, *tests'*. A 'word family' is a related concept and refers to words that share a common meaning. Read (2001: 19) provides an example:

leak, leaks, leaking, leaked, leaky, leakage, leaker

He explains that even though some of these words have a more metaphorical meaning than others, they are all closely related. Read (2001: 19) does warn, however, that some word families are not as easy to define. For instance the words *socialist* and *socialite* may originate from the same underlying form 'soci-' but they are so distinct in their meaning that they should probably be classed in different word families.

Estimations of lexical richness further involve the calculation of:

1. lexical variation – the variety of different words used, or what might be described as the 'range of expression' (Read, 2001: 200). This is usually measured by calculating the type-token ratio i.e. the number of different words in the text divided by the total number of words in the text. It is important to note here that, because this is a measure of **lexical** variation, researchers focus their *type* measures on 'content' words only rather than also counting 'grammatical'/'function' words such as articles or prepositions.
2. lexical sophistication – the use of low-frequency words such as technical terms or other uncommon words. This is calculated by dividing the number of sophisticated (low frequency) word families in the text by the total number of word families in the text. When calculating this measure, it is usually important to compare the words used to a list of words that the test-takers might be expected to know e.g. by looking at an official vocabulary list for a particular ability level.
3. lexical density – this involves a comparison between the number of grammatical words and the number of content words and is usually calculated by dividing the total number of content (lexical) words by the total number of words in the text.
4. number of lexical errors – this involves counting the number of errors. These errors can take different forms e.g. choosing the wrong word to express a particular meaning, the use of the wrong form of the word, and the stylistically inappropriate use of a word (for instance a very informal word in a formal piece of writing).

All these calculations seem relatively straightforward, but Read (2001: 201) cautions that the results are premised on a number of key decisions. These include, as has already been mentioned (above), decisions about how words might be classified into word families. Other decisions involve deciding whether a word is a content word or a grammatical one and whether multi-word items (such as idioms or phrasal verbs) should be counted as single units. An example is provided from the Slovenian Primary School Leaving Exam (Alderson & Pižorn, 2004: 156) to demonstrate the decisions that need to be made.

Read the text and find out if the statements below the text are true (T), false (F), or not given in the text (NG). Circle the right answer. The example has been done for you.

**AMAZING TIGERS**

Tigers often have to hunt day and night to get enough to eat. You may think that a tiger can easily bring down any animal it goes after. But that's just not true. In fact, most of the time, the tiger's prey gets away. The great cat succeeds just once in 15 to 20 tries. That's why it sometimes doesn't eat for weeks.

A tiger's body is packed with muscles. So it can leap the distance of two cars parked one in front of another. Despite its huge muscled body, a tiger moves very gracefully through the forest. Its claws are mostly hidden in its paws. It glides on its soft, padded feet.

Like other cats, tigers clean their fur with their rough tongues. A tiger's tail is about half the length of its body. Tigers "talk" to other tigers with tails. An upright tail, shaking slowly back and forth, says "Hello". A lowered tail, moving quickly from side to side says "Better be careful".A tail straight back and moving quickly from side to side says "What's happening? I'm excited."

Tigers mark their home ranges with their scent and urine. These markings act as a special kind of communication between tigers. A female scent also lets the males know when she is ready to mate. A tiger roars as a warning to other tigers to stay away.

| | | | | |
|---|---|---|---|---|
| 0 | A tiger can easily catch any animal. | T | (F) | NG |
| 1 | Tigers are good hunters. | T | F | NG |
| 2 | A tiger's tail is the same length as its body. | T | F | NG |
| 3 | Tigers are dangerous to people. | T | F | NG |
| 4 | Tigers communicate with their tails. | T | F | NG |
| 5 | A tiger's tail can show a tiger's excitement. | T | F | NG |
| 6 | Males never know when to approach females. | T | F | NG |
| 7 | Females are not as strong as males. | T | F | NG |

**Figure 3: Example Task No. 4/25 – English**
**(Slovenian Primary School Leaving Exam)**
**Extracted from Alderson & Pizorn (2004: 156)**

Consider the following phrases in the text: *day and night*, *goes after*, *back and forth*, *in fact*, *most of the time*. Would you consider all of these phrases to be multi-word items (which should be counted as single units) or do you think that one or more should be counted as separate words? Similarly, what would you do with the contractions in the text (*that's*, *doesn't*)? Are these single units or are they two separate words? Read (2001: 201) makes clear that there are no 'wrong' answers. It is more important to be meticulous in your recording of the decisions you take and to spend time at the beginning of the analysis setting up the rules that you intend to follow. Read (2001: 201) further suggests the use of corpus analysis tools such as a concordance (perhaps WordSmith). This will list all the words in the text and how frequently they are used. It is also possible to compare the words used in the text with a larger corpus such as the British National Corpus (BNC - http://www.natcorp.ox.ac.uk/). Doing so will reveal the words that might be considered low frequency in relation to the large corpus. To do this you will need to use a corpus analysis tool such as WordSmith (http://www.oup.com/elt/global/isbn/6890/). If you do not have easy access to a corpus of spoken and written language nor to a tool such as WordSmith, you might find it helpful to refer to Leech et al. (2001). This volume presents frequency lists based on an analysis of the BNC. It presents rank-ordered and alphabetical frequency lists for the whole corpus and for various subdivisions (e.g. informative vs. imaginative writing, conversational vs. other varieties of speech). Words are presented according to their grammatical use. For instance, 'round' may be used as a preposition or as an adjective. These two uses of the word 'round' are presented separately.

Even when decisions have been made about how to classify words and phrases it is important to note that other issues might need to be addressed. The first is that lexical variation (the type-token ratio of the lexical words in the text) is affected by the length of the text; it tends to drop as texts get longer. This is particularly problematic when analysing test-taker writing output since some test-takers will inevitably write more than others. Researchers have approached this problem differently. Laufer (1991) decided to take the first 250 words of the scripts that she analysed whereas Arnaud (1984) randomly selected 180 words from test-taker scripts for his analysis.

The second issue that should be addressed is how errors might be treated (quite apart from measuring them as suggested above). For instance, when calculating the lexical variation of a test-taker's writing output, do you wish to take account of all the words that the test-taker has written or only the ones he/she has used correctly? It is also sometimes difficult to decide whether an error is a vocabulary error or a grammatical one. Additionally, it is important to bear in mind that if every error carries the same weight this might skew the results that you get. Therefore, should you ignore minor errors (such as spelling) or should you count every error?

As the foregoing discussion of just one feature has demonstrated, the analysis of test language is a serious undertaking and its exploration requires much preparatory work in order to take defensible decisions. Indeed, the questions that inevitably arise are who is the judge and who has the right to be the judge? Certainly, one way to ensure that your decisions are defensible is to have your categories confirmed by an independent observer (i.e. perform a reliability check) but it is clear that this further lengthens an already complex process. This suggests that it might not be feasible to include analyses of test language as part of your routine checks of test quality. However, as the following descriptions of research will demonstrate, it would certainly be useful if you have a specific question about your test.

O'Loughlin (1995) investigated the comparability of test-taker output in two versions (face-to-face and tape-mediated) of a speaking test. He analysed data gathered from performances on the *Australian Assessment of Communicative English Skills* (henceforth referred to by its acronym - *access:*) comparing the lexical density of the performances on each version. An earlier study by Shohamy (1994) had shown that the language in face-to-face speaking tests (OPIs) tends to contain a higher percentage of grammatical/function words (60% grammatical and 40% lexical words) than the language in tape-mediated speaking tests (SOPIs). This suggests that test-taker output in a SOPI tends to be more 'literate' whereas test-taker output in an OPI tends to be more 'oral'. It further suggests that OPIs and SOPIs do not tap the same underlying construct of speaking. This is of some concern to test developers since they want to ensure that all versions of a test have the same underlying construct. O'Loughlin's (1995) study probed Shohamy's (1994) conclusions by considering the effect of task type on lexical density. The *access:* test was well suited to this exploration because the face-to-face and tape-mediated versions had been developed in parallel and incorporated the same task types.

O'Loughlin's (1995) first step was to develop a comprehensive framework for analysing the test-taker performances (see figure 4, below). Note that O'Loughlin (1995) decided that the verbs 'to be' and 'to have' plus all modals and auxiliaries should count as grammatical items whereas other verbs should be classed as lexical items. Note also his decision to count all contractions as two items (particularly since this was an analysis of speaking output).

O'Loughlin (1995) developed this framework after careful consideration of his data set of 20 speaking performances from 10 test-takers who each took both forms of the *access:* test. He examined this data for the effect on lexical density of both test format (face-to-face or tape-mediated) and task type. To do this, O'Loughlin (1995) focused on four tasks that were roughly parallel in both version of the test – a description, narration, discussion and a role-play. Each task was analysed separately for lexical density. O'Loughlin (1995) was also concerned that his results might differ depending upon the relative frequency of the lexical items used. Therefore, he calculated lexical density using two methods. In the first, he weighted all the lexical items equally regardless of their frequency. In the second, he gave all the high-frequency items half the weighting of the low frequency items.

*A. Grammatical items*
Verbs 'to be' and 'to have'. All **modals** and **auxiliaries**
All **determiners** including articles, demonstrative and possessive adjectives, quantifiers (e.g., some, any) and numerals (cardinal and ordinal).
All **proforms** including pronouns (e.g., she, they, it, someone, something), proverbs (e.g., A: Are you coming with us? B: Yes I *am*), proclauses (e.g., this, that when used to replace whole clauses).
**Interrogative** adverbs (e.g., *what, when, how*) and **negative adverbs** (e.g., *not, never*).
All **contractions**. These were counted as two items (e.g., *they're* = they are) since not all NESB speakers regularly or consistently use contractions.
All **prepositions** and **conjunctions**.
All **dicourse markers** including conjunctions (e.g., *and, but, so*), sequencers (e.g., *next, finally*), particles (e.g., *oh, well*), lexicalised clauses (e.g., *now, then*), spatial deities (e.g., *here, there*) and quantifier phrases (e.g., *anyway, anyhow, whatever*).
All **lexical filled pauses** (e.g., *well, I mean, so*).
All **interjections** (e.g., *gosh, really, oh*).
All **reactive tokens** (e.g., *yes, no, OK, right, mm*).

*B. High-frequency lexical items*
Very common lexical items as per the list of the 700 most frequently used words in English (accounting for 75% of English text) identified in the COBUILD dictionary) project. This list is included in the *Collins COBUILD English course, level 1, student's book*) Willis and Willis, 1988: 111 – 12). It includes **nouns** (e.g., *thing, people*), **adjectives** (e.g., *good, right*), **verbs** (e.g., *do, make, get*), **adverbs of time**, **manner and place** (e.g., *soon, late, very, so maybe, also, too, here, there*). Not items consisting of more than one word are included in this category as the COBUILD list consists of words not items.
**Repetition of low-frequency lexical items** (see below) including alternative word forms of the same item (e.g., *student/study*).

*C. Low-frequency lexical items*
Lexical items not featuring in the list of 700 most frequently used English words cited above including less commonly used **nouns, adjectives, verbs** including participle and infinitive forms (all multiword and phrasal verbs count as one item). Adverbs of **time, place and manner** and all **idioms** (also counted as one item).

**Figure 4: Lexical density – classification of items**
**Taken from O'Loughlin (1995: 228)**

The analyses resulted in data sets comprising percentages of the amount of lexical words/items in the test-takers' output in comparison to the grammatical words/items. Since each test-taker had taken both versions of the test, this meant that there were 8 measures of lexical density for each test-taker. O'Loughlin (1995) reported that the method of calculating the lexical density of test-taker output provided only slightly different results but he argued that the weighted approach was probably more accurate. He also reported that the lexical density of the performances was generally higher for the tape-mediated test. For both test versions, lexical density was lower for the narration task than for the description and discussions tasks. The role-play appeared to be most affected by the test format. In the tape-mediated version, the lexical density was similar to the description and discussion tasks but in the face-to-face version it was lower than all the other tasks analysed. O'Loughlin (1995) concluded that differences between the OPI and the SOPI are more dependent upon the relative interactiveness of the tasks that test-takers are required to perform than upon the test format itself.

Apart from examining the lexical density of two speaking test formats (OPI and SOPI), Shohamy (1994) also conducted a number of other analyses. She first analysed the ideational functions (e.g. describing, elaborating, complaining) of the tasks in the two test formats. She found that the SOPI generally required more functions than all the versions of the OPI analysed i.e. those for low, middle and high level test-takers. Shohamy (1994) then analysed the topics covered by the different versions. She found that low-level test-takers taking the OPI tended to be tested in a narrower range of topics and also on fewer topics. She argued that these results indicated that the OPI implicitly assumed that higher level test-takers were

more able to discuss serious issues. She further argued that because the SOPI presented the same tasks and topics regardless of the level of the test-taker, it gave the test-takers equal opportunities to show what they could do.

Shohamy (1994) then analysed 20 test-taker performances. She calculated the number of errors per performance in relation to the number of words produced, looking particularly at certain error types such as word order, tenses, verb structure and gender. She found that this did not differ significantly between the two test formats. Shohamy (1994) then compared, for each performance, the communicative strategies of shift of topic, hesitation, self-correction, paraphrasing, and switch to L1. She and two independent assessors counted the frequency of occurrence of each of these strategies and then calculated the means for each test performance. The results indicated that paraphrasing was used significantly more frequently in the SOPI. Self-correction also tended to be used more frequently in the SOPI whereas switch to L1 was used more frequently in the OPI.

Shohamy's (1994) final set of analyses compared a number of discourse features of the test-taker performances in each test version. These were:

1. lexical density
2. rhetorical structure of the two test formats
3. genre
4. speech moves e.g. expansion, reporting, description, negotiation for meaning
5. communicative properties e.g. dialogue or monologue, smooth or sharp topic shifts
6. discourse strategies e.g. turn-taking, hesitation, silence
7. content/topics (n.b. this applied the same analyses as had been conducted on the test tasks)
8. prosodic/paralinguistic features e.g. intonation, laughter, hesitations, silence
9. speech functions (n.b. this applied the same analyses as had been conducted on the test tasks)
10. discourse markers e.g. connectors
11. register e.g. level of formality

As a result of this comprehensive analysis, Shohamy (1994) concluded that the SOPI is characterised by concise language that is very similar to a monologue. It is lexically more dense than the OPI and is also more formal. She suggested further that, despite their potential to elicit more functions (as indicated by the analysis of the tasks), the test-taker performances showed that SOPI tasks were more likely to elicit only narrative, reporting and description whereas the OPI had the potential to elicit a wider variety of speech functions. Finally, she argued that the test format could influence the type of language elicited from test-takers.

Wigglesworth (1997) also analysed the language test-takers produced during a tape-mediated speaking test in order to explore the effect of planning time on test-taker output. She was particularly interested in this because the provision of planning time can add considerably to the length of the test. It would also affect the underlying construct of the test. For instance, the question would need to be addressed of whether planning time makes the test more or less authentic. It is therefore important to establish whether such a change to the test is justified by the language that is elicited. Taking a 6-part tape-mediated test, Wigglesworth's (1997) methodology was as follows:

1. She prepared two versions of the test. For both versions two parts (parts 2b and 4) were presented with planning time. For version A planning time was also provided for sections 2a and 3 whereas for version B of the test planning time was provided for sections 2c and 5.
2. She then collected test-performances from 107 test-takers, divided roughly equally between the two test versions.
3. After the test performances had been rated, Wigglesworth (1997) selected a sub-set of 28 performances on each test version dividing these into high and low proficiency candidates.

Once the selected performances had been transcribed, Wigglesworth (1997) divided the texts into clauses. She did this because the dataset was very large and this focus on the clause helped her with the analysis. Wigglesworth (1997) subsequently analysed the texts for:

1. complexity (defined in this case as the number of subordinate clauses used per task)
2. accuracy (i.e. the use of bound morphemes (plural *s*), verbal accuracy, the distribution of definite and indefinite articles)
3. fluency (a type-token analysis was used to measure the number of words used in relation to the number of words used in conjunction with false starts, repetitions and hesitations. The number of clauses containing self-repair was also calculated).

As a result of these analyses, Wigglesworth (1997) reported that high proficiency learners benefited from planning time when performing more difficult tasks. Low proficiency learners did not benefit from planning time on these tasks. She also said that planning time is less beneficial to either group of test-takers when the task is easy, suggesting that this might be because the cognitive load on the students is not heavy in such cases. Her tentative conclusions were that it might be justifiable to provide planning time for complex tasks but not to do so when the tasks were relatively straightforward.

The remaining two examples of research show how analyses of test language can be used to achieve insights into writing test performances. The first, by Ginther & Grant (1997) considered the effects of test-taker ability level and language background and the topic of the task upon the written output. Ginther & Grant analysed 180 exam scripts from the Test of Written English (TWE). Each of these essays had already been rated by two independent assessors using the TWE scale of 1 to 6 where 6 is the highest possible score. The selected scripts had all been given a score of 3, 4 or 5 on the scale (there were insufficient numbers of scripts at the other levels to allow sampling) and represented test-takers with three different L1 backgrounds (Arabic, Chinese and Spanish). Half of the group had written on topic 1 and the other half had written on topic 2.

The essay scripts were then tagged by two independent judges (to allow for a reliability check) for parts-of-speech and for errors. The parts-of-speech coding followed the categories presented in figure 5.

| | |
|---|---|
| Definite article | BE |
| Indefinite article | BE able to |
| Demonstrative adjective | BE going to |
| Adjective | Verb |
| Count noun | Infinitive |
| Noncount noun | Phrasal verb |
| Possessive noun | Preposition |
| Gerund | Multi-word preposition |
| Pronoun | Conjunction |
| Possessive pronoun | Subordinate 1: complement |
| Adverb | Subordinate 2: relative pronoun |
| Multi-word adverb | Subordinate 3: conditional |
| Conjunctive adverb | Subordinate 4: adverbial subordinator |
| Negation | Subordinate 5: present participal subordinator |
| Auxiliary (do/have) | Subordinate 6: wh-interrogative |
| Modal auxiliary | |

**Figure 5: Parts-of-Speech Coding**
**Taken from Ginther & Grant (1997: 388 – 389)**

The categories of error identified were:

1. word form i.e. if the wrong form of a verb, adjective or noun is used (n.b. if there was only one possible correct form, the correct form was also indicated. If there was more than one possible correct answer, then a code was used to indicate this.)
2. word choice e.g. the selection of the wrong preposition
3. word omission .e.g. if the test-taker omitted the article (n.b. omission error codes were placed on the word immediately following the place where the omitted word should have been)
4. spelling

Ginther & Grant (1997) used their analyses to answer the following questions:

1. the influence of test-taker proficiency level on essay errors
2. the influence of test-taker L1 on essay errors
3. the effect of topic on the production of selected parts of speech

They reported that more proficient test-takers (i.e. those who had been rated at level 5 on the TWE scale) wrote longer essays and also produced fewer errors than lower ability test-takers. Additionally, the more proficient test-takers tended to make spelling errors rather than other types of errors whereas the most common error for lower ability test-takers was word form errors. Ginther & Grant also found that the patterns of error by L1 reflected the relative differences or similarities between the test-takers' L1 and English. For instance, the Arab L1 test-takers had the highest percentage of errors per essay and the Spanish L1 speakers the lowest. Chinese and Arabic L1 test-takers were more likely to produce errors of word form whereas the Spanish L1 test-takers most frequently made spelling errors. Interestingly, the Spanish L1 test-takers made more word choice errors than either of the other two L1 groups. Finally, Ginther & Grant (1997) found that the two topics elicited slightly different categories of parts of speech. For instance, topic 1 elicited more examples of negation, gerunds, modal verbs and conditionals than topic 2 whereas topic 2 elicited more adverbs than topic 1. They suggested that this had implications for the equivalence of the topics presented particularly if the mark that the students received was influenced by the presence/absence of certain structures.

Ginther & Grant (1997) suggested a number of avenues for further research. For instance, they said that further analyses should be conducted in order to understand better the effect of certain language features on the marks awarded by assessors. They also suggested that "larger, phrase and sentence-level constructions" should be investigated in order to "evaluate the claim that more complex constructions (such as subordination) are indicative of more mature writers" (1997: 394).

Kim (2004) took a step in this direction in her study of a collection of 33 writing performances by students on an English for Academic Purposes (EAP) course. Her purpose was to describe changes in the grammatical complexity of students' writing that had been placed at different CEF levels. In this small-scale study Kim (2004) focused on three adjacent CEF levels: A2, B1, B2. She conducted three different measures of syntactic complexity:

1. the variety of use of structures
2. the number of subordinate clauses
3. the shift from clauses to phrases

She expected that a comparison of the results of each of these measures would better explain developmental changes between the CEF levels under investigation.

Kim (2004) adopted an analytical framework suggested by Wolfe-Quintero et al (1998), which took the *T-unit* as the basic unit of analysis. The T-unit is also referred to as the terminable unit. It is an independent clause with all its dependent clauses. Take, for example, the following sentence:

> The girl who is getting married tomorrow morning just ran in front of a bus in her haste to collect her wedding dress on time and she was lucky not to be run over.

This sentence comprises two T-units as follows:

i. The girl who is getting married tomorrow morning just ran in front of a bus in her haste to collect her wedding dress on time
ii. She was lucky not to be run over

Kim (2004) conducted the following analyses of each T-unit (ignoring test-takers' errors):

| Measure of syntactic complexity | Analysis |
|---|---|
| variety of use of structures | adverbial clauses per clause (AdcC/C)<br>adjective clauses per clause (AdjC/C)<br>nominal clauses per clause (NoC/C) |
| Number of subordinate clauses | clauses per T-unit (C/T)<br>dependent clauses per T-unit (DC/C)<br>dependent clauses per clause (DC/T) |
| shift from clauses to phrases | prepositional phrases per clause (PP/C)<br>participial phrases per clause (PaP/C)<br>gerund phrases per clause (GP/C)<br>infinitive phrases per clause (IP/C) |

Kim (2004) was then able to compare the analyses for each of the three CEF levels she was investigating. Her results showed a progression from A2 to B2 in all but two of the measures (nominal clauses per clause and gerund phrases per clause). She also found that the results were clearest when comparing A2 and B2. The differences between adjacent levels A2 and B1 were far less clear but there appeared to be a marked increase in syntactic complexity (across measures) when going from B1 to B2.

It is clear from the examples provided in this section that an analysis of test language can provide insights into the:

1. effect of a particular test method upon test-taker performance (for instance, the tape-mediated speaking test)
2. effect of a particular task-type on the language sample elicited
3. influence of topic on the language sample elicited
4. effect of planning time (and other test conditions) upon test-taker performance
5. influence of ability level upon the language sample produced

Unlike CA and DA, an analysis of test language can be performed upon both speaking and writing output. Though no examples have been reported here, I have suggested that it is also possible to analyse the language of the input (for instance in a listening or reading test). I will discuss the analysis of test input in more detail in relation to task characteristic frameworks (see sub-section 4.2, below).

These examples also suggest the following points:

1. The size of the data set can vary. Ginther & Grant (1997) analysed 180 writing scripts while Kim (2004) analysed 33. However, analyses of speaking test language tend to involve relatively small data sets. For instance, O'Loughlin (1995) and Shohamy (1994) studied 20 transcripts of test-taker speaking performances.
2. It is important to define the language features you are using in your analysis. Where competing definitions exist (e.g. O'Loughlin, 1995) I would recommend that you offer a comparison of more than one. In each case, show how the definition affects the results you get and discuss the implications of each for your claims about the quality of your test.

3. Ensure that all your analyses are checked for rater reliability (e.g. Shohamy, 1994; O'Loughlin, 1995; Ginther & Grant, 1997 and Kim, 2004). This will provide proof of the defensibility of your judgements.

Finally, it is important to reiterate that the analysis of language samples is time-consuming and should be used strategically.

## 4. Analytical frameworks

This chapter (particularly section 3) has already made reference to a number of analytical frameworks that can be used to investigate test quality e.g. Conversation Analysis, measures of syntactic complexity and measures of lexical density. Section 5.2 will describe how you might design checklists as a guide for data collection and analysis (usually as part of a study of the test-taking context or of the test-taking process). This section, therefore, will focus on the use of analytical frameworks to analyse test input. The most influential of these is the Framework of Task Characteristics developed by Bachman & Palmer (1996) (see section 4.1, below). However, a recent study involving the CEF has developed a framework that can be used to analyse tests and test specifications (Alderson, personal communication). A brief description of this study is available at http://ling.lancs.ac.uk/groups/ltrg/projects.htm (follow the link for the Dutch CEF construct project).

### 4.1 Task characteristic frameworks

Task characteristic frameworks can help you to analyse your test tasks in some detail in order to explore the extent to which they reflect the test's purpose or perhaps to compare test tasks from two or more versions of a test. The frameworks present a number of 'dimensions' along which the tasks can be analysed or compared. For instance, Weigle (2002: 63) presents a framework that she adapted from Purves et al. (1984: 397 – 8) and Hale et al. (1996) for analysing and comparing writing test tasks. She presents 15 dimensions along which tasks can be described including subject matter, type of stimulus (e.g. graph, table or text), specification of audience, specification of tone, time allowed and choice of prompts.

Fulcher (2003: 57) offers a framework for analysing speaking tasks that includes the following dimensions:

1. Task orientation (for instance is it an open task where the test-taker(s) can decide on the outcome or is the response guided by the rubric? Alternatively, is the task closed and are responses heavily circumscribed?)
2. Interactional relationship (i.e. is there interaction? If there is, how many speakers are involved?)
3. Goal orientation
4. Interlocutor status and familiarity (n.b. in the case of tape-mediated tests it can be argued that there is no interlocutor)
5. Topics
6. Situations

Both Weigle's (2002) and Fulcher's (2003) frameworks are very useful because they are skill specific and therefore take into account characteristics of writing and speaking respectively. A more generic framework is that developed by Bachman & Palmer (1996).

Bachman & Palmer (1996) describe their framework of Task Characteristics as a starting point for task analysis. They list a number of characteristics that should be carefully analysed and described for every task including:

i. the setting (including the physical setting, the participants, and the time of the task)
ii. the test rubrics (including the language of the instructions, the number of parts to the task, the time allotted and the scoring method)
iii. the test input (including the channel of delivery, the length and the characteristics of the language)

iv. the expected response (including the format and the language characteristics)
v. the relationship between the input and the response (including its reciprocity, scope and degree of directness)

(see Bachman & Palmer, 1996: 48 – 57 for more details)

Bachman & Palmer (1996: 57 – 58) suggest that the task characteristics framework can be used as follows:

1. To compare the characteristics of tasks in the target language use situation with test tasks.
2. To analyse existing test tasks in order to make changes or improvements to them.

The Bachman & Palmer framework (as it is commonly referred to) develops on an earlier framework developed by Bachman (1990) called Test Method Facets. This framework was used in a comparison between the Test of English as a Foreign Language (TOEFL) and the Cambridge First Certificate in English (Bachman et al., 1995). Bachman et al. (1995) convened a group of expert judges. These judges were trained to use the framework and subsequently analysed a number of tasks from both tests. The process of training and analysis was as follows:

1. Each judge was given a pair of tests, one from each of the test batteries being studied (FCE and TOEFL). They were asked to study each test carefully and to consider how similar or different they were (and in what ways).
2. The judges were then asked to familiarise themselves with the Test Method Facets framework.
3. They then went through a part of the test, describing it using the Test Method Facets framework. While doing so they were asked to make notes on how well the various descriptive categories in the framework captured their intuitions about the characteristics of the two tests. These notes were used to make revisions to the Test Method Facets framework.
4. The judges then used the revised framework to perform their final analyses of the two tests. For each facet, the judges were asked to place the test task or input text on a three-point scale. For instance, they were asked to rate the rhetorical organisation of the input text on a scale of very simple to very complex. Alternatively they were asked to the number of occurrences of a feature in a test task or input text. For instance, for the facet cultural references, they were asked state whether there were *no occurrences*, *one occurrence* or *two or more occurrences*.

The judges' analyses were used to establish the differences between the tasks on the two tests and to make claims about differences in their underlying constructs. Bachman et al. (1995) reported that agreement between the judges was very high, this implying that the framework helped the experts to pay attention to the key features of the test tasks that were being compared. Their study also demonstrated that the framework allows expert judges to make very detailed judgements about tasks.

However, Clapham (1996) experienced rather more difficulty in applying the Test Method Facets framework in the analysis and comparison of different reading tasks. She tailored the original framework to suit her analysis of IELTS reading tests, reducing the number of facets to 35. However, she found that this was too daunting for her volunteer judges and was forced to reduce the framework further by amalgamating some facets and eliminating others. The final instrument contained only 17 facets. Her procedure consisted of a familiarisation phase and a rating phase. However, despite the familiarisation, Clapham (1996: 149 – 150) remained unsure about their judgements. They commented that some of the categories were not always self-explanatory and were particularly concerned that their analyses would not be stable over time. Finally, Clapham's (1996: 150 – 153) reliability analyses of her judges' ratings revealed quite high agreement for the facets 'grammar' and 'cohesion' but little agreement on facets related to topic specificity. She commented also that her modified Test Method Facets framework did not suit matching and gap-filling tasks (1996: 162).

Any difficulties experienced by researchers are probably because, as Alderson (2000) comments, the framework still needs to be thoroughly investigated through empirical studies and to be modified in the light of the research outcomes. He suggests some possible areas of modification. For instance, the parts of the framework that focus on the characteristics of the test input might not be easy to apply in the analysis of reading test tasks. This is because reading test input comprises both a text and the items that are based upon it. A text might be relatively difficult but the item might be quite straightforward (such as remembering the main 'facts'). Conversely, the text might be quite easy but the item might be rather challenging.

You will have gathered from the discussion so far that there is little published empirical work on the use of task characteristics frameworks. However, despite her own difficulties, Clapham (1996: 162) believes that task characteristics frameworks could be very useful in the content validation of new tests. Indeed, these frameworks have a lot of potential to help us systematise our analyses of test input providing that you bear in mind two guiding principles:

1. You will need to adapt the frameworks already available to suit your test and your context. You will also need to trial and adjust your modified framework until you find that it is practical to use and that your judges understand exactly what they need to do.
2. Remember that the framework is only as good as the judges who use it. Since it is difficult to ensure that a framework is entirely self-explanatory it is important to select your judges carefully and then to familiarise them with the analytical instrument and to also give them sufficient practice in using it before they make 'live' analyses. An issue often debated is whether or not familiarisation and training results in 'cloning' of judgements. This is inevitable and perhaps to some extent some 'cloning' is necessary to ensure the comparability of judgements across raters.

## 5. Feedback methods

Feedback methods such as questionnaires, checklists (particularly observation checklists) and interviews are probably the most familiar methods for gathering qualitative data. They are also typically used in conjunction with each other or with other methods. For instance, in their study of the relationship between students' language proficiency test scores and their subsequent performance on academic degree programmes, Allwright & Banerjee (1997) sent a questionnaire to each student participant at the end of each academic term. The questionnaires were designed to complement each other in order to gather information about each student's study performance and experiences at equally spaced intervals in time. This was to ensure, for example, that the results of the questionnaires at time 2 (in this case the end of the second term of study) could be compared to the results at the end of time 1 (the end of the first term of study) and so on. However, Allwright & Banerjee (1997) also conducted an in-depth interview with each student at the end of their third term of study. During this interview, Allwright & Banerjee (1997) drew on the questionnaire results, probing areas for which the responses had been particularly interesting and also checking their interpretation of the data. They also used the face-to-face meeting to explore aspects of the students' study experiences that were not easy to probe via a questionnaire.

From this example, therefore, it is clear that the different feedback methods are complementary rather than interchangeable. Whenever the circumstances allow, it is often good to 'triangulate' your data by using more than one method (see 7.4 for more discussion). This was the guiding principle behind a set of instruments designed for an International English Language Testing System (IELTS) impact study project (Banerjee, 1996; Herington, 1996; Horák, 1996 and Winetroube, 1997). One set of instruments focused on the classroom. It included a classroom observation schedule, an interview schedule to be used when speaking to the teacher after the observation, and a students' post-observation questionnaire. Further questionnaires were also designed to capture data from teachers and students who were not observed. It is clear from this example that these instruments were intended to complement one another, gathering data from a number of different perspectives and combining different methods of data collection.

The remainder of this section will look more closely at how questionnaires, checklists (including classroom observation schedules) and interviews might be designed.

## 5.1 Questionnaires

Questionnaires gather data that could otherwise also be collected through interviews or focus groups. Their advantage, however, is that they allow researchers to collect views from large numbers of respondents. It can also be easier to manage the data (though this is partly dependent on the questionnaire design) and it is possible to ask face-threatening questions and provide a certain degree of anonymity. Since questionnaires can be completed at any time, respondents also have time to consider their responses.

There are two basic question types – open or closed. Consider the following question pair:

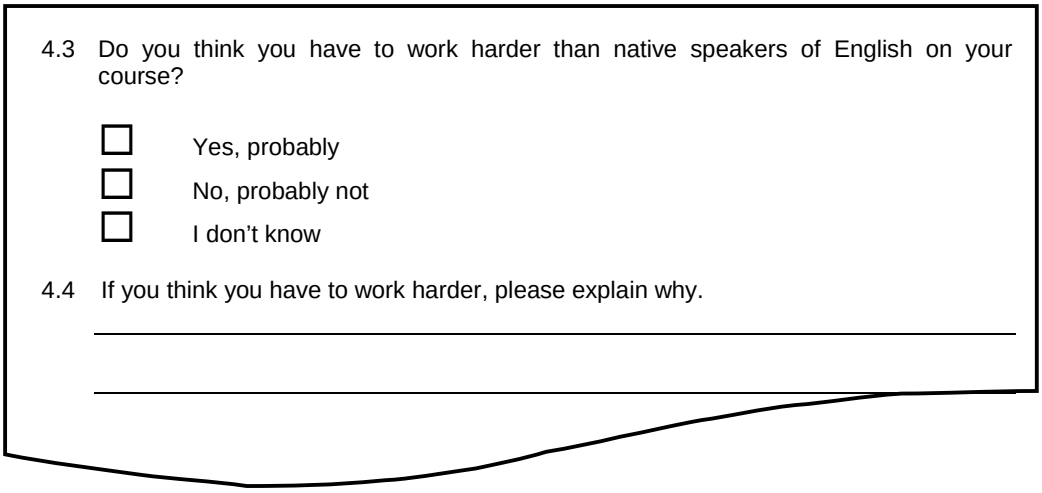



**Figure 6: Open and closed questions**
**Taken from Allwright & Banerjee (1997)**

The first question (4.3) is an example of a closed question. The respondent is asked to choose from one of three responses. Another common closed question type is one that uses a scale:

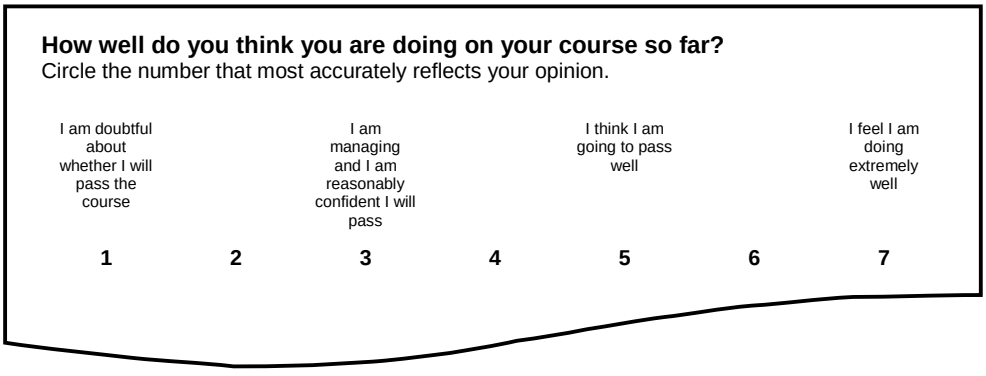



**Figure 7: An example of a questionnaire item using a Likert-scale**
**Taken from Allwright & Banerjee (1997)**

Note that only four points on the scale have been described. Some scales describe all the points and others describe only the two extreme points. You will need to decide how much guidance to give your respondents. It is important to bear in mind, however, that you cannot guarantee that your question will be clearer (and less open to interpretation) if you provide more guidance. Low (1996) has demonstrated the minefields within rating scale wording (e.g. Likert scales), pointing out a number of pitfalls, including:

1. describing the midpoint. If you offer your respondents a midpoint on the scale (e.g. '2' on a three point scale), you need to think carefully about whether the midpoint represents neutrality (i.e. neither agreement nor disagreement with the proposition) or undecidedness about the proposition (i.e. 'I don't know').
2. the number of dimensions that your options capture. Low (1996: 71) provides an interesting example where respondents have to say whether a course has helped them or not. However, the options that they can select include other dimensions such as enjoyment (e.g. 'I've had a lot of fun') and changes in proficiency (e.g. 'I've improved immensely').

The only way you can check that your questionnaire items are clear and are likely to be interpreted similarly by most respondents is by validating them (see section 7.6 for further discussion).

The follow-up question in figure 6 (4.4, above) is an example of an open question. Here, the respondent is asked to explain their answer and they can decide how much or how little they would like to say and what information they would like to provide.

Open questions can also be used on their own. For instance, in order to gauge attitudes to the IELTS test, questionnaires in the IELTS impact study (Banerjee, 1996; Horák, 1996 and Winetroube, 1997) asked both students and teachers to describe three things that they liked most about the IELTS test. Respondents were also separately asked to describe three things that they liked least about the test. Both these questions were deliberately open so that respondents could decide for themselves what they wished to include.

Open questions are particularly useful when you are not sure what the range of responses is likely to be (i.e. if your research is an initial exploration of issues) or if you want to avoid 'suggesting' answers to your respondents. You will find it easier to use closed questions when you are certain of the possible range of responses and/or when you want to make sure that you gather information on all the possibilities. In other words, you want to make sure that no possible response is accidentally forgotten.

It is important to note, however, that each question type has advantages and disadvantages. The advantages of closed questions are that they are quick to answer, process and compare. However, closed questions provide no scope for other answers and can reflect the researcher's bias in the categories provided. For instance, if you look more closely at the closed question presented in figure 6 you will see that the responses assume that the students' should compare their **overall** effort to that of their native-speaking classmates. However, further research by Banerjee (2003) has shown that students' experiences differ from subject to subject within a particular degree programme. For instance, MBA students with a background in Engineering find the more quantitative courses such as Management Science relatively easy. They find that they do not need to work harder than their native-speaking classmates on these courses. However, these students find that they struggle with less familiar and more language-oriented subjects such as Behaviour in Organisations. Therefore, the students will find it hard to give a single answer to the question 'do you think you have to work harder than native speakers of English on your course?'. Indeed, respondents could become frustrated or irritated if the response options did not suit what they wanted to say.

Open questions, on the other hand, provide more scope for a variety of answers and also allow the researcher to probe answers (e.g. 'please explain your answer'). But, such questions are time-consuming to complete and demand more effort and commitment from respondents. It is also more time-consuming and difficult to code and analyse the responses. In particular you will need to interpret responses in order,

for instance, to decide whether two differently worded answers from two respondents mean the same thing.

The foregoing discussion has revealed that open and closed questions are equally useful and both have drawbacks. Indeed, there is no perfect question type. Rather, you should select the best type for your purposes. In most cases, you will decide to use a combination of open and closed questions as this will allow you to combine focused and proscribed questioning with some more exploratory prompts. Regardless of the question type you select you also need to think carefully about the wording of your questionnaire. Check your draft questionnaire for the following pitfalls:

i. double-barrelled questions – your respondents are likely to find the question difficult to answer and you will find it impossible to determine whether the answer refers to only one (indeed which one) or both parts of the question.
ii. unclear instructions – so respondents are not sure what to do.
iii. questions that do not apply to the respondent – it is important to allow respondents to indicate when a particular item does not apply to them.
iv. questions that rely on memory or are hypothetical – e.g. the responses to such questions are unlikely to be stable or accurate.
v. biased options – respondents might be uncomfortable about selecting an option that has been presented in a negative light.

Beware also of mixing positively phrased items with negatively phrased ones. If your respondents do not read each question carefully, they might give the wrong response:

> I think it is important to check the dictionary when I do not understand a word
> I do not think it is important to check my work after I have finished writing

Oppenheim (1992) and Dörnyei (2003) provide good overviews of questionnaire design. Dörnyei (2003) gives particularly practical advice on the length and layout of the questionnaire. In particular he advises researchers to resist the temptation to include every question that they think might be useful. He warns that a questionnaire should not take more than 30 minutes to complete. He also reminds us that we need to take into account the reading speed of our respondents (2003: 17 – 18). Therefore, if you are gathering questionnaire data from young learners (e.g. 10 – 12 year olds) or are administering your questionnaire in a student's L2, then you need to consider how quickly they will be able to read and respond to the questions. Indeed, you should also to make your wording simple and accessible to the lowest level student you are gathering data from.

Dörnyei's (2003) advice makes it clear that questionnaire design is very complex and requires you to be very clear about the information you are trying to gather and also to think carefully about how to elicit that information in the most economical way possible. I would suggest the following six-step procedure for questionnaire design:

1. Brainstorm all the areas and possible questions that your questionnaire should cover.
2. Write questions to address each of these areas.
3. Return to the original purpose of your questionnaire. Eliminate all the questions that do not address that purpose.
4. Group the questions so that you can see where overlaps exist. Examine the overlaps in order to decide whether or not they are necessary. Bear in mind that you might want to ask the same question twice (in slightly different ways) in order to check the stability of your respondents' views.
5. Format the questionnaire and administer it to a small group of target respondents. Ask them to mark the questions that they do not understand. Time how long it takes for each respondent to complete the questionnaire.

6. Re-work the items that were difficult to understand. If the questionnaire was too long, consider carefully whether you can remove any questions without damaging the coverage of your questionnaire.

Questionnaires can be used to investigate test quality in a number of ways. For instance, they can be used to gather feedback from test-takers. Brown (1993) explored the usefulness of test-taker feedback questionnaires for the test development process. She gathered feedback from 53 test-takers during the trialling of a tape-mediated test of spoken Japanese for the tourism and hospitality industry – the Occupational Foreign Language Test. The questionnaire had two parts. In part one, the test-takers were asked for their overall attitudes to the test. For example they were asked if the test reflected accurately how well they spoke Japanese and whether they believed that the test reflected the type of language they would need in the tourism and hospitality industry. In part two, the test-takers were asked to comment on individual sections of the test. They were asked to rate each section for its usefulness and difficulty and also to say whether they had had enough time to respond. The test-takers were also encouraged to make comments on any items that they found problematic. Brown (1997) commented that the survey results confirmed that the content and level of the test was appropriate for the target language use situation. She also reported that the results revealed a lot about the expectations of the test-takers and indicated that much more advance information was needed. This feedback was used to improve the test handbook.

Clapham (1997) also used questionnaires during the test development process. She presented the revised IELTS test and specifications to different stakeholders, along with a detailed survey instrument that asked for their views on the extent to which the revised test sampled the test specifications. The questionnaires are presented in full in Clapham (1997: 133 – 140). One questionnaire was sent to academic subject specialists who would teach students who were admitted to university on the basis of their IELTS scores. This instrument included questions about whether the texts were comparable to the sorts of texts that students would have to read in their academic courses and whether the reading tasks were comparable to the reading tasks that students would have to perform on their academic courses. The subject specialists only had to look at one version of the IELTS test in order to answer these questions. The second questionnaire was sent to language teachers, testers and applied linguists. It contained the same questions as the questionnaire for subject specialists but the language specialists were asked to look at more versions of the IELTS test. The results of these questionnaires were used to make changes and improvements to the specifications of the IELTS test.

Marinič (2004) has demonstrated how they can be used during the test piloting phase to gather feedback from test-takers. She showed that test-takers can be asked for their views on the topics and methods of the test-tasks, the clarity of the instructions and also whether they were given sufficient time in which to complete the tasks. Additionally students can be asked whether they found the task difficult. In some studies, students have been asked to estimate whether or not they got the item correct as well. Marinič (2004) explained that this data could be analysed alongside the item statistics available for the tasks in order to judge the quality of individual test tasks.

Data can also be gathered routinely after live administrations. Halvari & Tarnanen (1997) described a study of the Finnish National Certificate language tests. The National Certificate tests can be taken in a number of different languages but the most commonly taken languages are Finnish, Swedish and English. It is not uncommon for a test-taker to sit for a test in more than one language. Such test-takers are a good source of information about the comparability of tests at the same level but in different languages. Halvari & Tarnanen (1997) distributed questionnaires after the test administration to test-takers who had taken tests in more than one language. The test-takers were asked whether they agreed with the scores that they had obtained (both the overall score and their score for each sub-test). They were also asked to identify any differences between the contents of the tests in the different languages. Halvari & Tarnanen (1997:

134) categorised the comments they received into three basic groups. They found that test-takers commented on:

1. differences in the test-taking context (e.g. one test-taker commented that the room for the German test was very cold).
2. the relationship between their test result and their 'true' language ability.
3. differences between the content of the tests (n.b. some of these were comments about test difficulty i.e. the English test was more difficult than the Swedish test).

Despite some interesting results, Halvari & Tarnanen (1997) found that the response to their questionnaire was rather low. This made it difficult for them to draw specific conclusions. Nevertheless, they argued that such data can throw light on the tests from the test-takers' perspective and can be used to make improvements to the test conditions and tasks.

Another use of questionnaires is to gather background information about test-takers. Test-takers routinely provide information when taking the IELTS through the Candidate Information Sheet (CIS). This instrument asks test-takers for their gender, age, language background and other language learning information. Herington (1997) developed a more detailed version of the CIS as part of the IELTS impact study project (described above). This questionnaire included questions about the students' attitudes to learning English and to taking English tests. They were presented with a list of statements about learning English and taking English test and asked to indicate how strongly they agreed or disagreed with each statement. For instance:

| | -3 | -2 | -1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| English is an easy language to learn | | | | | | | |
| I feel nervous when I see new words in an English test | | | | | | | |

Herington (1996: 48)

Here *–3* represented 'strongly disagree' and *3* represented 'strongly agree'.

Herington's (1996) instrument also asked test-takers to describe their learning strategies and their test-taking strategies. For instance:

| | 0 | 1 | 2 | 3 | ☺ |
|---|---|---|---|---|---|
| I learn new words in English by translating them into my language. | | | | | |
| During an English test the first thing I do when I read a passage is to look for the main ideas. | | | | | |

Herington (1996: 49 - 51)

The scale for this section ranged from 0 (never) to 3 (always). It also included an interesting additional option - ☺. This symbol meant 'a good idea but I don't do it'. Herington (1996) hoped that this would help test-takers to be very accurate in their claims about the strategies they used.

Background information questionnaires such as the one Herington (1996) designed can be used when analysing test-takers' performances on the test. The results can be categorised according to country of origin, language background and gender. Such analyses are routinely performed by testing organisations such as the Educational Testing Service (ETS – http://www.ets.org). You might even analyse the better (or worse) performers in more detail to see if they use common learning or test-taking strategies. This information can be used to give advice to future test-takers about how to prepare better for the test.

You might even wish to gather specific background information if you are considering major changes to your test. When ETS was preparing to introduce the Computer-based Test of English as a Foreign

Language (TOEFL CBT) they conducted a number of computer familiarity studies across the world (Kirsch et al., 1998; Eignor et al., 1998 and Taylor et al., 1998). In their first study they surveyed 90 000 test-takers. Each test-taker was asked to provide some background information such as their country of origin, educational background and language background. They were also asked to complete a computer familiarity scale. For instance, test-takers were asked how often they had access to a computer. They were also asked where they had access to a computer (e.g. at home, at work etc.). Test-takers were also asked to indicate how often they used the computer for specific tasks such as surfing the Internet. The responses to this familiarity scale were analysed to give profiles of the computer familiarity of test-takers in different parts of the world and from different backgrounds. A second study was then carried out to compare the test-takers familiarity with computers to their performance on a set of 60 computer-based TOEFL tasks. Each test-taker first took a computer familiarisation tutorial which trained them in the computer skills that they would need in order to take the computer-based TOEFL (e.g. how to use a mouse). Taylor et al. (1998) report that there was no evidence that the computer delivery of test items affected test-taker performance (regardless of the test-taker's previous computer familiarity). This indicated that the familiarisation tutorial provided sufficient support to test-takers who were unfamiliar with computers.

Other test-taker characteristics might also affect the construct validity of a test. For instance, Allan (1992) developed a scale of test-wiseness in order to explore the effect of test-taking strategies upon test-takers' performance on a reading test. He argued that the test-taking skills of L2 learners had little to do with their reading abilities yet could affect their final reading test scores. Allan (1992) developed a 33-item instrument and administered it to 51 students in a Hong Kong polytechnic. Each item was a multiple-choice question. The test-takers had to answer the question by choosing the most appropriate option from the choices. The items were designed such that the test-takers would not be able to answer them from their background knowledge. The correct answer was 'cued' in one of the following ways:

1. stem-option (there is an association between a word in the stem and a word in one of the alternatives. This association is usually semantic or grammatical).
2. grammatical cue (the option grammatically matches the stem e.g. the form of the article might suggest that the option should begin with a vowel sound)
3. similar option (this is when all but one of the options are similar in meaning. This makes the 'odd' option stand out)
4. item giveaway (the answer to the item can be found in another item)

Approximately one third of the students were also asked to provide brief explanations for their answers. This data was used to throw light upon the responses. Allan (1992) found that the items in the 'grammatical cue' and 'item giveaway' sets appeared to correlate well with one another. The results for the other two sets ('stem-option' and 'similar option') were less clear. Nevertheless, he argued that some students were more sophisticated test-takers. He further suggested (1992: 109) that this was particularly problematic for teacher-designed tests because these were less likely to be carefully piloted and validated.

Questionnaires can also be used to investigate the processes used by test-takers to complete different items. Li (1992) administered a questionnaire within a reading test in order to explore which reading strategies each test-taker used to complete individual items. The test-takers first completed an item and then indicated which of a list of reading processes they had used to do the item. He also asked them to indicate whether they found the item difficult or easy. Li's (1992) analysis of the questionnaires confirmed the findings of Alderson (1990) that test-takers use a variety of reading skills to complete test items. While some overlap may exist, in general it is very difficult to predict the reading skills that test-takers will use to complete a particular test item. This research cast doubt on whether test constructors can design items that test specific skills.

The studies described in this section have shown that questionnaires can be used in a number of ways to examine test quality:

1. To canvas test-taker views on the difficulty and/or appropriacy of test items.
2. To explore the views of other stakeholders such as teachers, test designers and applied linguists on the suitability of test input and test tasks for the target group of test-takers.
3. To gather information about test-takers in order to profile the test-taking population.
4. To establish the need for test-taker training or familiarisation as well as the nature of that training.
5. To investigate possible threats to construct validity (such as the influence of test-wiseness or computer familiarity upon test-taker performance).
6. To explore test-taking processes and strategies.

Questionnaires can also be used at various stages in the test development process as well as during live administrations. It is important to note, however that questionnaire response rates can be low. Indeed, Halvari & Tarnanen (1997) report that only 63% of the questionnaires they distributed were returned and return rates can sometimes be as low as 30%. It is therefore better to ask respondents to complete questionnaires in your presence (either in class or immediately before test-takers are released from the testing venue). This ensures that they have to hand in the questionnaire before they leave.

### 5.2 Checklists

If you have ever taken a car for a routine service you will probably have noticed that the mechanic has a form that must be filled during the procedure. The form comprises a list of features that must be checked. The mechanic is required to tick every item off and also to note any problems in a space provided. This is a checklist.

Checklists are used in a variety of contexts including store inventories and quality control inspections. They are also very useful in investigations of test quality. The key feature of checklists is that they structure observations. As such they can vary in format from very clearly defined lists, where the researcher simply ticks for the presence or absence of a particular feature or characteristic, to more open grids. In their more open form, checklists might simply comprise a list of column or row headings with space in which to make notes. The checklist for validating speaking tasks developed by O'Sullivan et al. (2002) falls into the former category, while the Communicative Orientation of Language Teaching (COLT) observation instrument developed by Allen et al. (1984) falls into the latter category. Alternatively, a checklist might combine elements of the two as does the Classroom Observation Instrument designed for the IELTS impact study project (Banerjee, 1996). The first three pages of this instrument comprised an open grid that asked observers to note the time taken for each activity, what the teacher did, what the students did and the nature of the interaction. The remaining pages listed different task types and text types as well as different interaction patterns. The observer was asked simply to tick the task types, text types and interaction patterns that he/she observed.

It is rare for a checklist to be adopted directly from another context. Instead, researchers usually survey and analyse other checklists, paying attention to the features that might be useful in their context. Banerjee (1996) used this process when she designed the Classroom Observation Instrument for the IELTS impact study project. She first analysed the COLT observation instrument (Allen et al., 1984) and an instrument designed for the Sri Lankan impact study (Wall & Alderson, 1993). These proved very useful in suggesting an overall design for the observation instrument. In order to identify specific items to include in the checklist, Banerjee (1996) needed to decide what washback from the IELTS might look like. To achieve this she closely examined the test materials and published teaching materials available for the test (in this case the IELTS test). She also brainstormed the content of the checklist with other researchers, teachers and students. Though this was not possible in the case of the IELTS impact study project, it is also advisable to analyse the test specifications. Additionally, it is always useful to attend a typical lesson in order to document the teaching and learning that takes place (either through field notes or a video-

recording). This will enable you to identify categories of data that you would like to capture. All these sources of information (test materials, specifications, published teaching materials, brainstorming etc) will help you to compile a full list of the activities, interactions, text-types etc. that could occur in a typical lesson. A detailed checklist can then be produced.

The checklist should then be extensively trialled and revised until you are sure that it is easy to use and will also help the observer to capture all the information being sought. Banerjee's (1996) observation checklist was reviewed by the Language Testing Research Group at Lancaster University, a group of researchers, teachers and students with a lot of experience in designing research instruments. Banerjee (1996) also trialled her observation checklist with an IELTS-type class in order to ensure that it was practical to use in a live observation. She conducted this observation exercise with a colleague with whom she was later able to compare notes. This comparing of observation notes revealed the extent to which the observation checklist helped the two observers to make note of the same features of the lesson (a reliability check).

As has already been stated (above) Banerjee's (1996) final instrument combined an observation sheet and a checklist of activities, interactions and text-types. It was very similar in structure to the observation checklist that Wall & Alderson (1993) used when they investigated the effect of the introduction of a new Secondary School leaving test ('O' level) upon the teaching that took place in Sri Lankan classrooms. At the time little empirical research had been carried out to establish the influence of a test upon teaching and learning in the language classroom. Wall & Alderson's (1993) study was also innovative in that it included direct observation of classrooms whereas previous research had been based on questionnaires and interviews. Indeed, it is important to note that the data gathered from questionnaires and interviews is self-report data i.e. what teachers, students and test-takers 'say' they do or believe. It is often useful to complement such data with direct observation such as classroom observation or the observation of live test administrations in order to, as Wall & Alderson (1993: 42) argue, take into account not only what study participants report about the effect of an exam upon their teaching, learning and/or test-taking practices, but also to capture what those practices might look like in reality.

Wall & Alderson (1993) hoped to examine the extent to which the new Sri Lankan English 'O' level had influenced the types of teaching activities that took place as well as the interaction patterns (e.g. teacher-student or student-student interaction) and the input text types. Therefore, their observation instrument included checklists of different teaching activities, interactions and input text types. These lists included activities, interactions and text-types that occurred in the test as well as other activities, interactions and text types that were not represented in the test and which it was hoped would not occur in the classroom because they were considered to be poor teaching practice. A copy of this observation checklist can be found in Alderson & Wall (1992).

The observations were conducted by seven Sri-Lankan teachers, each of whom visited 7 schools six times over a period of two years. It is important to note that the six rounds of observation were carefully timed to capture different 'moments' in the academic year. For instance, round 1 took place at the start of the first year, round 2 was scheduled for the middle of the school year (four months after the first observation round and three months before the examination). Round 3 took place shortly before the examination. Rounds 4 – 6 followed the same pattern in the following academic year. Wall & Alderson (1993) encountered a number of difficulties in the data-gathering for this study. Firstly, the round 1 observations were disrupted by political instability in Sri Lanka. The round 3 observations were also affected, this time by the fact that students were released from regular classes more than one month before the examination so that they could study for the exams. Wall & Alderson (1993) also had to cope with changes in the team over the two-year period of the study. Finally, the Sri-Lankan teacher-observers sometimes had difficulty in being released from their regular teaching duties in order to conduct the observations. These difficulties are instructive because they are not unusual. Any study will have to take into account the 'rhythm' of the

teaching year (including the fact that teaching might be suspended early for examination classes) as well as the availability of research participants and helpers. It is always important to gain the support of governing bodies so that you can maximise the co-operation you might expect for your study.

Despite the difficulties they encountered Wall & Alderson (1993) reported that they had a full data set (i.e. 6 rounds of observation) for 18 schools. Also, at its largest the sample contained 64 schools (the observations from round 5). Even though the smallest round of observations contained data from only 18 schools, the second smallest round of observations included a creditable 36 schools. Most of the data that was gathered through the observations was analysed using the statistical software tool SPSS (http://www.spss.com) to calculate the frequency of occurrence of particular features. For instance, Wall & Alderson (1993) calculated the percentage of classes that were devoted to the different language skills (reading, writing, listening, speaking and language form). This amounted to a quantitative analysis of data that had been collected using a qualitative data collection method. This is not unusual for the analysis of questionnaires and checklists. Indeed, quantitative analysis of data is a useful complement to qualitative analyses and Wall & Alderson (1993: 55 - 57) also looked carefully at patterns in the teaching methodology, reporting a tendency towards a lockstep approach where the teacher dominated the interaction. As a result of this combination of analyses, Wall & Alderson (1993: 66) reported that the Sri Lankan 'O' Level examination had some effect on the content of teaching and upon the design of in-class tests in Sri Lankan classrooms. However, they could not find evidence of the effect of the examination upon the method of teaching.

A recent and rather different example of a checklist is the observation checklist developed by O'Sullivan et al. (2002) to validate speaking tasks. O'Sullivan et al. (2002) were motivated by the fact that most speaking test validation requires detailed and time-consuming analyses of test language as has been described in section 3 (above). They wanted to develop a framework that could be used during live administrations to analyse the language elicitation tasks (LETs). They argued that the performances elicited by LETs should match the predictions of test designers if we are to make valid interpretations of test-takers' scores but also contended that analyses of test language (the most common method for analysing speaking test performances) were time consuming and demanded considerable expertise. Consequently, the sample of test performances subjected to such analyses tended to be small and was therefore not easily generalisable. O'Sullivan et al. (2002: 39) argued for a methodology that complemented more detailed analyses of language samples but could be applied to larger numbers of test takers.

O'Sullivan et al. (2002) began by reviewing the literature in spoken language, second language acquisition and language testing in order to identify a set of informational and interactional functions that can occur in spoken language. Three lists were written initially and these were then refined via a number of meetings in which participants used the checklists and then commented on their usability. Through this process, items on the checklist that could not achieve a high degree of agreement were discarded and other items were improved to make them more transparent. The final version of the checklist is presented in O'Sullivan et al. (2002: 54). It consists of three categories of functions: informational functions (including providing personal information, speculating and describing), interactional functions (including agreeing, modifying and asking for information) and managing interaction functions (including initiating, reciprocating and deciding). This checklist is a good example of the way in which a data collection framework can be developed and used in post-hoc analyses of test output. It is important to note, however, that the final form of the checklist was influenced by the Cambridge ESOL tests on which it would be used. This is further evidence of my earlier claim that observation instruments like checklists are rarely adopted directly from another context. They are more likely to be customised to the test being investigated.

The research reported so far has demonstrated that checklists can be used to investigate test quality in the following ways:

1. To explore the impact/washback of a test upon the teaching and learning in the language classroom.
2. To investigate the match between test-designers predictions and the actual language elicited by test tasks.

Checklists can also be used during item moderation meetings. Observers can use them to record the decisions that are taken with respect to individual items and the test as a whole. Similarly, checklists can be used during rating scale development. The resulting data can reveal a great deal about the construct of the test as well as the thought processes of item writers and test and scale developers. Additionally, test-takers can be observed while they are taking the test and assessors can be observed during the rating process (as a complementary procedure to verbal reports). It is clear, however, that checklists are used in these contexts to structure observation. Finally, you will probably also find it useful to audio or video record events such as item moderation meetings and assessor moderation exercises. The transcripts from these recordings can later be analysed in greater detail.

The studies reported here also indicate that checklists (like questionnaires) can be used to collect larger samples of data in a systematic and easily comparable manner. However, there are also some key considerations:

1. Stability of the group that conducts the observations. Wall & Alderson (1993) found that their observation team changed from one study year to the next. Additionally, their observers were also teachers and sometimes found it difficult to get leave from their teaching responsibilities in order to carry out the observations.
2. Training for the observers. O'Sullivan et al. (2002: 46) argue that observers should be trained to use the checklists "if a reliable and consistent outcome is to be expected". As with the use of task characteristics frameworks (see section 4.1), training will inevitably result in 'cloning' of observers. However, this is important if you intend to compare and combine different observations.
3. Observation checklists should be piloted extensively and validated carefully to ensure that they are performing appropriately in the context for which they are used. Validation issues will be discussed in section 7.6 (below).

### 5.3 Interviews

The final feedback method to be discussed is the interview. It is probably best described as "a conversation between interviewer and respondent with the purpose of eliciting certain information from the respondent" (Moser and Kalton, 1971: 271) and has many of the same purposes as questionnaires. It differs from questionnaires primarily because it is a more flexible data collection method; a questionnaire item is pre-prepared and cannot be altered at the point of administration whereas an interview question can be altered to suit the flow of the interaction between the interviewer and the respondent. Yet questionnaires and interviews should not be viewed as polar alternatives. You will probably find that they combine well with each other. Questionnaires can be used to gather information on a set of clearly defined themes from a large number of respondents (some sample sizes exceed 1000 respondents) while interviews can be used to probe some themes in greater depth and detail with a sub-set of the questionnaire respondents.

Interviews can take a number of different forms. They can be individual (where there is one respondent and one interviewer) or group (where there are two or more respondents and one interviewer) interviews. Individual interviews have the advantage of your being able to focus in considerable detail upon the views of a single respondent and to build a picture of an individual test-taker or stakeholder. However, group interviews can be used to brainstorm ideas and to establish group viewpoints. One advantage of the group

interview is that the interaction between respondents can sometimes spark revelations that you, as the interviewer, might not succeed in eliciting from a single respondent.

Interviews can also vary in their degree of structure. Regular census data is often collected by structured interview. The interviewer either contacts you by telephone or by coming to your front door. He/she has a fixed schedule of questions to ask. The wording and the order of the questions is pre-determined. At their most structured, such interviews closely resemble questionnaires. Shohamy et al. (1996) conducted structured interviews with teachers and inspectors as part of their investigation into the impact of two national tests - an Arabic as a second language test and an English as a foreign language test. The interviews included questions about preparation for the test, stakeholders' knowledge about the test and the impact of the test upon teaching and testing practices (1996: 302). This data was complemented by data from questionnaires administered to students and an analysis of test documentation such as bulletins issued by the Ministry of Education.

Unstructured interviews fall at the opposite end of the continuum. The ground covered in these interviews is dependent upon the interaction between the respondent and the interviewer. The latter rarely has more than a set of themes to guide the discussion. Though this is the most flexible of the interview structures, it is also the most demanding. If poorly handled, interviewers risk that the interview data will not result in helpful or interesting revelations. Indeed, such interviews are usually best conducted by highly experienced and well-prepared interviewers.

Taking the middle ground are semi-structured interviews where the interviewer has an interview schedule to guide the discussion but where there is some room for the respondent to negotiate the pace and coverage of the interview. Allwright & Banerjee (1997) used this type of interview in their investigation of the study experiences of non-English speaking post-graduate students at a British university. They selected this interview type for a number of reasons:

1. They were each going to interview half the students in a series of individual interviews. Consequently, they needed to have a structure to follow so that their respective interviews yielded comparable data.
2. Though their concern for having comparable data suggested the use of a structured interview, Allwright & Banerjee (1997) wanted to retain some flexibility to respond to the themes that emerged during the interviews.

Since the semi-structured and unstructured interview allow the interviewer to respond to the data as it emerges, this also means that these interview types have a distinct social dimension. Consequently, their direction and success can be influenced by the interaction between the interviewer and the interviewee. Banerjee (1999) compared the interviews she conducted as part of the Allwright & Banerjee (1997) study with those conducted by Joan Allwright (the lead researcher on the project). Banerjee's (1999) analysis of the transcripts revealed that the interviews between herself and the study respondents were slightly strained in comparison to those conducted by Joan Allwright. The students she interviewed appeared unwilling to respond to questions that probed their responses. In contrast, the students interviewed by Joan Allwright seemed generally more willing to elaborate and often stayed well beyond the agreed time limit for the interview. Banerjee (1999) viewed this experience as an example of what Mishler (1986) describes as the co-construction of the interview by the participants. She argued that the interviews were different because the people involved were different and the interpersonal dynamic therefore differed. She contended further that the key to that different dynamic lay in the relationship she had with the respondents compared to Joan Allwright's relationship with them. At the time she was the research assistant on the project and a research student. As such she was the respondents' equal – a fellow student. In contrast, Joan Allwright was a member of staff. Banerjee (1999) argued that this power differential at least partly determined the tendency of the respondents to be more forthcoming with Joan Allwright and

less impatient to end the interview. They possibly wanted to appear co-operative for the interviewer they perceived to be in a superior position to them.

It is, of course, also possible that one interviewer may be more experienced and therefore more skilled than the other. This underscores the importance of preparing thoroughly for interviews. Borg & Gall (1983) advise that it is important to eliminate any bias that might be introduced by factors such as the length and location of the interview, the attitude of the informant to being interviewed and/or to the researcher and the behaviour of the researcher. It is clear, therefore, that interviews should be designed and piloted carefully. Always ensure that the interviewer has had an opportunity to practice conducting interviews before he/she begins collecting data. Indeed, if you plan to use a team of interviewers, it is useful to conduct an interviewer training session in which each interviewer can practice his/her interview technique as well as analyse and reflect upon the practice interview. If combined with a piloting procedure, the interviewer training can be used to refine and clarify the aims of the interview for all the interviewers.

It is important to note, however, that training and piloting will not eliminate (or render inconsequential) the effect of the interpersonal dynamic between interviewer and respondent upon the interview. I would recommend that, where possible, you should try to include familiarisation questions that allow the interviewer and respondent to relax in one another's company. You will probably also find it useful if you systematically note details about the interview situation such as the place, physical setting (arrangement of furniture, position of participants relative to one another) and the relationship between the interviewer and the respondent. This is because, as Stimson (1986) argues, data analysis should take account of the effect that the data collection setting might have upon the respondent.

As I have already said, interviews are rarely the only data collection method in a study. They tend to be combined with at least one other method such as observations (e.g. Alderson & Hamp-Lyons, 1996) or questionnaires (e.g. Shohamy et al., 1996 and Allwright & Banerjee, 1997). They are useful in investigations of test quality because stakeholders (including test-takers, teachers, administrators and parents) can be asked for their views about the test including the overall quality of the test (the extent to which they believe the test gives a true picture of language ability), the difficulty of specific tasks, items or input texts and the extent to which the input texts and tasks are interesting and/or authentic. Interviews can also be used to examine how test scores are interpreted and used by receiving institutions and other stakeholders.

Clearly, the advantage of interviews is that the interviewer can concentrate on a single respondent and thoroughly explore his/her views on the test. The interviewer can also probe responses in order to better understand the respondents' views. In this way interviews can provide a wealth of detail that might not be available from a questionnaire. However, interviews can be time-consuming (an interview can take an hour or more to complete). This means that fewer informants can be studied, which can in turn affect the generalisability of your results.

## 6. Using qualitative methods for standard-setting

I suggested at the start of this chapter that the qualitative methods described here could also be used for standard-setting. You will have read in the chapter on standard-setting (see Section B) that the establishment of cut-off scores involves expert judgements. You will also know that it is important to safeguard the validity of these judgements. This can be done using qualitative procedures. This area of research is still rather new so there is little published guidance on how to use qualitative methods to establish the validity of standard-setting procedures. This sub-section will suggest a few procedures that could be applied during the judgement phase (when standards are set) as well as during the specification phase (when the content coverage of the test is examined).

During the judgement phase it is necessary to establish benchmark performances for the productive skills (writing and speaking) and to establish benchmark texts, items and responses for the receptive skills (reading and listening) as well as for tests of linguistic competence (e.g. grammar and vocabulary) (for more details see Chapter 5 of the Manual). Expert judges establish these benchmarks by placing texts, items, responses and/or performances in the CEF bands A1 – C2. This process can be monitored and investigated as follows:

1. Judges can be asked to prepare their assessments individually. A meeting can then be convened in which each judgement is discussed.
2. The discussion can be recorded and observation notes can be taken.
3. The observation data and the transcripts of the recordings can be analysed later to explore the sources of agreement and disagreement more closely. This will throw light on the characteristics of test items, input texts, test-taker responses and/or performances that signal a particular benchmark. It will also help to explain the features of test items, input texts, test-taker responses and/or performances that can cause variation in expert judgements.
4. Additionally, selected participants could be asked to perform a retrospective verbal protocol. It might be helpful to use a stimulated recall protocol if the verbal protocol takes place a few days or weeks after the benchmarking meeting. This data could explain how the judges made their benchmarking decisions. It might reveal criteria unrelated to the performance or test input that have influenced the benchmarking decision. The latter could constitute a threat to the validity of the benchmarking.

This data could also be used to establish the validity of the final benchmarks and could inform future training and familiarisation programmes for expert judges.

Cut-scores are also estimated during the judgement phase. Subsequently, test-takers who receive scores above the cut-score will be presumed to have met a particular performance standard. Test-takers whose scores fall below that cut-score will be presumed not to have fulfilled the requirements for that standard. Yet, as Kaftandjieva (Section B of this volume) points out, cut-scores are arbitrary. It is necessary, therefore, to gather evidence of the validity of the final cut-scores in order to legitimise them. But the validation of standards is not achieved by an appeal to external criterion (Kane, 2001). Instead it is important to gather evidence to support the cut-score decision. This can be done by demonstrating that the decision-making process was logical and reasonable and that the decision is plausible. Qualitative evidence could be gathered at the following points in the process of setting a cut-score:

1. The meeting at which individual judges discuss their individual conclusions about the cut-score can be recorded and observation notes can be taken. The observation data and the transcripts of the recordings can be analysed later to explore the sources of agreement and disagreement more closely. This will throw light on the characteristics of test-taker responses that signal a particular level of performance. It will also help to explain the features of test-taker responses that can cause variation in expert judgements.
2. The transcripts and observation notes can also be analysed to demonstrate that the cut-score procedure was carried out correctly and with appropriate attention to detail.
3. It might also be useful to conduct follow-up interviews with the judges. The interview questions should ask for their views on the cut-score procedure. The judges should also be asked if they believe the final cut-score was appropriate and whether they felt able to be honest in their judgements during the setting of the cut-score. These interviews will provide evidence of the credibility of the procedure followed and also of the extent to which the final judgement is plausible.
4. Additionally, selected participants could be asked to perform a retrospective verbal protocol or a stimulated recall protocol of their own judgement process. This data could explain how the judges made their cut-score decisions. It might reveal criteria unrelated to the performance or test input

that have influenced the cut-score decision. The latter could constitute a threat to the validity of the cut-score.

During the specification phase judges are likely to be asked to examine the content coverage of the test. The judges will examine each input text and item to answer a number of questions such as:

i. Which situations, content categories, domains are the test takers expected to show ability in?
ii. Which communication themes are the test takers expected to be able to handle?
iii. Which communicative tasks are the test takers expected to be able to handle?
iv. What kind of communicative activities and strategies are the test takers expected to be able to handle?

(examples taken from Form A10, Council of Europe, 2003: 43)

The validity of this process can be established in similar ways to those described for the judgement phase:

1. The exemplar judgement sheet provided in the Manual, Form A10 (Council of Europe, 2003: 43), requires judges to provide evidence for their judgements. This evidence could be compared across judges to identify similarities and differences in the evidence selected to justify the judgements made.
2. A few judges could be asked to perform a retrospective verbal protocol or a stimulated recall protocol of their own judgement process. This data could explain how the judges performed the analyses and selected their supporting evidence. It might also provide additional insight into the judgement process that the judges had not written down.
3. It might also be useful to conduct follow-up interviews with the judges to explore the evidence provided in more detail. For instance, judges could be presented with the evidence that they did not provide and asked to discuss the suitability of that evidence. This will explain differences in the evidence provided.

The verbal protocol and interview data may also provide you with feedback on the usability of the forms.

## 7. General issues arising

The discussion so far has revealed that qualitative methods of investigating test quality share a number of theoretical and practical concerns. The more practical issues include deciding what language to collect the data in, how to go about piloting and trialling the instruments and what level of detail to provide in transcriptions. The more theoretical issues include decisions about triangulating data sources, analysing the data, the validity of the instruments and procedures and the generalisability of the results. In this section I will return briefly to each of these issues.

### 7.1 Language that the data is collected in

I commented in 2.1 that, when collecting qualitative data, the choice of language is not necessarily straightforward. It is relatively common for diary studies, interviews and questionnaires to be conducted in the respondents' L1 but the language of verbal reports has varied from study to study. Key issues to consider are:

1. The respondents' L2 proficiency. If you are gathering data from respondents with low language proficiency you might find it more productive to gather the data in their L1. This will enable you to probe for more sophisticated answers. Indeed, if you conducted the interview or verbal report in the respondents' L2 you might worry that the depth of responses was adversely affected by the respondents' L2 proficiency (regardless of their level of ability in their L2).
2. Your own ability in the respondents' L1. There are contexts in which the researcher does not speak the respondents' L1 well enough or at all. This could be because the researcher has not learned that language sufficiently well to conduct interviews or verbal protocol procedures with study participants. In such circumstances you might wish to work with a native speaker of the respondents' L1 who could gather the data on your behalf. However, this might not be a practical solution in cases where the study participants come from a wide variety of language backgrounds.

For example, in Allwright & Banerjee's (1997) study the 38 respondents represented 20 different nationalities, and spoke a range of 13 different languages. It would have been impractical to arrange for these respondents to receive questionnaires in their L1 and to be interviewed in their L1. Indeed, this might have further complicated the interpersonal considerations that arose with using two interviewers working separately to gather the data (see 5.3, above, for more discussion).

3. The cognitive load of performing a task in the L2 but talking about it in the respondents' L1 might affect the processes that you are trying to capture. In such circumstances, you might wish to gather the data in the L2 so that this cognitive load is controlled.

## 7.2 Piloting and trialling

It is important to pilot all the instruments that you use and to train everyone who will be involved in collecting data. Piloting of instrumentation is particularly important when you are gathering data using feedback methods such as questionnaires, observation checklists and interviews. Piloting is usually on a smaller scale than the main data collection phase but must be conducted with a comparable context and with a similar sample group of respondents. The purpose of the piloting stage is to check that the questions or observation prompts are eliciting the data that you are trying to capture and that your respondents understand the wording of the questions. Piloting also gives you feedback on your procedures for gathering the data. For instance, you can use piloting to establish the best time to administer a questionnaire or to check that your instructions and procedures are clear and efficient.

Observer- and interviewer-training is also important for successful data collection. Though the training phase could be combined with the piloting phase it is probably best to conduct observer and interviewer training after the instruments are finalised. As with piloting, training must be conducted in a comparable context to the live data collection context. In the case of observer training the data used can be pre-recorded. Observers can be asked to complete the observation checklist while watching a video recording of a class, test performance or test administration. They can then discuss the notes they have taken, using the video-tape record to discuss the aspects of the lesson, test performance or test administration that they did not capture. This discussion should alert the observers to aspects of the observation context that they should pay particular attention to. It should also familiarise them with the observation instrument. This process can be repeated until you and the observers are confident that they are ready for live data collection.

Interviewer training is rather more complex. Though video-recordings are useful for familiarising interviewers with the interview structure and alerting them to possible pitfalls, it is also important to give interviewers one or two practice interviews. Each practice interview should be recorded so that they can be reviewed. The practice should help the interviewers to internalise the interview structure and should help them to conduct the interview more naturally (with less recourse to notes). The discussion should alert the interviewers to possible pitfalls in their own interviewing style.

## 7.3 Transcribing the data

If you intend to analyse your data using Conversation analysis you will need to adopt the detailed transcription scheme that I described in 3.1. For other types of analysis, however, you need to pick the most appropriate level of detail for your purposes (Silverman, 1993: 124). Silverman also advises that you should adopt transcription conventions that are achievable within your constraints of time and resources (1993: 124). For instance, in her study of the influence of different language proficiency levels upon students' experiences on academic courses, Banerjee (2003) was primarily interested in **what** her respondents said about their study experiences rather than in the nature of the interaction between herself and her research participants. Consequently, she adopted a very simple transcription scheme for her interview data:

| | |
|---|---|
| **,** | pause for breath during a thought. |
| **.** | pause at the end of a thought. |
| **?** or **(?)** | a question either to self or to other speaker. |
| **!** or **(!)** | particular emphasis placed during utterance. |
| **mmm** or **um** | sounds usually indicating thinking. |
| **mhmm** | sound indicating agreement. |
| **…** | pause of any length. |
| **[unclear]** | speech that could not be decoded. |
| **[ ]** | action/event occurring or co-occurring e.g. [laughs] = laughter from speaker; [tape ends] = end of side A or recording. Also used for my own clarifications of what is being referred to e.g. [1998/199 class] clarifies which MBA class the speaker is referring to when she says 'class'. |

(Banerjee, 2003: Appendix 5J)

Banerjee (2003) captured repetitions and fillers (such as 'you know') but did not need, for her purposes, to capture the pace of delivery or pronunciation of her interview respondents. Similarly, she did not attempt to capture overlapping speech as this was not relevant to her analysis. Instead, she used standard punctuation (e.g. commas and full stops) to indicate natural pauses in delivery. However, she felt that non-verbal behaviour (e.g. laughter or a pause to check or read from a file) was relevant to her analysis so this was noted. Banerjee (2003) developed this transcription scheme iteratively while simultaneously analysing a subset of her data. This helped her to develop a transcription scheme with an appropriate level of detail.

It is important to note, however that you may not need to transcribe all (or perhaps any) of your data. In some cases it may be enough to listen to the recordings several times, taking detailed notes and transcribing only the most illuminating or colourful extracts. You can then report the broad themes thrown up by the analysis, flavouring it with appropriate extracts.

### 7.4 Triangulation of data sources

The perennial question that needs to be answered in any study is whether the data that was gathered was a true reflection of the reality it was intended to study. The triangulation of data sources refers to the gathering of data about a particular event or context from a number of different angles. If the data gathered from each of these perspectives or angles all suggests the same interpretation or conclusions, this can help to corroborate your claims.

Triangulation can be achieved in a number of ways. First, you could use two or more methods to collect your data from your respondents. For example, in their study of the effect of the TOEFL test on teaching Alderson & Hamp Lyons (1996) first interviewed the teachers and then followed this up by observing the teachers in both TOEFL-preparation and non-preparation classes. Another way of triangulating your data is to collect data from more than one source. For instance, if you were exploring the appropriacy of the content of a test you might ask three different groups to provide judgements – test developers, teachers and test-takers.

In addition to corroborating your analysis, triangulation provides opportunities for probing certain aspects of your data in more depth such as when you follow up a questionnaire with in-depth interviews with a sub-set of your sample.

## 7.5 Analysing the data

Arguably, good analysis begins with the appropriate and accurate storage and transcription of data. Dey (1993: 74) argues that "[g]ood analysis requires efficient management of one's data". It is important, therefore, that data is stored in a format that allows you to search it easily and to compare different transcripts. This can be done manually by using a system of filing cards and annotated transcripts. You might begin by highlighting and annotating your transcripts with themes and codes. Quotations could be transferred onto a filing card and labelled with the theme that they represent. If a quotation represents more than one theme, you could either complete two filing cards (one for each theme) or you could devise a cross-referencing system.

The manual approach is easy to transport but clearly very labour intensive and could involve a lot of repetitive work. Therefore, researchers are increasingly using electronic tools. There are a number of software packages that support qualitative data analysis, some of which interface with statistical tools like SPSS (see 5.2, above). Two examples of these are Atlas-ti (http://www.atlasti.de) and QSR NUD*IST (http://www.qsrinternational.com/products/ productoverview/product_overview.htm). These programmes help researchers to apply multiple codes to their data and to build theories about how the codes might be related to one another.

Nevertheless, data analysis tools cannot actually perform the analyses. They simply support the analysis being done. This phase of the research process can be very daunting for, as Denzin argues, data analysis "is a complex, reflexive process" (1998: 316) that involves making sense of the data and then representing it in a coherent way that explains the interpretation taken. The first question that must be addressed, however, is how to approach the coding. Indeed, the assembled data can be very overwhelming (cf. Buck, 1994 and Feldman, 1995). It is important, therefore, to find a way into the data perhaps by first looking for answers to your initial research questions or by inspecting your data for themes that have emerged from your review of the literature. For instance, Buck (1994) used his initial research hypotheses as his starting point when analysing his data. Another alternative would be to adopt the Grounded Theory approach (Strauss & Corbin, 1998). Grounded theory refers to theory that is data driven. It demands that researchers should look for patterns in the data rather than attempting to impose a pre-existing theory or explanation.

Regardless of the approach you adopt, however, Brown & Rodgers (2002) emphasise the importance of coding data in a way that helps you to reveal its underlying patterns. While the coding categories that emerge are usually specific to the research being conducted (e.g. Alderson (1990) coded for reading processes), Brown & Rodgers suggest three important considerations:

i. Are the code-categories clear and unambiguous?
ii. Is the coding scheme reliable? Will alternative analysts code data in the same way?
iii. Do the results of coding lead to useful analyses?

(2002: 64)

## 7.6 Validity, reliability, generalisability

This focus of this chapter has been on validity and how to establish that a test is valid. It follows, therefore, that the methods used to establish test validity should themselves be valid. As the Manual argues, "[i]n an empirical validation, the data have to be analysed and interpreted thoughtfully and with full awareness of possible sources of uncertainty and error" (Council of Europe, 2003: 99). Indeed, Maxwell (1992: 279) warns that the legitimacy of qualitative research is threatened when it cannot consistently produce valid results. Indeed, this is true of all research but the problem is perhaps more acute for qualitative research because of its interpretive nature.

Alderson & Banerjee (2001) provide a practical approach to instrument validation. Drawing on the procedures already used in test validation, they suggest a number of simple measures that can reveal the transparency and clarity of the language used in the instrument as well as whether the options offered (e.g.

never, sometimes, often) mean the same thing to different users and whether there are any gaps in the instrument. These measures include:

1. Reliability measures such as internal consistency (split-half measures), response stability (test-retest), and consistency within and between raters.

   Internal consistency measures are useful if you are gathering information about stake-holders attitudes towards the test or the effect of the test on their attitudes towards the language being tested. Such questionnaires typically include more than one item that is measuring the same issue. One would expect respondents to give comparable responses to items that are measuring the same issue.

   Response stability is also most useful when validating questionnaires. Respondents can be asked to complete the questionnaire on day one and then again the next day. Alderson (1992) used this method in his study of the effect of an exchange programme upon students' language proficiency. He warned, however, that response stability must be checked item by item rather than by aggregating responses across items. Response stability measures can also be used in a modified form for interviews. In this case the respondents would be interviewed twice on consecutive days. The researcher and the respondent could then review the interviews together. Differences in the responses to each question could be discussed in order to establish whether the change in the response had been prompted by a difference in the phrasing of the question or the approach taken by the interviewer. It is important to recall, however, that interviews are a social event and some variability is to be expected and must be borne. The key, nevertheless, lies in minimising the effect of the interpersonal interaction between interviewer and respondent upon the data that is collected.

   Establishing consistency within and between raters is important for the use of checklists and analytical frameworks. It is also important in all aspects of language analysis. To establish intra-rater consistency, judges will need to complete the data collection instrument twice. The stability of the judges' decisions could then be inspected. For instance, if a judge were applying Bachman & Palmer's (1996) Task characteristics framework to a reading test, he/she would need to complete the judgement on two separate occasions (perhaps on consecutive days). His/her judgements could then be compared for consistency. Similarly, if the judge was using a classroom observation instrument he/she would need to complete the checklist twice. In this case, the consistency check would have to be carried out with a video-recorded class. Similar procedures could be applied to establish intra-rater consistency. In this case the completed assessments/observations of two or more different judges would be compared. In both cases, it would be important to interview the judges as well in order to explore inconsistencies in the judgements. It will be important to establish whether any inconsistencies that occur have been caused by changes in the judges' interpretation of what they have been seeing (perhaps a training issue) or by problems with the wording of the instrument.

2. Validity measures such as investigations of content relevance and coverage, and of interpretations of question wording.

   Investigations of content relevance and coverage are useful for questionnaires, checklists, task characteristics frameworks and interviews. For instance, if you were designing a speaking test observation checklist similar to the one designed by O'Sullivan et al. (2002), you could ask expert judges (item writers, teachers etc) to discuss what they would expect the test to include and what they would expect a validation checklist to include. You could then show the judges the actual checklist and ask them to assess the content relevance and coverage of the instrument. This discussion might reveal areas of construct under-representation and/or construct irrelevant items.

Explorations of the way in which respondents interpret questions will help you to establish whether the respondents have understood the question in the way that it was intended. This is particularly useful for questionnaires and interviews but might also be useful in the validation of checklists. In the latter case you want to ensure that your observers have understood the categories they need to gather data under. Clearly, one way of exploring how respondents interpret interview and questionnaire prompts or observation categories would be to conduct a verbal protocol (e.g. Alderson, 1992 and Block, 1998). Alderson (1992) had designed a questionnaire to explore the benefits for their language proficiency of an exchange programme for university students across Europe. Alderson (1992) used verbal protocols to explore respondents' interpretations of the questionnaire items. Block (1998) replicated this methodology in his validation of an end-of-course evaluation form. Block (1998) was particularly interested whether different respondents interpreted the questionnaire items in the same way and also in whether they interpreted the points on the 1-5 rating scale in the same way. Block (1998) reported a high degree of variability in the respondents' interpretations of the questionnaire items and the rating scale. This had implications for Block's ability to aggregate and interpret the questionnaire results.

Foddy (1993:186) suggests an alternative approach to verbal reports, where respondents are asked to rephrase the questions in their own words. You could then analyse these reformulations according to four parameters:

i. fully correct - leaving out no vital parts
ii. generally correct – no more than one part altered or omitted
iii. partially wrong – but indicating that the respondent knew the general subject of the question
iv. completely wrong and no response

(Foddy, 1993: 186)

This approach is interesting because it could be less time-consuming than verbal protocols and might also circumvent some of the problems associated with gathering verbal report data (see 2.1 for this discussion).

Apart from validity, another area of concern for qualitative research is our ability to generalise from the study sample to the wider population. The key to this lies in the representativeness and size of our sample. However, as Lazaraton (1995: 465) argues, even if a result has been established on a large, randomly selected sample, this does not guarantee that it will apply to a particular individual. More importantly, Cronbach (1975) argues that all analyses are context bound:

> Generalizations decay. At one time a conclusion describes the existing situation well, at a later time it accounts for rather little variance, and ultimately is valid only as history.
>
> (Cronbach, 1975: 122)

Cronbach suggests instead that, instead of focusing upon the generalisability of our results, we should make clear the effect of context upon the results, giving "proper weight to local conditions" (1975: 125). He also believes that "any generalization is a working hypothesis, not a conclusion" (1975: 125). These comments are important for they remind us that research is systematic, observant and reflective. It is important to be persuasive and to be seen to have paid attention all the data and to have attempted to account for all of it (rather than just the convenient bits of it). They also highlight the importance of the results having "explanatory power" (Strauss & Corbin, 1998: 267).

## 8. Conclusion

This section of the reference supplement has provided an overview of the range of qualitative methods available for investigating test quality. It has demonstrated the variety of options available and explained the key features of each. In addition, examples of research using the methods have been provided so that you can see how specific qualitative methods have been implemented. The final sub-section (7.1 – 7.6) has also addressed more general issues such as transcription and triangulation of data. The key messages of this section have been:

1. Qualitative methods have enormous power to explain and augment the statistical evidence we might gather to establish test quality.
2. Many of the methods are complimentary and can be used for the triangulation of data sources.
3. It is important to safeguard the validity and generalisability of your data collection methods in order to legitimise the inferences you draw from them.

**References**


Alderson, J.C. (1990) Testing reading comprehension skills (part two): getting students to talk about taking a reading test (a pilot study), Reading in a Foreign Language, 7(1), 465 – 503.

Alderson, J.C. (1992) Validating questionnaires, CRILE Working Papers 15, Lancaster: Department of Linguistics and English Language, Lancaster University.

Alderson, J.C. (2000) Assessing reading, Cambridge: Cambridge University Press.

Alderson, J.C. and Banerjee, J. (2001) Impact and washback research in language testing, in Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. and O'Loughlin, K. (eds.) Experimenting with uncertainty: essays in honour of Alan Davies, Cambridge: University of Cambridge Local Examinations Syndicate, 150 – 161.

Alderson, J.C. and Hamp-Lyons, L. (1996) TOEFL preparation courses: a study of washback, Language Testing, 13(3), 280 – 297.

Alderson, J.C. and Pižorn, K. (eds.) (2004) Constructing school leaving examinations at a national level – meeting European standards, Ljubljana, Slovenia: The British Council & Državni izpitni center.

Alderson, J.C. and Wall, D. (1992) The Sri Lankan O-Level evaluation project: fourth and final report, Lancaster University.

Allan, A. (1992) Development and validation of a scale to measure test-wiseness in EFL/ESL reading test takers, Language Testing, 9, 101 – 122.

Allen, P., Fröhlich, M. and Spada, N. (1984) The Communicative Orientation of Language Teaching: An Observation Scheme, in Handscombe, J., Orem, R.A. and Taylor B.P. (eds) On TESOL '83: The Question of Control, Washington D.C.: TESOL.

Allwright, J. and Banerjee, J. (1997) Investigating the accuracy of admissions criteria: a case study in a British university, CRILE Occasional Report 7, Lancaster: Lancaster University, Department of Linguistics and Modern English Language.

Arnaud, P.J.L. (1984) The lexical richness of L2 written productions and the validity of vocabulary tests, in Culhane, T., Klein-Braley, C. and Stevenson, D.K. (eds.) Practice and problems in language testing, Occasional Papers No. 29, Department of Language and Linguistics, University of Essex, 14 – 28.

Bachman, L.F. (1990) Fundamental considerations in language testing, Oxford: Oxford University Press.

Bachman, L.F., Davidson, F., Ryan, K. & Choi, I.C. (1995) An investigation into the comparability of two tests of English as a foreign language. The Cambridge-TOEFL comparability study, Cambridge: Cambridge University Press.

Bachman, L.F. and Palmer, A.S. (1996) Language testing in practice, Oxford: Oxford University Press.

Banerjee, J.V. (1996) UCLES Report: The design of the classroom observation instruments, unpublished report commissioned by the University of Cambridge Local Examinations Syndicate (UCLES), Cambridge: UCLES.

Banerjee, J.V. (1999) Being an insider – a double-edged sword?, paper presented at the BAAL/ CUP Seminar 1999, Lancaster, U.K.

Banerjee, J.V. (2003) Interpreting and using proficiency test scores, unpublished PhD dissertation, Lancaster University.

Block, D. (1998) Exploring interpretations of questionnaire items, System, 26, 403 – 425.

Borg, W.R., & Gall, M.D. (1983) Educational Research: An Introduction (4th ed.) New York: Longman Inc.

British National Corpus, maintained by the Oxford University Computing Services (http://www.natcorp.ox.ac.uk/)

Brown, A. (1993) The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese, Language Testing, 10(3), 277-303.

Brown, A. (2003) Interviewer variation and the co-construction of speaking proficiency, Language Testing, 20(1), 1 – 25.

Brown, A. and Hill, K. (1998) Interviewer style and candidate performance in the IELTS oral interview, in Woods, S. (ed.) IELTS Research Reports: Volume 1, Sydney, ELICOS, 173 – 191.

Brown, A., and Lumley, T. (1997) Interviewer variability in specific-purpose language performance tests, in Huhta, A.., Kohonen, V., Kurki-Suonio L. and Luoma S. (eds.) Current Developments and Alternatives in Language Assessment, Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä, 137 - 150.

Brown, J.D. and Rodgers, T.S. (2002) Doing second language research, Oxford: Oxford University Press.

Buck, G. (1994) The appropriacy of psychometric measurement models for testing second language listening comprehension, Language Testing, 11(2), 145 – 170

Clapham, C. (1997) IELTS Research Report 3, The British Council, the University of Cambridge Local Examinations Syndicate and the International Development Project for Australian Universities and Colleges, Cambridge.

Clapham, C. (1996) The development of IELTS: A study of the effect of background knowledge on reading comprehension, Cambridge: Cambridge University Press.

Cohen, A (1984) On taking language tests: what the students report, Language Testing, 1(1), 70 – 81

Cohen (1994) English for academic purposes in Brazil: the use of summary tasks, in Hill, C. and Parry, K. (eds.) From testing to assessment: English as an international language, London: Longman, 174 – 204.

Council of Europe (2003) Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF), Strasbourg: Language Policy Division, Council of Europe

Cresswell, J. W. (2003) Research design: qualitative, quantitative, and mixed methods approaches (2nd Edition), Thousand Oaks, CA: Sage Publications.

Cronbach, L. (1975) Beyond the two disciplines of scientific psychology, American Psychologist, 30, 116 – 127.

Denzin, N.K. (1998) The art and politics of interpretation, in Denzin, N.K. and Lincoln, Y.S. (eds)
Strategies of qualitative inquiry, Thousand Oaks, CA: Sage Publications, Inc., 313 - 344.

Dey, I. (1993) Qualitative data analysis: a user-friendly guide for social scientists, London: Routledge.

Dörnyei, Z. (2003) Questionnaires in second language research: construction, administration and processing, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Eignor, D., Taylor, C., Kirsch, I. and Jamieson, J. (1998) Development of a scale for assessing the level of computer familiarity of TOEFL examinees, TOEFL Research Reports 60, Princeton, NJ: Educational Testing Service.

Feldman, M.S. (1995) Strategies for interpreting qualitative data, Qualitative research methods series 33, Thousand Oaks, CA: Sage Publications, Inc.

Foddy, W. (1993) Constructing questions for interviews and questionnaires: theory and practice in social research, Cambridge: Cambridge University Press.

Fulcher, G. (2003) Testing second language speaking, Cambridge: Polity Press.

Gass, S.M. and Mackey, A. (2000) Stimulated recall methodology in second language research, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Ginther, A. and Grant, L. (1997) Effects of language proficiency and topic on L2 writing, paper presented at the annual conference for Teachers of English to Speakers of Other Languages, Orlando, Florida, March 1997.

Green, A. (1998) Verbal protocol analysis in language testing research: a handbook, Studies in Language Testing 5, Cambridge: University of Cambridge Local Examinations Syndicate.

Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B. and Kantor, R. (1996) A study of writing tasks assigned in academic degree programs, TOEFL Research Report No. 54, Princeton, NJ: Educational Testing Service.

Halvari, A. and Tarnanen, M. (1997) Some aspects on using qualitative procedures to ensure comparability across languages within a testing system, in Huhta, A., Kohonen, V., Kurki-Suonio, L. and Luoma, S. (eds.), Jyväskylä: Centre for Applied Language Studies, University of Jyväskylä, 127 – 136.

Hambleton, R. (2001) Setting performance standards on educational assessment and criteria for evaluating the process, in Cizek, G. (ed.) Setting performance standards: concepts, methods and perspectives, Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 89 – 116.
Heritage, J. (1984) Garfinkel and ethnomethodology, Cambridge: Polity.
Herington, R. (1996) Test-taking strategies and second language proficiency: is there a relationship?, unpublished MA dissertation, Lancaster University.
Horák, T. (1996) IELTS impact study project, unpublished MA assignment, Lancaster University.
Hutchby, I. and Wooffitt, R. (1998) Conversation Analysis: An Introduction, Cambridge: Polity Press.
Kane, M.T. (2001) So much remains the same: conceptions and status of validation in setting standards, in Cizek, G.J. (ed.) Setting performance standards: concepts, methods and perspectives, Mahwah, NJ: Erlbaum, 53 – 88.
Kelly, P. (1991) Lexical ignorance: the main obstacle to listening comprehension with advanced foreign language learners, IRAL, 24, 135 – 149.
Kim, S. (2004) A study of development in syntactic complexity by Chinese learners of English and its implications on the CEF scales, unpublished MA dissertation, Lancaster University.
Kirsch, I., Jamieson, J., Taylor, C. and Eignor, D. (1998) Computer familiarity among TOEFL examinees, TOEFL Research Reports 59, Princeton, NJ: Educational Testing Service.
Kormos, J. (1999) Simulating conversations in oral-proficiency assessment: a conversation analysis of role play and non-scripted interviews in language exams, Language Testing, 16(2), 163 – 188.
Laufer, B. (1991) The development of L2 lexis in the expression of the advanced language learner, Modern Language Journal, 75, 440 – 448.
Laufer, B. and Sim, D.D. (1985) Measuring and explaining the reading threshold needed for English for Academic Purposes texts, Foreign Language Annals, 18, 405 – 411.
Lazaraton, A. (1995) Qualitative research in Applied Linguistics: a progress report, TESOL Quarterly, 29(3), 455 – 472.
Lazaraton, A. (2002) A qualitative approach to the validation of oral language tests, Cambridge: UCLES/CUP.
Leech, G., Rayson, P. and Wilson, A. (2001) Word frequencies in written and spoken English: based on the British National Corpus, London: Longman.
Li, W. (1992) What is a test testing? An investigation of the agreement between students' test taking processes and test constructors' presumption, unpublished MA Thesis, Lancaster University.
Low, G. (1996) Intensifiers and hedges in questionnaire rating scales, Evaluation and Research in Education, 2(2), 69 – 79.
Lumley, T. (2002) Assessment criteria in a large-scale writing test: what do they really mean to the raters?, Language Testing, 19(3), 246 – 276.
Marinič, Z. (2004) Test quality, in Alderson, J.C. and Pižorn, K. (eds.) (2004) Constructing school leaving examinations at a national level – meeting European standards, Ljubljana, Slovenia: The British Council & Državni izpitni center, 179 – 192.
Maxwell, J.A. (1992) Understanding and validity in qualitative research, Harvard Educational Review, 62(3), 279 – 300.
Mishler, E.G. (1986) Research interviewing: context and narrative, Cambridge, Mass.: Harvard University Press.
Moser, C.A. and Kalton, K. (1971) Survey Methods in Social Investigation (2nd ed.) Aldershot, Hants.: Gower.
O'Loughlin, K. (1995) Lexical density in candidate output, Language Testing, 12(2), 217-237.
O'Loughlin, K. (2002) The impact of gender in oral proficiency testing, Language Testing, 19(2), 169 – 192.
O'Sullivan, B, Weir, C.J. and Saville, N. (2002) Using observation checklists to validate speaking tasks, Language Testing, 19(1), 33 – 56.
Oppenheim, A.N. (1992) Questionnaire design, interviewing and attitude measurement, London: Pinter Publishers Ltd.

Potter, J. (1996) Discourse analysis and constructionist approaches: theoretical background, in Richardson, J. (ed.) Handbook of qualitative research methods for psychology and the social sciences, Leicester: BPS, 125 – 140.
Potter, J. (1997) Discourse analysis as a way of analysing naturally-occurring talk, in Silverman, D. (ed.) Qualitative research: theory, method and practice, London: Sage Publications Inc., 144 – 160.
Potter, J. and Wetherall, M. (1987) Discourse and social psychology: beyond attitudes and behaviour, London: Sage Publications.
Purves, A.C., Soter, A., Takala, S. and Vähäpassi, A. (1984) Towards a domain-referenced system for classifying assignments, Research in the Teaching of English, 18(4), 385 – 416.
Read, J. (2001) Assessing vocabulary, Cambridge: Cambridge University Press.
Sarig, G. (1987) High level reading tasks in the first and in a foreign language: some comparative process data, in Devine, J., Carrell, P.L. and Eskey, D.E. (eds) Research in reading in English as a second language, Washington, D.C.: TESOL, 105 – 120.
Shohamy, E. (1994) The validity of direct versus semi-direct oral tests, Language Testing, 11(2), 99-123.
Shohamy, E., Donitsa-Schmidt, S. and Ferman, I. (1996) Test impact revisited: washback effect over time, Language Testing, 13(3), 298 – 317.
Silverman, D. (1993) Interpreting qualitative data: methods for analysing talk, text and interaction, London: Sage Publications, Ltd.
Silverman, D. (2001) Interpreting qualitative data: methods for analysing talk, text and interaction, London: Sage Publications Ltd.
Stimson, G.V. (1986) Viewpoint: Place and space in sociological fieldwork, Sociological Review, 34(3), 641 – 656.
Strauss, A. and Corbin, J. (1998) Basics of qualitative research: Techniques and procedures for developing grounded theory, Thousand Oaks, CA: Sage Publications, Inc.
Symon, G. (1998) Qualitative research diaries, in Symon, G and Cassell, C. (eds.) Qualitative methods and analysis in organisational research: a practical guide, London: Sage Publications Inc., 94 – 117.
Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I. (1998) The relationship between computer familiarity and performance on computer-based TOEFL test tasks, TOEFL Research Reports 61, Princeton, NJ: Educational Testing Service.
ten Have (1999) Doing conversation analysis, London: Sage Publications Ltd.
Wall, D. and Alderson, J.C. (1993) Examining washback: the Sri Lankan impact study, Language Testing, 10(1), 41-69.
Weigle, S.C. (1994) Effects of training on raters of ESL compositions, Language Testing, 11(2), 197 – 223.
Weigle, S.C. (2002) Assessing writing, Cambridge: Cambridge University Press.
Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse, Language Testing, 14(1), 85 – 106.
Winetroube, S. (1997) The design of the teachers' attitude questionnaires, unpublished report commissioned by the University of Cambridge Local Examinations Syndicate (UCLES), Cambridge: UCLES.
Wolfe-Quintero, K., Inagaki, S. and Kim, H.Y. (1998) Second language development in writing: measures of fluency, accuracy and syntactic complexity, Hawaii: University of Hawaii.
WordSmith Tools, developed by Mike Scott (http://www.oup.com/elt/global/isbn/6890/)

# Section E

# Generalizability Theory

**N.D. Verhelst**
**National Institute for Educational Measurement (Cito)**
**Arnhem, The Netherlands**

This report contains four sections. The first two sections give a non-technical introduction into generalizability theory (G.T.). In the third and fourth sections the same problems are treated in a somewhat more technical way.

It is interesting to notice that a very basic term of Classical Test Theory is not well defined. In explaining the concept of measurement error in the manual and in Section C, reference was made to repeated observations under 'similar' conditions, but 'similar' was not defined precisely. An often used example of a cause of (negative) measurement error is the noise in the testing environment. But suppose a student is only tested in his school. If the school is located in a very noisy environment, and if noise has indeed a negative impact on test performance, it will maintain this negative impact (because it is constant) on retesting or administration of a parallel test. In such a case the noise is to be considered systematic influence, and its impact cannot be conceived of as measurement error; it will lower the true score of the student. If one wants to have an idea about the magnitude of the negative impact of noise, one will have to conduct an experiment to find out. (A good experiment would be to administer the test to two equivalent samples in two different conditions - quiet and noisy - and to compute the differences between the average test scores in both conditions.)

An important way of controlling for such systematic effects is the standardization of the test administration, which, for example in the case of a listening test, could prescribe that headphones are to be used. It is, however, impossible to control for all possible sources of disturbance. A typical example occurs when the item scores have to be determined by means of ratings by some rater, e.g., by the teacher. Some teachers are more lenient than others, and if a candidate happens to get (always) a lenient teacher his true score will get higher than with a harsh teacher.

To find out whether differences in leniency of the raters make a lot of difference in the scores, one has to investigate this in a special study. Such an investigation can be supported by a psychometric theory that is able to quantify these differences. A theory which is especially created for this purpose is **Generalizability Theory** (G.T.), which was published in a series of articles in the 1960s, and as a book in 1972[1].

In this theory, measurements are described in terms of the conditions where they are observed. A set of conditions that belong together is called a **facet**. In this way, 'items' is a facet of the measurement procedure. The measurement object is usually the person who is tested, and the basic observations are usually collected by observing all persons in the sample with all items in the test, i.e., persons are crossed with a number of conditions (specific items) from the facet 'items', and such a set-up is called a single-facet crossed design. But sometimes more facets are involved: it is possible that the answers by persons to items are to be rated by a number of raters. If the answer of each person to each item is rated by each rater (from a well-defined group of raters), we have a crossed two-facet design: the facets are 'items' and 'raters'. (At least, this is the description one usually finds in textbooks on G.T.; we will come back to this example in later sections.)

[1] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley. A more recent and more accessible book is: R.L. Brennan, (2001). *Generalizability Theory*. New-York: Springer-Verlag.

Two important views can be taken with respect to the conditions of the facets: they can be considered as **fixed** or as **random**. (The principle is the same as in the analysis of variance; in generalizability theory the concepts of analysis of variance are used throughout.). In the fixed case, the conditions are taken as they are: if items are considered as fixed, this means that we are interested in the very items that are part of the test. In the random case, the conditions are considered as a random sample from a much bigger collection of conditions. Such a collection is called a **universe**.

Two conditions will be considered in detail: the one-facet crossed design (persons by items) and the two-facet crossed design (persons by items by raters).

**E.1. The persons by items design**

Consider the Classical Test Theory as it was presented in the manual and in appendix C. The true score was defined as the average or expected score **on the very same test** (under repeated administrations). This means that the items are considered as fixed. But we could also draw a new random sample of (the same number of) items from the universe of items at each administration, and then compute the expected score of the person over these test administrations. This expectation is called the **universe score** of that person. It is clear from this that a particular observed score will deviate from the universe score not only because of measurement error but also because of the factual composition of the test that has been used: if the items in the test happen to be relatively easy, the observed score will probably be higher than in case the items in the (randomly composed) test happen to be relatively difficult. This means that in the random view, the difficulty of the items now has to be considered as an extra source of variance (of the observed scores).

But generalizability theory considers yet an extra source of variance. To see this, imagine that the basic observations are arranged in a two-way table, the rows associated with the persons and the columns with the items. A particular cell contains the observed item score for the person of that row on the item of that column. Then four sources of variability are (in principle) distinguished:
- the persons (having different universe scores);
- the items (having different difficulties);
- the interaction between persons and items (John is especially good on items 1 and 2, while Mary performs especially poorly on item 3 but especially well on item 17, etc.);
- the measurement error.

With each source of variability there is a corresponding variance, and the theory makes it clear that the total variance in the two-way table is the sum of those four variances. These four variances are called **variance components**. The main purpose of the analysis of the two-way table is to estimate (from a single sample) these variance components. But since there is only one observation per cell, it is impossible to estimate the interaction component and the error component separately (interaction and error are confounded); only their sum can be estimated. This sum is usually called the residual component. The variance components are usually estimated by techniques of analysis of variance.

If the variance components are known, then some interesting correlations can be predicted. In the case of a two-way design (one facet), two correlations are interesting:

1. The correlation between the actual scores of the persons and their scores on an independent replication **with the same items.** Notice that this correlation is the reliability of the test (see Section C). Unfortunately, to predict this correlation one has to know the interaction component and the error component, and since they cannot be estimated separately, one has to be satisfied with an approximation. The approximation used in G. T. happens to be identical to Cronbach's alpha, and it can be shown that this approximation equals the true coefficient only when the interaction component is zero.
2. The correlation between the actual scores of the persons and their scores on another test (with the same number of items). This latter test has to be randomly drawn from the universe of items. It will be clear that in this case the items are considered as random.

These two correlation coefficients are called generalizability coefficients. More technical details and questions of interpretation are discussed in Section E.3.

## E.2 The persons by items by raters design

In the one-facet design, it is usually not difficult to construct the two-way table needed for estimating the variance components, since the data (the responses to the items) are commonly collected on the calibration sample. Moreover, to get a stable estimate of the variance components one needs a reasonable number of persons and a reasonable number of items, but in the usual procedures of internal validation this is no problem (40 items is a reasonable number). If one uses a second facet (raters) in a crossed design, things become more complicated: for the analysis one needs a **three-way** table, which one can consider as a piling up of a number of two-way tables. Each two-way table (a layer in the pile) has the same structure as in the one-facet design, but corresponds to a single rater. To estimate the variance components, one needs at least two layers, but to have stable estimates, one needs more. Suppose that a test constructor can use ten raters. Usually, it is a lot of expensive work to have all raters rate the responses of all persons in the sample to all items in the test. Therefore, one uses only a subset of persons (drawn at random from the calibration sample), and (if there are many items) a subset of items. For this (these) subset(s), all available raters rate all answers in order to have a completely filled three-way table. (Incomplete three-way tables are very difficult to handle when estimating variance components[2].) This special data collection together with its analysis to estimate the variance components is called a G-study. It is good practice to carry out a G-study when using raters.

In the two-facet crossed design there are eight variance components: three components associated with main effects, three first order interactions, one second order interaction and one error component. The three main components are associated with persons, items and raters, respectively. The raters component refers to different degrees of leniency of the raters. The three first order interaction terms are listed below, together with a typical example to illustrate the ideas:
- person-item interaction: John is especially good on item 1;
- person-rater interaction: Rater A is especially lenient with Mary;
- item-rater interaction: Rater A is especially lenient when rating item 1.

A second-order interaction then occurs when rater A is especially harsh with John when rating item 1. Since in the three-way table, we have only one observation per cell, the second order interaction and the error are confounded, so that their variance components cannot be estimated separately; only their sum (the residual component) can.

At this point, however, a serious problem with respect to the correct interpretation of the variance components must be noted, because the three-way table (students by items by raters) may come about in two quite different ways, which we illustrate by the following example. A number of young musicians has to play a number of fragments from different composers, and each performance has to be scored by a number of jury members. The fragments play the role of items; the jury members act as raters. The whole contest may be arranged (at least conceptually) in two different ways. Firstly, it may be that each student plays each fragment only once in the presence of the whole jury (which is what usually will happen); but, secondly, it might well be that each student plays all fragments in turn for each jury member. In both cases the data collection will be arranged in a similar three-way table, and in both cases the analysis will be carried out in an identical way, but the interpretation of the variance components is different. In the former case the jury members all judge the very same performances, and it may happen that a single performance (of John, say, playing a fragment of Brahms) is incidentally quite poor, which means the judged performance may be infected by a negative measurement error, but this will lead probably to a low score given by all raters. This means, in more general terms, that the scores given by the raters will be correlated. Because of this dependence on the same measurement error in a single performance it is better to conceive of such a set-up as a **nested**

[2] Special software to estimate variance components in the two facet design with missing observations can be obtained on request from Ton.Heuvelmans@citogroep.nl

**design (**the raters are nested under the student-item combinations; even if for all student-item combinations the same set of raters has been used[3]). In the latter case, where each student plays each fragment (independently) for each rater, the measurement errors in the performances are assumed to be independent, and we have a genuine crossed design. Of course such a set-up will probably never occur in educational settings, and it is remarkable that the nested design (which is the usual way of data collection) has been treated in G.T. as if it were a truly crossed design. A more technical treatment of this problem will be given in Section E.4.

As an example, the results of a G-study are given for a number of countries which participated in the first cycle of PISA[4]. The items were reading items (in the Mother Tongue) used for a scale that was called Retrieving Information. The number of students participating in the G-study varied between 48 and 72 (depending on the country), the number of items is 15 and the number of raters is 4. See Table E.1. Notice that in this case the students answered each item only once. In a G-study, the numerical values of the variance components are not important, only their relative contributions to the total variance matters. Therefore, one usually reports the different components as a percentage of the total variance. This is done in Table E.1: the numbers in each row add up to 100.

Table E.1. Variance components in the first cycle of PISA for a reading scale (expressed as a percentage of the total variance)

| | Students | Items | Raters | S x I | S x R | I x R | residual |
|---|---|---|---|---|---|---|---|
| Australia | 22.40 | 19.01 | -0.02 | 50.36 | 0.01 | 0.22 | 8.01 |
| Denmark | 13.24 | 24.56 | 0.01 | 54.22 | 0.16 | 0.25 | 7.56 |
| England | 14.79 | 22.14 | 0.00 | 59.71 | 0.01 | 0.00 | 3.35 |
| Finland | 18.97 | 18.30 | 0.02 | 55.93 | -0.11 | 0.07 | 6.81 |
| Norway | 15.66 | 17.79 | 0.00 | 61.43 | 0.21 | 0.17 | 4.74 |

A number of interesting observations can be made fromTable E.1. An extensive discussion can be found in Section E.4. We make only three observations here:

1. Two numbers in the table are negative. Although variances cannot be negative, their estimates can, which usually indicates that the true variances are near zero. It is customary to treat small negative values as zero.
2. The three shaded columns involve the raters: one as a main effect and two in interaction with either students or items. We see that in all three columns the contributions to the total variance are very small, and for all practical purposes negligible. This result was the basis for the decision taken to let the items be rated by a single rater (for all students not involved in the G-study). In Section E.4, some critical remarks on this decision will be made. For now, it is important to realize that the three shaded columns point to the almost complete absence of systematic rater effects: there are no systematic overall differences in leniency (the main effect component is almost zero), and there are no systematic interactions of raters with students and items. The low student-rater-interaction component is to be expected, since the students came from a national sample in each country and were unknown to the raters; the low rater-item-interaction component means that there were no systematic differences in scoring some of the items, and this may be due to a large part to the careful construction of the rating rules, and to all kinds of measures taken in the PISA project to check that these rules were followed meticulously. But it does not necessarily mean that the agreement between raters was very high, because there might have been **unsystematic** differences between raters which were not taken into account in the PISA study. A detailed discussion of this problem can be found in Section E.4.

[3] The usual way of conceiving nesting is where all instances of one facet are specific to each instance of the other facet. A typical example in educational measurement is the facet schools and the facet students. One says that students are nested within schools, and of course, one assumes that each student belongs to only one school. This unique assignment, however, is not necessary to have a nested design.

[4] PISA stands for Program for International Student Assessment. An overview of the first cycle is given in Knowledge and Skills for Life (2001). More details can be found in PISA 2000, Technical Report (2002), Edited by R. Adams and M. Wu. Both volumes are published by the OECD (Paris).

3. Probably, the most puzzling result in Table E.1 is that the most important variance component is the interaction component between students and items, accounting in each country for more than 50% of the total variance. This result is especially remarkable if it is compared to the residual component which takes relatively modest values in the PISA study. This finding will be commented upon in detail in Section E.4.

As a final comment of this section, it must be emphasized that in collecting the data for a G-study, the raters must work independently of each other. Joint decisions by the raters may look attractive for a number of reasons, but they make the results of a G-study misleading and useless.

### E.3. Generalizability Theory for the one-facet crossed design

Generalizability Theory is a statistical theory which is highly similar to Classical Test Theory, but it is more general. In every theory, the starting point consists of a number of assumptions. Because it is a mathematical theory, these assumptions are usually expressed by mathematical statements (as a formula). The whole of the assumptions is called a **model**. In Section E.3.1 the model will be introduced and some comments will be given on the estimation procedures, while section E.3.2 will be devoted to the use one can make of the results of the analysis.

#### E.3.1 The model

We start with the model for a one-facet crossed design (the facet being 'items'). Variables will have one or two subscripts; the subscript *p* refers to a person (a test taker), and the subscript *i* to an item. The basic observed score is the score of person *p* on item *i*, and this score is denoted by $Y_{pi}$. In the model this score is considered as the sum of five parts, called effects: a general effect, a person effect, an item effect, an interaction effect (between person and item) and a measurement error. Symbolically, this is written as

$$Y_{pi} = \mu + \alpha_p + \beta_i + (\alpha\beta)_{pi} + \varepsilon^*_{pi} \qquad \text{(E.1)}$$

1. The Greek letter $\mu$ symbolizes the general effect. It corresponds to the average item score, where the average is to be understood as the average in the population of persons and across all the items in the universe.
2. The person effect is $\alpha_p$. It is an unknown number and every person in the population can be characterized by a person effect. So, generally speaking, the person effect is a **random variable**, which has some distribution in the population of persons. The population average of the person effects is set to zero. (This is a technical restriction, without which the model cannot 'work'). The practical implication of this restriction is that person effects have to be considered as deviations from the mean: a positive person effect means an effect greater than the average, and a negative effect means an effect smaller than the average. The main problem in the analysis is to estimate the variance of the person effects. This variance will be symbolized as $\sigma^2_\alpha$.
3. The item effect is $\beta_i$. It is a random variable in the universe of items, with mean equal to zero. Its interpretation is completely analogous to that of the person effect. The variance of the item effects is symbolized as $\sigma^2_\beta$.
4. The interaction effect is symbolized as $(\alpha\beta)_{pi}$. A double symbol is used to indicate this interaction; it is not to be understood as a product. (The subscripts *p* and *i* refer to the whole symbol, and therefore the symbol is placed between parentheses.) So, like person effects and item effects, $(\alpha\beta)_{pi}$ is an unknown number which applies to the particular combination of person *p* and item *i*. For every possible combination of a person from the population and an item from the item universe, there is such an interaction effect. The average of these effects is set to zero, and the problem to be faced is the estimation of the variance $\sigma^2_{\alpha\beta}$ of the interaction effects.

5. The measurement error is symbolized as $\varepsilon^*_{pi}$, which is also a random variable with mean zero. Its variance is $\sigma^2_{\varepsilon^*}$.
6. There is one important assumption to be added: it has to be assumed that all random variables in the right hand side of equation (E.1) are independent of each other. Using this assumption, a very useful result from statistics follows directly: the variance of the item scores (across the population of persons and across the universe of items) is just the **sum** of $\sigma^2_\alpha$, $\sigma^2_\beta$, $\sigma^2_{\alpha\beta}$ and $\sigma^2_{\varepsilon^*}$.These four variances are called the **variance components**.

The main purpose of a so-called G-study is to estimate these four variance components. To do so, one needs to administer a **random sample** of items (from the universe) to a **random sample** of persons (from the population). One can store the item scores thus obtained in a rectangular table where the rows correspond to the persons and the columns to the items, and each cell contains the observed item score (obtained by the row person on the column item). If the items are administered only once to each person (as is commonly done), then, unfortunately, it is impossible to estimate the variance components of the interaction and the measurement error separately; only their sum can be estimated. (Technically one says that interaction effects and measurement error are confounded. This confounding can also be deduced from formula (E.1): the interaction effect and the error have the same pair of subscripts. If there were more than one observation for the same person-item combination, the error term (and only this one) would have an extra subscript indicating the replication.) Although we started with a model as detailed as reflected in equation (E.1), we will have to simplify it a little bit. We do so by defining

$$\varepsilon_{pi} = (\alpha\beta)_{pi} + \varepsilon^*_{pi} \qquad \text{(E.2)}$$

The random variable $\varepsilon_{pi}$ is called the **residual effect**, and its variance is called the residual variance.

The main purpose of the analysis to be carried out on the data table is to estimate the person variance ($\sigma^2_\alpha$), the item variance ($\sigma^2_\beta$) and the residual variance ($\sigma^2_\varepsilon$). The analysis can be carried out by standard software like SPSS. An important condition, however, is that the table is complete, i.e., there must not be any empty cells.

### E.3.2 Generalizability coefficients

In the literature on Generalizability Theory, much attention is given to so called generalizability coefficients. These coefficients are in some sense generalizations of the reliability coefficient from classical test theory. The latter, however, can also be expressed as a correlation: the correlation between two series of test scores from parallel tests. In the same way, generalizability coefficients can be considered as correlations between two series of tests scores, but to understand them well, we need to be rather precise as to how both tests are defined.

We need some more notation here. We will indicate the number of items in the test by the capital letter *I*. Of course, when we take decisions on persons, these decisions are based on the test score, and not on individual item scores. To arrive at relatively simple formulae, we will work with mean test scores, and we will denote them by the symbol $Y_p$ defined as

$$Y_p = \frac{1}{I}\sum_{i=1}^{I} Y_{pi}$$

Applying the model (E.1) to the mean test score in the one-facet design gives

$$Y_p = \mu + \alpha_p + \frac{1}{I}\sum_{i=1}^{I}\beta_i + \frac{1}{I}\sum_{i=1}^{I}(\alpha\beta)_{pi} + \frac{1}{I}\sum_{i=1}^{I}\varepsilon^*_{pi} \qquad \text{(E.3)}$$

Now we will distinguish three cases. In the first case we want to have an expression for the correlation between two series of test scores coming from administering the same test twice (and assuming that there are no memory effects, yielding scores on two parallel tests). In the second case two tests are used, one for the first administration and one for the second. The two tests have the same number of

items, but are randomly drawn from the universe of items. In the third case, we want the correlation between two series of test scores in a rather peculiar situation where every person gets his/her own pair of tests. All the tests consist of *I* items, but for each person two independent tests of *I* items are drawn randomly from the universe of items.

The right hand side of equation (E.3) contains five terms whose sum is the mean score. Now we can ask for each term if it contributes to the variance of the mean scores and if it contributes to the covariance between the two mean scores. In the first case (same items for everybody) the general effect μ and the average item effect are the same for all persons and do not contribute to differences in mean scores. The person effect, the average interaction effect and the average measurement error may differ from person to person and will thus contribute to the variance. Terms which contribute to the covariance are those terms which are identical in both test administrations: this holds for the person effects and for the average interaction effect, but not for the measurement error which is assumed to be independent in each test administration. In general, terms which contribute to the covariance also contribute to the variance. So we can summarize the preceding discussion in a table, like in Table E.2, in the column labelled 'one test').

Table E.2. Contribution to variance and covariance (one-facet design)

| | one test | two tests | 2n tests |
|---|---|---|---|
| Constant | $\mu, \beta$ | $\mu, \beta$ | $\mu$ |
| Variance and covariance | $\alpha, (\alpha\beta)$ | $\alpha$ | $\alpha$ |
| Variance only | $\varepsilon^*$ | $(\alpha\beta), \varepsilon^*$ | $\beta, (\alpha\beta), \varepsilon^*$ |

In the second case of two different tests, the only change is that the interaction effects will not contribute to the covariance, because the two tests are independently drawn from the universe. The two tests may be of unequal difficulty, but since the same test is used for all persons on each occasion, this difference in difficulty will not contribute to the variance within each test separately. In the third case, where everybody gets two independent tests, the item effects will contribute to the variance, because some persons will happen to get an easy test and some others will have a rather difficult test. The item effects and the interaction effects, however, will not contribute to the covariance, because they refer to two tests independently drawn from the universe.

To compute the correlation between the scores obtained in the two test administrations, we need the variances of the terms in the right hand side of (E.3). We take one term to illustrate how this variance comes about. To understand the result, we need two easy-to-prove but fundamental results from statistics, which we give here. Let X and Y represent two random variables and let *c* be a constant. Then

$$\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$$

and

$$\text{If X and Y are independent then } \mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

If we apply these rules to the variance of the mean item effect, we find that

$$\mathrm{Var}\left[\frac{1}{I}\sum_{i=1}^{I}\beta_i\right] = \frac{1}{I^2}\mathrm{Var}\left[\sum_{i=1}^{I}\beta_i\right] = \frac{1}{I^2}\sum_{i=1}^{I}\mathrm{Var}(\beta_i) = \frac{1}{I^2}\sum_{i=1}^{I}\sigma_\beta^2 = \frac{\sigma_\beta^2}{I}$$

To find the expression for the correlation we have to take a ratio: the numerator consists of the sum of all variance terms contributing to the covariance, and the denominator is the sum of all variance terms. Referring to Table E.2, we find that the correlation in the first case (symbolized by $\rho_1$) is given by

$$\rho_1 = \frac{\sigma_\alpha^2 + \dfrac{\sigma_{\alpha\beta}^2}{I}}{\sigma_\alpha^2 + \dfrac{\sigma_{\alpha\beta}^2 + \sigma_{\varepsilon^*}^2}{I}} \qquad \text{(E.4)}$$

Similarly referring to Table E.2, we find for the second case that

$$\rho_2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_{\alpha\beta}^2 + \sigma_{\varepsilon^*}^2}{I}} \tag{E.5}$$

and for the third case:

$$\rho_3 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \frac{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_{\varepsilon^*}^2}{I}} \tag{E.6}$$

There are a number of interesting observations to make about these three correlations:

1. If we know the variance components from a G-study (or have good estimates for them, which we substitute for the true but unknown values), we can compute the correlations for any value of the number of items. In all three cases it is true that the larger the number of items, the larger the correlation will be, and if the number of items is very large, all three correlations will become very close to one.
2. For any value of I, the three correlations are always in the same order:
$$\rho_3 \leq \rho_2 \leq \rho_1$$
3. Unfortunately, in a one-facet design $\rho_1$ cannot be computed, because we do not have separate estimates of the interaction component and the measurement error variance (measurement error and interaction effects are confounded). Therefore one uses $\rho_2$ instead, but from a comparison of formulae (E.4) and (E.5) we can easily see that both coefficients are equal if and only if the interaction component is zero; otherwise $\rho_2 < \rho_1$.
4. It has been shown mathematically that coefficient $\rho_2$ is equal to Cronbach's alpha, while we derived $\rho_1$ as the test-retest correlation (under the assumption of no memory effects). So $\rho_1$ is the reliability of the test in the sense of classical test theory. Cronbach's alpha will be smaller than the reliability unless the interaction term is zero.
5. Although one may regret that the coefficient $\rho_1$ is not available in one-facet designs, one should also be aware of the limitations of this coefficient, because it expresses the correlation between two test series based on exactly the same test. If interactions between students and items are really effective, the correlation $\rho_1$ will depend in a substantial way on the **specific** interaction effects in the test. If at the second administration the test is replaced by a parallel form, a quite different pattern of interaction effects may come about. One could think about this in very concrete terms: It is possible that John practiced hard last week, and he is lucky that some items in the test are very similar to the questions of his last-week exercises. So he profits from some coincidence. If, upon a second administration the very same items are used again, he will profit a second time, but in such a case the possibilities of generalization are quite narrow: we are in some sense only entitled to say that John is good at what the test measures if we stick to the very same set of items of which the test is composed. By dropping the item by person interaction term from the correlation formula (in the numerator), we just get rid of these coincidences, but that is precisely what is expressed by coefficient $\rho_2$. In Generalizability Theory $\rho_2$ is called the **generalizability coefficient for relative decisions**, because in principle it does not matter which items from the universe are chosen to compare (rank) different persons.
6. If one wants to know the level of proficiency in a more absolute way, of course it does matter which items are included in the test. A good example is a test of vocabulary. Suppose the test items ask for giving the meaning (e.g., by a translation) of 50 words. One might conceive the 50 items in the test as being randomly chosen from some lexicon or some corpus, the universe. The proportion of correctly answered items in the test is then to be seen as an estimate of the proportion of words mastered in the whole universe. This measure will not only show variation because of measurement error, but also because of sampling error in composing the test: scores will vary from test to test because of the varying difficulty of the included items and because of

interaction effects with the persons. The coefficient $\rho_3$ expresses the correlation between two series of test scores, based on randomly composed tests. In generalizability theory it is known as the **generalizability coefficient for absolute decisions**.

### E.4 Generalizability Theory for the two-facet crossed design

As was noticed in Section E.2, data which are collected in a complete three-way table (students by items by raters) are usually treated as data in a two-facet crossed design, but we have distinguished between a genuine crossed design (unrealistic but conceivable), and a special case of a nested design where the student answers each item only once and each such response is rated by the same set of raters. This latter case is ubiquitous in educational measurement, and will be denoted here as the two-facet nested design.
In section E.4.1 the genuine crossed design will be treated; in Section E.4.2 the nested design will be discussed.

#### E.4.1 The genuine two-facet crossed design

For the two-facet (items and raters, say) crossed design, the model is a straightforward generalization of model (E.1). But now we have to use three subscripts, *p* for the person, *i* for the item and *r* for the rater. The model is given by

$$Y_{pir} = \mu + \alpha_p + \beta_i + \gamma_r + (\alpha\beta)_{pi} + (\alpha\gamma)_{pr} + (\beta\gamma)_{ir} + (\alpha\beta\gamma)_{pir} + \varepsilon^*_{pir} \qquad \text{(E.7)}$$

The three double symbols between parentheses indicate **first order** interactions. There are three of them: a person-item interaction, a person-rater interaction and an item-rater interaction. The triple symbol indicates the **second order** interaction. Examples of the meaning of such interaction terms are given in Section E.2. The typical data needed to estimate the variance components are now the answers of a sample of persons to a sample of items (from the universe of items) as rated (independently) by a random sample of raters (from the universe of raters). All these ratings can be arranged in a three-dimensional array, with as many layers as there are raters. Each layer is a rectangular table just as in the one-facet crossed design. Since each cell of this table contains just one observation (the rating by rater *r* of the answer of person *p* to item *i*), the second order interaction effect and the measurement error are confounded, and we need to take them together as a residual which is now defined as

$$\varepsilon_{pir} = (\alpha\beta\gamma)_{pir} + \varepsilon^*_{pir}$$

Notice that in this case it is perfectly possible to estimate variance components of the three first order interactions. But this is only possible in the genuine crossed design where the student answers as many times to each item as there are raters.

With techniques of the Analysis of Variance one can estimate seven variance components: three for the main effects ($\sigma^2_\alpha, \sigma^2_\beta$ and $\sigma^2_\gamma$), three for the first order interactions ($\sigma^2_{\alpha\beta}, \sigma^2_{\alpha\gamma}$ and $\sigma^2_{\beta\gamma}$) and one for the residual ($\sigma^2_\varepsilon$). For tabulation purposes it is suitable to convert all components to percentages, by dividing each component by the sum of all seven components (and multiplying by 100). If some components are in reality very close to zero, it may happen that their estimates are negative. Usually one sets such estimates equal to zero.

As to the generalizability coefficients, a large number of different correlations may be predicted, and one should be very careful in defining precisely the conditions of the two test administrations and/or ratings. We will consider four different cases, which are described hereafter. In all cases mean test scores are used, which are defined as

$$Y_p = \frac{1}{I \times R} \sum_{i}^{I} \sum_{r}^{R} Y_{pir}$$

i.e., the average score across items and raters. Notice that in the description of the four cases a test arrangement is described which would deliver the correlation wanted, but such an arrangement does not have to be carried out: the correlations can be predicted from the results of a G-study.

1. One test administration with the same set of *R* raters. This case is easy to implement: after a second rating the item answers are given a second time to the same set of raters, who are requested to give their ratings again. To warrant independent ratings, one usually will not tell the raters that they have rated the performances already. The correlation to be predicted is the correlation between the mean test scores for the two ratings.
2. One test administration where the performances are rated twice, each time by an independent sample of *R* raters. The data collection design consists in administering the test once to the student and to let these performances rated by two sets of *R* raters.
3. Two independent test administrations (to the same students with the same items) and each series of performances is rated by the same set of *R* raters.
4. Two independent test administrations (as in case 3) and each series is rated by a different set of *R* raters.

In all cases the needed set(s) of *R* raters are to be considered as a random sample from the universe of raters. In Table E.4 the nine effects (the nine terms in the right-hand side of (E.7)) are assigned to a constant term, the covariance between the two series or only the variance within each series. An extra row is added to indicate the confounded terms.

Table E.4. Contribution to variance and covariance (truly crossed two facet design)

| Case | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| performances | Same | | different | |
| sets of raters | 1 set | 2 sets | 1 set | 2 sets |
| Constant | $\mu, \beta, \gamma, (\beta\gamma)$ | $\mu, \beta, \gamma, (\beta\gamma)$ | $\mu, \beta, \gamma, (\beta\gamma)$ | $\mu, \beta, \gamma, (\beta\gamma)$ |
| Var. and cov. | $\alpha, (\alpha\beta), (\alpha\gamma), (\alpha\beta\gamma)$ | $\alpha, (\alpha\beta)$ | $\alpha, (\alpha\beta), (\alpha\gamma), (\alpha\beta\gamma)$ | $\alpha, (\alpha\beta)$ |
| Variance only | | $(\alpha\gamma), (\alpha\beta\gamma)$ | $\varepsilon^*$ | $\varepsilon^*, (\alpha\gamma), (\alpha\beta\gamma)$ |
| Confounded | $\varepsilon^*$ | $\varepsilon^*$ | | |

We comment on this table:

1. The constant terms are the same in all four cases. Notice that the rater effects and the rater-item interaction effect are constant also in the case of two different sets of raters, because these effects are the same within each series of ratings.
2. The interactions containing persons and raters contribute to the covariance in the case of a single set of raters because these effects are systematic. So when there is a positive effect between student John and rater one in the first series, this effect will also be present in the second series, because the combination John and rater 1 appear in both series. In the case of two different sets of raters these effects contribute only to the variance of the test scores.
3. The interaction between persons and items is always common in the two series, and therefore contribute to the covariance.
4. The most intriguing effect is the measurement error, which represents unsystematic effects which are associated with the triple combination student-item-rater. But such a combination comes about in two steps: the performance of the student on a particular item may be incidentally (in an unsystematic way) poor, for example, and this poor performance may then be incidentally rated as reasonably good by some particular rater. The total measurement error should be conceived as the sum of these two step effects, or to say it more correctly, the measurement error has two sources of variation: the student-item combination and an effect attributable to the rater. In the truly crossed design each cell represents an independent replication of a student-item-rater combination, but in the prediction of the correlations in the cases 1 and 2, the student-item combination is held constant, while only the part of the measurement error that is due to the raters is really needed. So

to be used, the variance of the measurement error should be split into two parts: one part going to the covariance row, and one part being measurement error due to the raters. But in a truly crossed G-study with only one observation in each cell of the data table, this splitting is impossible. So from such a design, the correlations in the cases 1 and 2 cannot be predicted.

5. In cases 3 and 4, where two independent test administrations are used, the two sources that influence the measurement error are active. Nevertheless, the correlation in case 3 cannot be computed, because the second-order interaction ($\alpha\beta\gamma$) is needed separately for the covariances and the measurement error for the variance term. So, only case 4 is applicable.

This correlation, symbolized here as $\rho_4$, is given by

$$\rho_4 = \frac{\sigma_\alpha^2 + \frac{\sigma_{\alpha\beta}^2}{I}}{\sigma_\alpha^2 + \frac{\sigma_{\alpha\beta}^2}{I} + \frac{\sigma_{\alpha\gamma}^2}{R} + \frac{\sigma_\varepsilon^2}{I \times R}} \tag{E.8}$$

where the last term in the denominator refers to the residual component, the sum of the measurement error and the second-order interaction.

It should be emphasized that the preceding formula is of little practical use because the genuine crossed design is almost never applied in educational settings with raters as the second facet. Applying formula (E.8) to the estimates given in Table E.1 (for the PISA study) does not make sense, since the G-studies to estimate the variance components were based on a special case of a nested design, where the students responded only once to each item. This case is discussed in the next section.

### E.4.2 The special nested two-facets design

To model data from this design care must be taken to separate the two sources of variability in the measurement error. Therefore we will split the model in a two-step model: the first step models what happens when the student answers an item (with a given performance as the output), and the second step will model what happens when a rater rates such a performance. So the output of the first step will be the input of the second step, and the output of the second step is the observed item score given by rater $r$: $Y_{pir}$. The output of the first step will be conceived as a quantitative variable $K_{pi}$ which is unobserved, but which will be treated as a kind of auxiliary variable.

To distinguish the present model from the model used in the crossed design, the symbols for the effects will be Roman letters instead of Greek letters For the first step (at the student level) upper case letters will be used, and for the second step, random variables will be denoted by lower case letters.

The first step of the model is identical to the one facet crossed design model:

$$K_{pi} = M + A_p + B_i + (AB)_{pi} + E^*_{pi} \tag{E.9}$$

i.e., the unobserved output variable is the sum of a constant $M$, a main effect due to the person ($A_p$), a main effect due to the item ($B_i$), an interaction effect of person and item $(AB)_{pi}$ and a measurement error $E^*_{pi}$. The main effects, the interaction and the measurement error are conceived as independent random variables with a mean of zero and with variances $\sigma_A^2$, $\sigma_B^2$, $\sigma_{AB}^2$, and $\sigma_{E^*}^2$ respectively.

In the second step, one might conceive as if the output of the first step, $K_{pi}$, is amended by the rater to produce the observable rating $Y_{pir}$. Such amending may be influenced by a main effect of the raters, or an interaction effect between rater and person or between rater and item, or a second order effect (rater by item by person) and an unsystematic effect, a measurement error (at the rater level). Of course one can split all these effects into a mean effect (across raters, persons and items), and a deviation from the mean, and all the mean effects can be collected into a grand mean $m$. So we get as the second step

$$Y_{pir} = K_{pi} + m + b_i + c_r + (ac)_{pr} + (bc)_{ir} + (abc)_{pir} + e^*_{pir} \tag{E.10}$$

The models (E.9) and (E.10) cannot be used separately, because the variable $K_{pi}$ is not observed. So, both models have to be merged in some way. We do this by replacing $K_{pi}$ in the right-hand side of

(E.10) by the right-hand side of equation (E.9), and be grouping all the terms with the same set of subscripts. The result is this (with brackets placed around sums with the same subscripts):

$$\begin{aligned} Y_{pir} &= [M + m] \\ &+ A_p + [B_i + b_i] + c_r \\ &+ [(AB)_{pi} + E^*_{pi}] + (ac)_{pr} + (bc)_{ir} \\ &+ [(abc)_{pir} + e^*_{pir}] \end{aligned} \qquad \text{(E.11)}$$

where $M$ and $m$ are constants, and all ten subscripted variables are random variables whose variances one might wish to estimate. But this is impossible: random variables with the same set of subscripts are confounded, and all one can achieve is to estimate the sum of their variances. We take $[B_i + b_i]$ as an example. $B_i$ is a systematic item effect which influences the unobservable variable $K_{pi}$ and which one might call the inherent difficulty of the item, while $b_i$ is a systematic item effect which comes about during the rating of the performances, and which one might call the perceived item difficulty (by the raters). Confounding means that there is no way (in the nested design) to disentangle both effects, and that the only thing one can do is to estimate the variance of their sum. There are two other pairs of confounded variables. One is the second-order interaction effect and the measurement error at the rater level and the other is the confounding of the person-item interaction and the measurement error at the student level. Now, if we count the terms in the right-hand side of (E.11), counting bracketed terms as one single term, we see that we have one constant (first line), three main effects (second line), three first order interactions (third line) and a residual in the last line, which is just the same decomposition as in the genuine crossed design. This means that we can arrange the observed data in the nested design in a three way table which takes the same form as in the crossed design, and we can analyse this table in just the same way. The interpretation of the variance components, however, is different, as can be deduced from Table E.5

Table E.5 Correspondence between variance components in crossed and nested designs

| Crossed design | | Nested design | |
|---|---|---|---|
| Constant | $\mu$ | $[M+m]$ | Constant |
| Persons | $\alpha_p$ | $A_p$ | Persons |
| Items | $\beta_i$ | $[B_i+b_i]$ | Items |
| Raters | $\gamma_r$ | $c_r$ | Raters |
| Persons x items | $(\alpha\beta)_{pi}$ | $[(AB)_{pi}+E^*_{pi}]$ | Persons x items + error at person level |
| Persons x raters | $(\alpha\gamma)_{pr}$ | $(ac)_{pr}$ | Persons x raters |
| Items x raters | $(\beta\gamma)_{ir}$ | $(bc)_{ir}$ | Items x raters |
| Sec. order int. + error | $\varepsilon_{pir}=[(\alpha\beta\gamma)_{pir}+\varepsilon^*_{pir}]$ | $e_{pir}=[(abc)_{pir}+e^*_{pir}]$ | Sec. order int. + error at rater level |

Now, we are ready to reconsider the four cases of generalizability coefficients that were discussed in the previous section. We reproduce Table E.4 here as Table E.6 but with the symbols used in the present section.

Table E.6. Contribution to variance and covariance (nested two facet design)

| case | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| performances | same | | different | |
| set of raters | 1 set | 2 sets | 1 set | 2 sets |
| Constant | $M, m, B, b, c, (bc)$ | $M, m, B, b, c, (bc)$ | $M, m, B, b, c, (bc)$ | $M, m, B, b, c, (bc)$ |
| Var. and cov. | $A, (AB), E^*, (ac), (abc)$ | $A, (AB), E^*$ | $A, (AB), (ac), (abc)$ | $A, (AB)$ |
| Variance only | $e^*$ | $e^*, (ac), (abc)$ | $E^*, e^*$ | $E^*, e^*, (ac), (abc)$ |

Comparing Tables E.4 and E.6, we see that the row with confounded terms has disappeared in the nested design, but at the same time we see that not all four coefficients can be computed: case 1 is

excluded because the components *(abc)* and $e^*$ are needed separately, case 4 is excluded because the components *(AB)* and $E^*$ are needed separately, and case 3 is excluded for both these reasons jointly. Therefore only the correlation for case 2 (same student performance rated by two sets of R raters) can be predicted from a G-study using a nested design.

This correlation, denoted here as $\rho_5$ , is given by

$$\rho_5 = \frac{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad \text{(E.12)}$$

As an example, we apply this formula to the case of Australia in the PISA study (see Table E.1), for I = 10 items and $R$ = 1 rater (and replacing the negative variance component by zero), giving

$$\rho_5 = \frac{22.4 + \frac{50.36}{10}}{22.4 + \frac{50.36}{10} + \frac{8.01}{10 \times 1}} = 0.972$$

This is the prediction of the correlation one would find between two ratings (each by one rater) of the performances of a (random) sample of students on 10 items. However, one should be careful here, and not confuse this case with case 4, where the same sample of students takes the test twice, and each test performance is rated by an independent rater, which is case 4 with $R$ = 1. In this case the correlation is given by

$$\rho_6 = \frac{\sigma_A^2 + \frac{\sigma^2_{AB}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB} + \sigma^2_{E^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad \text{(E.13)}$$

and it is immediately seen that the interaction component needed in the numerator is not available from the G-study. Nevertheless, we can make good use of (E.13) if we have a reasonable estimate of the person-by-item-interaction component. In the PISA study the Rasch model (see Section G) has been used as IRT model, and this model presupposes absence of interaction between persons and items[5]. So we might assume quite reasonably that the component 'person by item interaction plus error at the person level' is to be attributed (almost completely) to measurement error at the person level. Or, in other words, that the person by item component is zero. If we apply formula (E.13) with this assumption to the case of Australia with $I$ = 10 item and $R$ = 1 rater, we obtain

$$\rho_6 = \frac{22.4}{22.4 + \frac{50.36}{10} + \frac{8.01}{10 \times 1}} = 0.793$$

which is a marked difference with the previous result of 0.972[6,7].

[5] This absence of interaction is at the level of the latent variable, and does not preclude interaction at the level of the observed scores. Extensive simulation studies (with a crossed design) have shown, however, that the person-by-item-interaction component at the observed score level usually is below 5% of the total variance.

[6] In the Technical Report of Pisa 2000, a formula similar to (E.12) was used, but the result was erroneously interpreted as a correlation with two independent administrations, like formula (E.13). Moreover, the formula used in the Pisa report also contains an error, because the rater effect and the rater by item interaction were erroneously considered as contributing to the variance. But since the estimates of these effects were negligible, this latter error had no noticeable effect on the results.

[7] If the interaction component is set to 5% of the total variance (and consequently the error at the person level at 50.36% - 5% = 45.36%), the result for $\rho_6$ is 0.811

The use of the results of a G-study, however, is much more versatile than the preceding example suggests. On can use the formulae (E.12) and (E.13) (and many others) to predict the correlations for different values if *I* and *R*. One might, for example, investigate whether the correlation $\rho_6$ would increase more by doubling the number of items or by doubling the number of raters in a future application. Applying any of these strategies will lead to doubling the total amount of rating time and costs while the first strategy will lead to doubling of the test taking time. In Table E.7, formula (E.13) has been computed with the results of the G-studies displayed in Table E.1 for 10 and 20 items and for 1 and 2 raters, and using the assumption that the true person by item interaction component is zero throughout.

The results are very easy to interpret in this case: doubling the number of raters do increase the correlations marginally, while doubling the number of items leads to a much more impressive increase of the correlation. This is consistent with the order of magnitude of the residual components in Table E.1: the measurement error attributable to the students (given in the column 'student by item interaction') is much larger than the error attributable to the raters (the column 'residual' in Table E.1). To reduce the impact of the former, the number of items has to be increased (see the denominator in formula (E.13): the confounded student-level error and first order interaction component is divided by the number of items, and since this is the largest component, the impact of changing the number of items will be the most drastic. Changing the number of raters diminishes the impact of the student by rater interaction component, but since this component is negligibly small in all countries, the impact on the change of the correlation will be negligible as well. The residual term is influenced in an equal way by doubling either the number of items and the number of raters.

Table E.7 The coefficient $\rho_6$ for the results in Table E.1 (student by item interaction set to zero)

| | $I = 10$ | | $I = 20$ | |
|---|---|---|---|---|
| | $R = 1$ | $R = 2$ | $R = 1$ | $R = 2$ |
| Australia | 0.793 | 0.805 | 0.884 | 0.892 |
| Denmark | 0.676 | 0.692 | 0.803 | 0.816 |
| England | 0.701 | 0.707 | 0.824 | 0.828 |
| Finland | 0.751 | 0.762 | 0.858 | 0.865 |
| Norway | 0.696 | 0.707 | 0.817 | 0.826 |

In conclusion we can summarize the results of the G-studies carried out in the PISA project as follows:

1. From Table E.1 we see that there are almost no systematic effects in the data due to the raters: rater main effect and first order interactions where raters are involved (the shaded columns) are negligible.
2. If the genuine student by item interaction component is assumed to be negligible, the big component in the column (S x I) has to be interpreted as measurement error at the student level, while the residual term is to be interpreted as a residual at the rater level (measurement error confounded with second order interaction). Although there is some confounding, it is reasonable to assume that the genuine interactions are much smaller than the measurement error.
3. This separation of two kinds of measurement errors (in the analysis of G-study data) is only possible in the special nested design (all raters judge on the same performances of the students), and not in the truly crossed design, where the two kinds of measurement errors are confounded.
4. Two different correlations, issuing from the nested design were studied. One ($\rho_5$, formula (E.12)) predicts the correlation between two series of independent ratings based on the very same student performances; the other ($\rho_6$, formula (E.13)) predicts the correlation between two series of independent ratings based on two independent test administrations. The former is an exact formula, the latter can only be used as an approximation, because one has to add an assumption about the student-item interaction component.
5. In the PISA study all $\rho_5$ correlations were very high (the present text gives only one example), while the $\rho_6$ correlations are substantially lower and also show substantial variation across countries. The reason why they are lower is due mainly to measurement error at the student level, which is much more important than the error at the rater level. In the light of this finding it would

have been of little use to let the performances of all students in the study to be rated by two (or more) raters. This can be clearly seen from Table E.7.

The example used in this Section may be atypical for many educational settings. In general one has to pay attention to a number of aspects when one carries out a G-study, using raters as one of the facets. We discuss these in turn.

1. The notion of random sampling in such studies is quite important. Especially the raters should be drawn randomly from the universe of raters which are possible candidates to do the rating work in large scale applications. Using only the best or most motivated raters for the G-study may invalidate the generalizability of the conclusions from such a study. Particularly, the use of only volunteers in the G-study may result in a non-representative sample. Moreover, the conditions for the rating work (allowed time, instructions, amount of training, etc.) should be the same in the G-study as in real applications.
2. In the PISA study the systematic effects associated with the raters were negligible, but this is not necessarily the case in G-studies.
    a. A substantial main effect component for the raters indicates differences in leniency. If in real applications of the test, the test score is to be compared with a pre-established standard (to succeed or to fail, for example), such differences may lead to incorrect decisions about the candidates.
    b. A substantial item-rater interaction component may be caused by different interpretation of the scoring rules by different raters. A more detailed search into the data (or an interview with the raters) may reveal that some rules are unclear or ambiguous. Although this interaction and the main effect do not appear in the formulae for $\rho_5$ and $\rho_6$, they may lower the reliability in other cases which are not discussed in detail in the present report. Here is an example. Suppose the work of 1000 students has to be rated (in an application), and one uses 10 raters to do the rating work, each rater rating 100 performances. If there are systematic differences between the raters, these will cause irrelevant (and therefore unreliable) variation in the test scores.
    c. A substantial student-rater interaction component is a serious problem. It may show up if some of the raters happen to know (and can identify) some of the students. This is important to remember when one tries to generalize the results of the G-study to future applications. It may be that in the G-study the students are anonymous to the raters and that no such interaction appears, but in future applications most of the rating may be done by the students´ own teacher. In such a case one cannot be sure that in the application this interaction will be absent.
3. The coefficient $\rho_5$ is the correlation between two independent ratings (each by $R$ raters) of the same student performances. One can compute it for different values of $R$ (usually the values 1, 2 and 3 will suffice). If this correlation is deemed too low if $R = 1$, but acceptable for $R = 2$, this means that in future applications one has to use two independent raters for each student, which can be very costly. Of course, one could also revise the scoring rules or provide better training or supervision of the raters, but one should realize that taking such measures does not automatically remove the problem. One can only be sure about this by doing a new G-study after these measures have been implemented.
4. It may be useful to compare $\rho_5$ to $\rho_6$ for different values of $R$ and $I$. The coefficient $\rho_6$ can be interpreted as a test-retest correlation. We have seen that its departure from the ideal value of one is due partly to the students and partly to the raters. By comparing it to $\rho_5$, one gets an impression whose contribution is the most important, and one can take measures to improve the reliability either by increasing the number of raters or the number of items administered to the students. The construction of a table like Table E.7 may be helpful in such a case.

# Section F

# Factor Analysis

**N.D. Verhelst**
**National Institute for Educational Measurement (Cito)**
**Arnhem, The Netherlands**

The performance on a test is usually summarized by a single number, the test score. This test score is a composite score, because it is built (by taking a sum) from item scores. In general one might ask whether it is meaningful to put a number of items together in a test and to let the performance be represented by a single number. What if the test consists of a mixture of two kinds of items, each kind measuring a different concept? Is reporting a single score meaningful or should one treat this composite test as two tests and report two test scores?

A model suitable to detecting if there are more dimensions responsible for the performance on the test is Factor Analysis (F.A.). The model originated in psychology, more than a hundred years ago and is still one of the most applied models in the social sciences. Although not defined originally as such, the model fits very well in the family of IRT-models to be discussed in Appendix G. But since the model and the techniques to carry out the analyses are so wide-spread (as well as a lot of misunderstandings about them), a separate, though short appendix is devoted to F.A.

The basic observation from which F.A. originated is the non-zero (but also not perfect) correlation between several measures that belong to some broad domain, like cognitive tests. F.A. is a model which explains the pattern of correlations that issues from observations in testing (or other measurements). Basically it says that since the correlations are not zero, the measurements must have something in common, and, since the correlations are not perfect either, the measurements must have also something unique. This is the general idea, which will be made more concrete next.

The common thing that tests share is called a factor (or, as the case may be, several factors). A factor is conceived as a non-observable (or latent) continuous variable, and every person taking the test can be represented by a value on this variable, called a **factor score**. If there are more factors, every person has a factor score on each factor. The 'unique thing' can also be conceived of as a factor, where the person also has a score. The **observed score** on a test is conceived of as a weighted sum of the factor scores, including the unique factor. In Table F.1 an example is provided with three tests and two common factors (The notion of 'common' factors is explained by means of the table)

Table F.1. The basic model of Factor Analysis

| | weights for | |
|---|---|---|
| | factor 1 | factor 2 |
| test 1 | 0.4 | 0.2 |
| test 2 | 0 | 0.7 |
| test 3 | 0.7 | -0.3 |

Suppose John's factor scores on the two factors are +1.2 and 0.8, respectively. Then the model says that John's observed score on test 1 is $0.4\times1.2+0.2\times0.8+$ his score on the unique factor for test 1. But we know from Classical Test Theory that the observed score also contains a measurement error. Therefore we have to conceive the score on the unique factor as a mixture of something systematic (but unique to the test) and the measurement error. But these two are confounded and cannot (with the three tests at hand) be disentangled. The other two factors are called common factors, because for each factor there exist at least two different tests with a non-zero weight for that factor. These weights are called **factor loadings**, and the main purpose of Factor Analysis (as a technique) is to determine these

weights. All one needs to carrying out such an analysis is the table of correlations (or covariances[1]) between the tests.

The discussion in the present section will be restricted to points which are essential in the interpretation of factor analytical results.

1. **Unique factors**. Suppose that in the preceding example, test 1 is a reading test, test 2 is a writing test and test 3 is a listening test. Suppose further that the reading test contains a lot of items (or text passages) on history, while the other two tests have nothing to do with history. Suppose, finally, that John is particularly good at history, such that his score on test 1 is determined to a considerable extent by his knowledge of history, while Mary is not very good at history, such that her knowledge in that domain will not be of much help in answering the questions of the reading test. This makes clear that knowledge of history will account for some variability in the test scores of test 1. But since the other two tests have nothing to do with history, 'knowledge of history' is unique for test 1, and cannot appear as a common factor. If we add a fourth test to the collection (a history test, for example), then there will be two tests which have history as a common factor, and this will show up in the analysis, and we might end up with three common factors, where the third factor has loadings of zero for tests 2 and 3, but non-zero loadings for test 1 and the added history test. More generally this means that unique factors are to be considered relative to the collection of tests included in the analysis.
2. **Origin and unit**. Suppose the factor scores on factor 1 for all people are multiplied by 2, and at the same time the factor loadings in column 1 are divided by 2, then the product of the transformed factor scores and the transformed weights would not change. Multiplying the scores by 2 is choosing another unit of measurement (if one owns 1000 euros, one also owns 2000 'half-euros'). The unit of measurement is in principle free (arbitrary), but to make communication possible, the unit used must be specified. It is common practice to choose the standard deviation of the factor scores as unit, or in other words, the standard deviation (in the population) of the factor scores is one. With a similar reasoning, one can choose the origin of the scale in an arbitrary way. It is common practice to choose the average factor score (in the population) as origin. Therefore, it is a common convention (and not a metaphysical truth) to say that factors have a mean of zero and a standard deviation of one. (Notice that this is **not the same** as saying that the factor scores are normally distributed.)
3. **Correlations and covariances**. Factor analysis can be carried out on tables (matrices) of correlations and on tables of covariances. A covariance (between two variables) is a measure of covariation (meaning literally: varying together). Its value depends on the unit of measurement used for the two variables. A correlation is a kind of standardized measure of covariation and varies between –1 and +1. If the correlation matrix is used for a factor analysis (as we will assume in the sequel), then the factor loadings cannot be larger than one in absolute value.
4. **Orthogonal factors**. The indeterminacy of what are called factors is more complicated than only the freedom in the choice of the unit and the origin. Also the correlational structure of the factors in the population is arbitrary (not completely, but to a large degree). For example, if there are two common factors, they can always be defined in such a way that the correlation between the factor scores (in the population) has an arbitrary value (different from –1 and +1). But changing the correlation will also lead to a change in the factor loadings. In many applications, the factors are chosen in such a way that their correlation is zero. Any pair of factors with zero correlation is called orthogonal. Most software give factor loadings for orthogonal factors as their primary output.
5. **Communality**. The sum of squares of the factor loadings (on the common factors, and with orthogonal factors) of a particular test is called the **communality** of that test. From Table F.1, we see that the communality of test 3 equals $0.7^2 + (-0.3)^2 = 0.58$. The communality is the proportion of the test variance that is explained by the two factors. In this case 58% of the variance

[1] The **covariance** between two variables is the correlation **multiplied** by the product of the two standard deviations. Or, conversely, the correlation is the covariance **divided** by the product of the two standard deviations. If one of the standard deviations equals zero, then the covariance is also zero, but the correlation is not defined, because the division of zero by zero is not defined.

is due to the two factors, and the complement (42%) is explained by the unique factor, part of which (but unknown) is due to measurement error. Thus we see that from F.A. we get another lower bound for the reliability of the test: the reliability is at least as large as (but may be larger than) the communality. As can be deduced from the discussion on unique factors, this lower bound may change as more or other tests are analyzed jointly in a F.A.

6. **Contribution of factors**. One may also take the sum of the squared loadings for a particular factor across the tests. This sum is called the contribution of that factor (to the total variance). In Table F.1 the contribution of the first factor is $0.4^2 + 0^2 + 0.7^2 = 0.65$. The contribution of the second factor is 0.62. Their sum (0.65+0.62=1.27) can be compared to the total variance which is the number of tests, in the present case 3 (Since we use correlations, each variable has been standardized, and thus has a variance equal to one). So, in the example we see that about 42% $(100 \times 1.27/3)$ of the total variance is explained by the two common factors. The remaining part is due to the unique factors. Most techniques of factor analysis determine the factors in such a way that the first factor explains as much variance as possible, the second factor then explains as much variance of the variance not explained by the first factor, etc. The technical term used for the determination of factors is **extraction of factors**. Notice that this way of extracting factors is just a mathematical procedure; it does in no way justify any substantive meaning or interpretation whatsoever to be attached to these factors. We will come back to this point.
7. **Reproduced correlations**. If we have the factor loadings, we can reproduce the correlation matrix from them. The reproduced correlation between two tests is the sum (over factors) of the products of the factor loadings of the two tests. From Table F.1 we can compute that the correlation between test 1 and test 3 is $0.4 \times 0.7 + 0.2 \times (-0.3) = 0.22$. Factor analysis as a technique does the reverse in some sense: from the correlations it has to compute the factor loadings. This reverse operation (which is mathematically not simple), however, is not well defined, because there does not exist a unique solution but infinitely many of them, even if we require that the factors are standardized and mutually orthogonal. This is explained next.
8. **Orthogonal rotation**. The factor loadings of Table F.1 are displayed graphically (as points in a plane) in the left hand panel of Figure F.1: the loading on the first factor corresponds to the x-value of the point, the loading on the second factor to the y-value. The points representing tests 1 and 3 are connected to the origin by a dashed line. Although the reproduced correlation was computed as a formula involving the loadings, it can also be computed from the distances of the points to the origin (the length of the dashed lines) and the angle between the dashed lines. Now imagine that the points representing the tests are fixed on the paper surface, but that the axes of the system lie loosely on the paper surface, fixed at the origin, such that they can rotate. In the middle panel of the figure this is shown by the dashed lines: both axes are rotated 45 degrees clockwise. In the right hand panel then, the old axes are removed, the new (rotated) ones are displayed as solid lines now, and the whole picture is turned such that one axis is horizontal and the other vertical. Notice that the pattern of dashed lines connecting the two test points to the origin has not changed: the dashed lines have the same length as in the first case, and they form the same angle. But the values of the x- and y-coordinates have changed. Their values are given in Table F.2, together with the old ones. It can be checked easily that the reproduced correlation from either solution are identical. Of course, we could have rotated the original axes an arbitrary number of degrees, each rotation giving a different solution, and there is no best solution, because they are all equivalent.

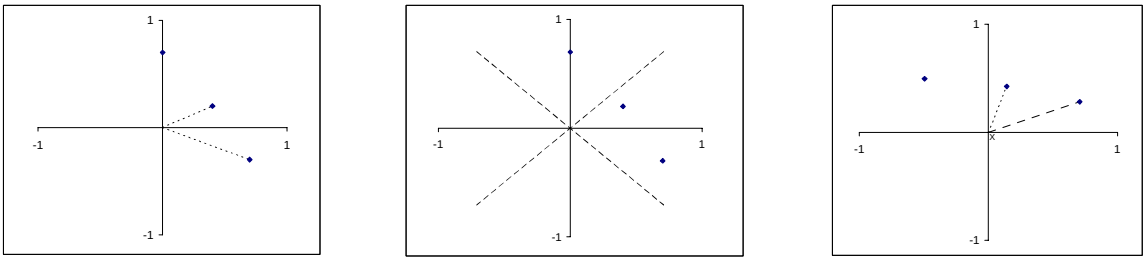



Figure F.1. Orthogonal rotation

Table F.2. Factor loadings before and after rotation

| | before rotation | | after rotation | |
|---|---|---|---|---|
| | factor 1 | factor 2 | factor 1 | factor 2 |
| test 1 | 0.4 | 0.2 | 0.141 | 0.424 |
| test 2 | 0 | 0.7 | -0.495 | 0.495 |
| test 3 | 0.7 | -0.3 | 0.707 | 0.283 |

9. **Interpretation**. Suppose the tests used in a single factor analysis consist of four reading tests and four listening tests. Suppose further that we can find a rotation such that the four reading tests have a positive loading on the first factor and a zero loading on the second factor, while the reverse holds for the listening tests. Then we could say (but this is a summary of our finding) that the first factor is a 'reading factor' and the second a 'listening' factor. This means that we can describe the covariation between eight original variables by a more parsimonious conceptualization which involves only two concepts. It does not follow, however, that there 'must' exist 'something real' (like a brain center) which is responsible for reading and another something which is responsible for listening. Conclusions like these are called reification, and they are not logically allowed: maybe there do exist such brain centers, but their existence does not follow from a factor analysis.
10. **Statistics and the number of factors**. All that has been said up to now is related to an analysis of a correlation matrix as it exists in the population. But the only thing one can analyze in practice is a correlation matrix computed on the data of a sample (usually the calibration sample). Therefore the correlations in the matrix are estimates of the population correlations, and the factor loadings will also be estimates of the population factor loadings. This all may sound quite familiar by now, but there is an extra (and quite difficult problem) associated with F.A. Suppose the population correlation matrix for 10 variables can be reproduced completely (i.e., without any error) with two factors. Then the matrix of estimated correlations will very likely not be reproduced with two factors. In general more factors will be needed, and in many cases the number of factors will be equal to the number of variables. This is caused by the estimation errors in the sample correlations. Usually one will not use as many factors as there are variables, but if we do not know the exact number of factors required for the reproduction of the population matrix (and usually we do not know), we have to guess it. There exist some mathematical criteria to help in this guessing but none is foolproof.
11. **Exploratory and confirmatory F.A.** Originally, F.A. was developed as an exploratory technique. A collection of tests is factor analyzed 'to see' the factorial structure. Much effort has been devoted to develop special rotation techniques which might be helpful in the interpretation of the factors. The best known, and still frequently used method of rotation is the varimax rotation. It is available in most statistical packages. The big problem with exploratory factor analysis is that it is quite difficult to determine the 'real' number of factors. (This number must be specified by the user in carrying out the analysis.) In the 1970's statistical theories were developed where one can impose a prespecified structure on the factor loadings as a hypothesis. Here is an example: suppose the test constructor wants to factor analyze jointly four reading tests and four listening tests, and he has the hypothesis that reading and listening should be conceived of as two distinct proficiencies. This hypothesis can be translated in a partial fixing of the table of factor loadings, by requiring that the reading tests have a loading of zero on the first factor (so this factor represents the 'listening factor'), while the listening tests have zero loadings on the second factor. So, eight of the sixteen cells of the table of factor loadings are filled already with numbers issuing from the hypothesis. With the software for confirmatory F.A. the non-specified loadings are estimated, but things are a little bit more complicated now: the researcher also has to specify if he thinks that these two factors are orthogonal (i.e., uncorrelated) or not. In the latter case, the software also estimates the correlation (in the population) between the two factors. But it does more: it performs a statistical test that can be used to decide whether the hypothesis put forward is tenable or not. In general the use of such models is not a simple matter, and special training is strongly advised.
12. **When tests are items**. There is no objection in principle to use one-item tests to carry out a F.A. So, one can use the items of a test under construction as one-item tests, compute the correlations

between the items on the calibration sample and submit it to a computer program for factor analysis. There are, however, a number of problems associated with this approach. Three of them are discussed briefly.

**a.** Since factors are conceived of as continuous variables, any weighted sum of factor scores (and the observed score is such a weighted sum) is also continuous. If the tests are items, and their score can assume only the values 0 and 1, this leads to an inconsistency which usually shows up in the following way. If the correlations between the items are computed using the usual Pearson correlation coefficient (also called φ-coefficient), F.A. will usually find (too) many factors which are hard to interpret. Therefore it is strongly advised to use tetrachoric correlations, which are based on the assumption that a binary variable is the result of a dichotomisation of an underlying continuous variable. There is no simple formula to compute these correlations but they can be computed with many software packages.

**b.** Tetrachoric correlations have relatively large standard errors. If the sample size is small, this may lead to a difficult decision as how to choose the correct number of factors and to large standard errors of the factor loadings, complicating the interpretation of the extracted factors.

**c.** There exist many mathematical methods to do a F.A. Most of them require that the correlation matrix to be analyzed has a special mathematical characteristic called 'positive semi-definiteness'. A matrix of tetrachoric correlations often does not possess this characteristic, so that the operation of extracting factors will fail. There are two methods that do not require this characteristic, the so-called MINRES method and Principal Axes method. One should choose one of these in carrying out an exploratory analysis, because other methods will fail if the matrix is not positive semi-definite. Confirmatory analyses will fail in such a case.

13. **The case of a single common factor**. If there is only one common factor (in the population), one might conclude that this is a 'proof' of unidimensionality, which makes the operation of summarizing the test performance by a single number meaningful. One should be very careful with such reasoning: a one-common-factor case is better interpreted as a necessary, and not a sufficient requirement. This is illustrated with a small example. Suppose a F.A. is carried out on three reading tests, where questions are asked on text passages. In the first test, the passages are on art, in the second on technology and in the third one on sports. The loadings on the common factor are 0.72, 0.70 and 0.40 respectively. Here are some comments:

**a.** Sometimes comments are heard like this one: "The performance on (my) reading tests are governed by a single proficiency irrespective of the content of the text passages; the fact that there is only one factor 'proves' that the tests measure reading ability and nothing else." Such reasoning, however, is a fallacy: it may be the case that the scores on the three tests are (partly) determined by specific knowledge of arts, technique and sports. If the amounts of knowledge in these three domains are not correlated in the population, their effect will be absorbed into the unique factors and cannot be distinguished from measurement error. So the only way to know is to add another three tests in the same domains. In that case, the systematic effect of the specific domain knowledge will show up as three common factors. This is an example of performing a thorough validation of a test, even without the technical tool of confirmatory F.A.

**b.** The example also gives a nice opportunity to help in the interpretation of the factor loadings. In principle, factor loadings have nothing to do with the difficulty of the tests that are analyzed, but they are indices of discrimination. It can be shown mathematically that a factor loading is the correlation between the test score and the common factor. So in the example considered, the tests on arts and technology correlate substantially higher with the common factor than the test on sports. If the tests used in the F.A. are single items, the same principle applies: the factor loadings express the correlations between the items and the underlying factor, and can thus be used instead of the correlation between items and test score as a measure of discrimination.

**c.** The problems associated with factor analysis on items are hard and in the literature no completely satisfactory solution to handle them is available in the framework of factor

analysis, i.e., in the approach which takes the correlation matrix between the items as the basic data to be analyzed. In a sense, students of factor analysis tend to consider F.A. on binary variables as a kind of nuisance. There is, however, a different approach possible which puts the binary character of the variables to be analyzed at the center of the approach. This approach is known as Item Response Theory (which developed historically quite independently from factor analysis). It is discussed in Section G.

# Section G

# Item Response Theory

**N.D. Verhelst**
**National Institute for Educational Measurement (Cito)**
**Arnhem, The Netherlands**

*This section consists of four non-technical sections (containing no formulae) where basic notions of IRT are explained and discussed. Following these, a number of notions and techniques are discussed in a more formal and technical style (sections G5 through G.7). To avoid the use of formulae as much as possible, we have made extensive use of graphical displays. It is possible to learn a lot from graphical displays used as examples in a textbook, but one learns a lot more by producing the graphs oneself and using one's own material. To help the reader in constructing graphs using modern computer technology, a special section (G.8) has been added where it is explained, step by step, how most of the graphs in the preceding sections are produced.*

## G.1 General characterization

The basic notion in Classical Test Theory is the true score (on a particular test). In Item Response Theory (IRT) the concept to be measured is central in the approach. Basically, this concept is considered as an unobservable or latent variable, which can be of a qualitative or a quantitative nature. If it is qualitative, persons belong to unobserved classes or types; if it is quantitative, persons can be represented by numbers or points on the real line, much like in factor analysis.

Approaches where the latent variable is qualitative are primarily used in sociology. The technique to do analyses of this kind is called latent class analysis. It will not be discussed further in this appendix.

In psychology and educational measurement the approach with quantitative latent variables is more widespread, and it will be the focus of the present section. We will start with a quite old approach by Louis Guttman. It contains a number of very attractive features and makes it possible to understand clearly the approach and theoretical status of IRT.

The concept to be measured (an ability, a proficiency, or an attitude) is represented by the real line, and a person is represented by a point on that line, or what amounts to the same, by a real number. The line is directed: if the point (of person) B is located to the right of the point (of person) A, we agree to say that B is more able, proficient, or has a more positive attitude than A. The basic purpose of measurement is to find as precisely as possible the location of A and B (and of everyone one might wish to measure) on that real line. To do this, one must collect information on these persons, and this is done by administering items to them. In this sense, an item response is considered as an indicator of the latent underlying variable. In the theory of Guttman, an item is **also represented by a point on the latent continuum**, where it has the status of a threshold: if the person's point is located to the left of the item point, then the item is (always) answered incorrectly; if the person's point is located to the right of the item, it is (always) answered correctly. So far the theory is somewhat trivial, but it does not remain so if we consider the responses to more than one item.

Consider the case of a three item test, with items *i, j* and *k*, and suppose the location of these items on the latent continuum is in this order: item *i* takes the leftmost position and item *k* the rightmost one. We can conceive of these three items as cut points of the real line (they cut the real line into four pieces). All persons having their representations to the left of threshold *i* give three incorrect answers, between *i* and *j*, only item *i* is answered correctly; between *j* and *k*, items *i* and *j* are correct, and to the right of *k*, all three responses are correct. In Table G.1 the four response patterns are displayed. Seen as a whole, the '1' scores form a triangular pattern, indicated by the shading. If the theory is adequate, then we can find an ordering of the items (in the present case the ordering of *i, j, k*) and an ordering of

the different response patterns such that this triangular shape arises. This solution is called a scalogram.

Table G.1. A scalogram

| item i | item j | item k |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 1 |

Is this a theory? Yes, it is and it is a very strong one. A theory is a coherent narrative about reality, which imposes certain constraints on possible phenomena. Guttman's theory (in the present example) says that a response pattern like (1,0,1), although possible, will not and may not occur. In general, Guttman's theory says that with $p$ items, only $p$+1 response patterns can occur (which, moreover, have to fit in a scalogram) while the number of possible response patterns is $2^p$. (If $p$ = 10, 11 different response patterns may occur, while 1024 different patterns are possible). This is a very strong prediction, and the theory can be **falsified** by a single occurrence of a single not-allowed pattern. The theory is so strong that it has to be rejected almost always in practice. Even one simple mistake in the recording of the item answers may suffice to reject the theory, and this is the weak point of Guttman's theory: it is **deterministic**, i.e., it claims that the response is predictable without error from the relative position of person and item on the latent continuum. The left hand panel of Figure G.1 shows this in a graphical way: to the left of the item point, the probability of a correct response is zero, to the right it is one (and at the point itself, it is left unspecified: the vertical dashed line is only added as visual support).

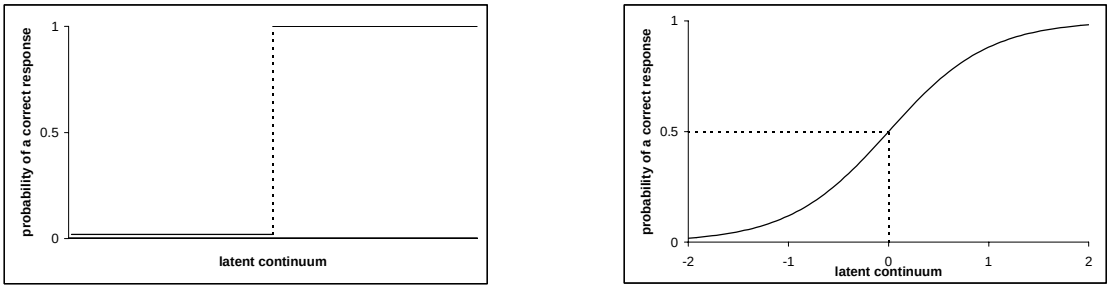



Figure G.1. A deterministic and a probabilistic model

An elegant way of getting rid of this deterministic character of the theory is to avoid this sudden jump from zero to one, and let the probability of a correct answer increase smoothly as the latent variable shifts from low to high values. This is shown in the right hand panel of Figure G.1. But eliminating the jump also makes the location of the item on the latent continuum unclear. Therefore one needs a convention, and the convention agreed upon in the literature is to define the location of the curve as that value of the latent variable that corresponds to a probability of ½ to obtain a correct answer. In the right hand panel of the figure, one can say that the curve is located at zero.

With the help of this curve, we can list a number of properties which are common to all models which are used in IRT:

1. The curve is increasing, meaning that the higher the value of the latent variable, the higher the probability of a correct response. (There are also models where this monotonicity is explicitly avoided, but these models seldom find useful application in educational testing.)
2. The probability of a correct answer is always greater than zero and always smaller than one. This means that there is always a positive probability of getting the answer right even for very low values of the latent variable, and always a positive probability of an error, even for very high values.
3. The curve describing the probability is continuous, i.e., it has no jumps like in the Guttman case.

4. The curve is ‘smooth’. For the discussion in this section, this is not important; for the mathematics to be done in IRT, it is.

In Figure G.2 two situations are displayed with two items. In the left-hand panel the two curves have exactly the same form, one is just a horizontal shifting of the other. In the right-hand panel, the rightmost curve has another location (see the dashed lines), but is also much steeper than the other.

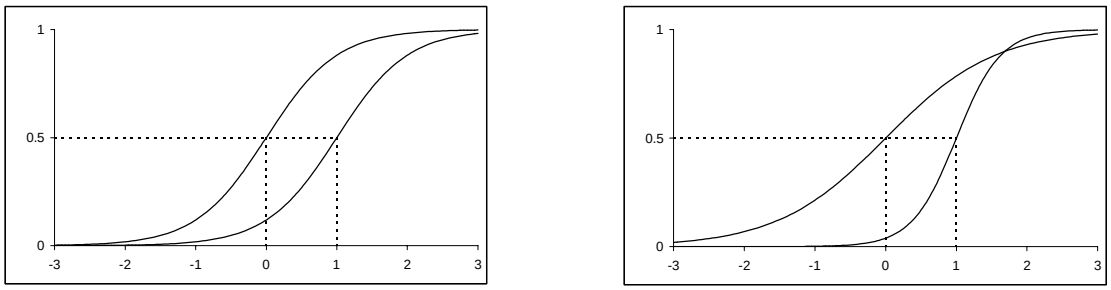



Figure G.2. Differences in difficulty and discrimination

In the left-hand panel one sees that one curve is located at zero and the other at the value of one. For the latter one, a higher value of the proficiency is needed to obtain a probability of ½ than in the former case, so one can say that the latter item is more difficult. This is what is generally done in IRT: the amount of proficiency to obtain a probability of ½ for a correct answer is defined as the index of difficulty of the item. In the right-hand panel the two items also have difficulty indices of zero and one respectively, but the more difficult item is also better discriminating than the easy one. This difference in discrimination is reflected by the differences in steepness of the two curves; the steeper the curve the better the item is discriminating. The two most important characteristics of the items are thus visually reflected in the figures: difficulty by location and discrimination by steepness. From the right-hand panel it is also clear that discrimintation is a local property of the item: the well discriminating item discriminates between people having a theta value lower than 1 (all having a low probability of getting the correct response) and higher than one (having a high probability); it does not discriminate for example between a theta value of –1 and –2, because at these two locations the probability of a correct response is very near zero (see also Section C).

Now we are ready for some terminology. In principle we can draw a curve like in Figure G.2 for each item in a test. These curves are called **item response curves**. The curves are graphs of a mathematical function which relates the value of the latent variable to the probability of a correct response. These functions are called **item response functions**. To be able to do mathematics with these functions, however, we need to know something more than only the graphs; we need a formula (a function rule) which expresses the exact relation between the latent variable and the probability. In such a formula the latent variable is usually represented by the Greek letter theta ($\theta$). There are many rules which result in a sigmoid graph like in the figure, and we could in principle choose a different rule for each item. But in the left-hand panel of Figure G.2, the two curves have the same form, only their location differs. So it is reasonable (and parsimonious) that their formulae are also very similar, but at the same time general enough for allowing differences in location. This is done by constructing a function rule where the precise value of the location is left unspecified, and is represented by a symbol. We will use the symbol $\beta$ for this. If zero is substituted for this symbol, the resulting function rule is the rule for the leftmost curve in the figure; if one is substituted, we get the rightmost curve. So $\beta$ is the symbol for a number, and since we leave it unspecified, it is called a **parameter**. So we may think of both curves as being described by the same rule, but with a different value of the $\beta$-parameter. In general we will say that the item response function of item 1, has parameter $\beta_1$, that of item 2 has parameter $\beta_2$, and in general that item *i* has parameter $\beta_i$. Since these parameters indicate the degree of difficulty of the item they are called **difficulty parameters**. One can also say that the general rule describes a family of curves, and the rule with a specific value of the difficulty parameter describes a particular member of this family.

In the right-hand panel of Figure G.2, the curves differ in two respects. To describe them as members of the same family, we will need a broader family, where members can differ not only in difficulty but also in discrimination. Therefore we will need two parameters, a difficulty parameter and a discrimination parameter. Details are discussed in Section G.5.

For the general function rule, many rules are applicable in principle, but one has become very popular, because of its mathematical elegance and because of a number of quite mathematical and philosophical reasons, which will not be discussed here. Its name is the **logistic** function. If it is used to characterize the item response functions, one says that the logistic **model** is used. The logistic model where it is assumed that all items in the test have the same discrimination (like in the left-hand panel of Figure G.2) is called the **Rasch** model (after the Danish mathematician G. Rasch who invented it). In case different discriminations are allowed as well, the model is called the two-parameter logistic model (2PLM).

One should clearly realize that all the above is a narrative (theory) about the world (admittedly a small piece of the world, but anyway), and that, although it may sound elegant and plausible, it is not necessarily true. Moreover, its basic entities – theta-values, difficulty parameters, probabilities – are not directly observable, although we need them in applications. The only observables we have are the observed answers to the items in the calibration sample, or more exactly, a summary of them: a table filled with ones and zeros. Using this table, we have three tasks that must be carried out:

1. Estimating the item parameters (difficulty parameters and possibly discrimination parameters);
2. Checking the truth (validity) of our narrative;
3. Estimating the theta-value of the persons in the calibration sample, and of future test takers.

These three steps are discussed in turn. Steps one and two are usually carried in a single run of a software program. The two steps jointly are usually designated as **calibration**.

## G.2 Estimation of parameters

The procedures by which parameters are estimated in IRT are generally quite complicated and cannot be carried out without a computer. There are, however, a number of features of this process which have direct implications for the practical use of the results. We will discuss them in a number of short paragraphs.

1. **Maximum Likelihood (ML)**. This expression refers to a general procedure to estimate parameters in probabilistic models. In general it chooses the values of the parameters in such a way that the data we have are as likely or probable as possible. How this is done, is a highly technical problem, but it is important to notice that the estimates your colleague obtains with his data will differ in general from the estimates you have with your data, even if both of you estimate the same 'true' parameters. Therefore, estimates always should be accompanied by a standard error which is a degree of accuracy of the estimate. The most important way to influence this accuracy is the sample size. In Section G.6, the principle of maximum likelihood is discussed in more detail..
2. **Joint Maximum Likelihood (JML)**. Suppose we use the Rasch model with $k$ items and $N$ persons. The unknown quantities in this problem are the $k$ difficulty parameters and the $N$ theta values of the test takers. We can treat these $N+k$ unknown quantities formally as parameters and estimate them **jointly** from the data by a maximum likelihood procedure. This is what was done in the first software that was developed for IRT in the U.S.A. This procedure, however, leads to problems: the bigger the sample size, the bigger the problem is, because each new person brings his/her own theta value. So, as the sample grows, the number of parameters grows at the same rate, and standard statistical theory is not valid in such a situation, although it is applied routinely in software that uses this approach. For example, the standard errors reported are not correct. It is strongly advised, therefore, not to use software which uses this method.
3. **Marginal Maximum Likelihood (MML)**. Instead of treating the individual theta values of the persons in the calibration sample as individual unknown parameters, we could also treat them as a random sample from a certain population of theta values. For example, we might think that in the

population the theta values are normally distributed, and that the sample we have is a random sample from this population. With this approach the number of parameters is limited: the unknown parameters in this approach are the item parameters and the two parameters of the normal distribution (mean and variance), which are estimated jointly by ML. This is a good and solid approach, but one should realize that in doing this, one has complicated the theory: one not only assumes that the items behave like in Figure G.2, but on top of that we have added the assumption that theta is normally distributed, and that the sample we have is a random sample from that distribution. If the latter assumption is not true, this will affect not only the quality of the estimates of the mean and the variance, but also of the item parameters. An example will be discussed in point 5.

4. **Conditional Maximum Likelihood (CML)**. In this method the parameters are estimated given that the score of each person is known. The concept is quite hard to explain without technical details, and only an intuitive approach with two items will be given. In Table G.2 the (fictitious) frequencies of the four response patterns with two items are given. From the margins of the table it is seen that item 2 is the hardest of the two: it has a *p*-value of 0.33 (100/300), while item 1 has a *p*-value of 0.5 (150/300). But we can deduce conclusions on the relative difficulty of the two items also from the shaded cells. Jointly, these cells indicate the persons who have one of the two items correct. There are 110 such persons, and of these 110 (with the same score on the two-item test), 80 had item 1 correct and only 30 have item 2 correct, indicating that item 2 is the most difficult of the two. The CML-method is based on this kind of comparison, but gets difficult when the test contains more items.

Table G.2. Frequency table for two items

| | | item 1 | | |
|---|---|---|---|---|
| | | 1 | 0 | total |
| item 2 | 1 | 70 | 30 | 100 |
| | 0 | 80 | 120 | 200 |
| | total | 150 | 150 | 300 |

The big advantage of this method is that the parameter estimates are not systematically influenced by the way the calibration sample is composed; it is immaterial whether the sample is a random sample from the population or not. This feature is sometimes called ‘sample independence’. Theoretically it is parsimonious, because it does not require any assumption about the distribution of theta in the population. The disadvantage, however, is that it cannot be applied with all IRT models. It is applicable with the Rasch model, but not with the 2PLM. The reason is that in the Rasch model the score is just the number of correct item answers, while the score in the 2PLM is a weighted sum, the weight being the discrimination parameter of the item. But if we do not know this weight (and we do not before the estimation), we cannot compute the score, and therefore we cannot apply CML, which requires that the score is known.

5. **OPLM**. In the Rasch model all items have the same discrimination. This is a very strict assumption which is almost never fulfilled in practice. On the other hand, being able to use the CML-method is a great advantage, because it frees the test constructor from the burden of sampling randomly from a population that often is not defined very sharply. The way out of this problem is to try to find a model which allows for different discriminations of the items and at the same time makes estimation by CML possible. Such a situation is created by applying formally the two-parameter model, but assuming at the same time that the discrimination parameters are known, i.e., they are no longer an unknown parameter, but just a known constant. This leaves only one parameter per item, although different discriminations are possible. (Hence the acronym OPLM, which stands for One Parameter Logistic Model.) Of course, this does not solve the whole problem: we have to know how to choose these constants, and we have to check whether they are an adequate choice. This is discussed in Section G.3.
6. **Test design**. In some cases the number of items is so large that it is unfeasible to administer every item to every person. So each person in the calibration sample responds to a subset of the items following a certain set up or design. Two examples of such an incomplete design are displayed in Figure G.3.

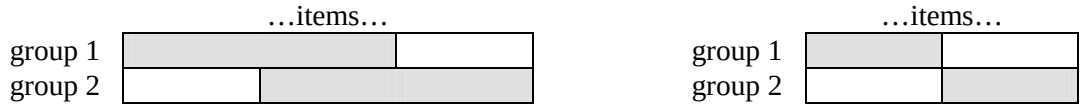


Figure G.3. Two incomplete designs

The groups refer to groups of persons. The shaded areas represent the items that are administered to the groups, the blank areas represent items not administered. There is an important difference between the two designs. In the left-hand panel, some items are administered to both groups. Such an overlap is not present in the right-hand panel. One says that the left-hand design is **linked**, while the right-hand one is not linked. These designs are simple because they involve only two groups. In Figure G.4 two linked designs with four groups are displayed. In the left-hand design a number of items are common to all groups. This set of items is called an **anchor**, and sometimes the design itself is referred to as an anchor design. The right-hand panel has no anchor, but it is linked anyway. Groups 1 and 2 can be compared to each other because they have some items in common; the same holds for groups 2 and 3. Groups 1 and 3 have no items in common, but they can be compared indirectly through group 2. This is why the design is linked: each pair of groups can be compared, directly or indirectly by some common items.

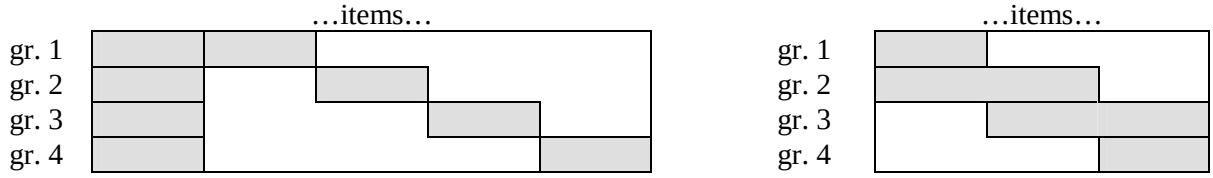


Figure G.4. Two linked incomplete designs

It is important to consider the sampling status of the groups of persons used to administer the items in an incomplete design. We consider two important cases: either the groups are planned to be 'equal', or they are planned to be 'unequal'. By 'equal' is meant statistically equivalent, meaning that the group a particular person belongs to is determined at random. Such a situation arises if there are too many items to be administered to a single person. In such a case both designs in Figure G.4 are suitable. But sometimes the groups are intentionally not equivalent. Suppose the items to be calibrated cover a broad range of proficiency, from A2 to C1, say. Then groups can be chosen in such a way that the items are adequate for their average level of proficiency. In the example of Figure G.4, the groups may be defined in terms of the number of years of instruction; e.g., group 1 having the fewest years therefore gets the easiest items. In this situation an anchor design will probably not be adequate, because the anchor must be administered to everybody. The design in the right-hand panel of Figure G.4 is more suitable.
Here are some rules for the estimation method to be used in different designs:

a CML can be used only with linked designs, be it with statistically equivalent groups or not. It can even be used in cases where some persons happen to belong to several groups. This may occur, for example, in the rightmost design of Figure G.4, if the data are collected at different time points. If the data for groups 1 and 2 are collected this year and for groups 3 and 4 next year, it may happen that the same person (with a possibly different theta value) participates twice. In the estimation procedure such a person is treated as two different persons. One should be careful, however, in administering twice the same items to the same person, because in such a case the effects of proficiency and memory are confounded, and if there are strong memory effects, the estimates of the item parameters will be distorted systematically.

b MML can be used with linked and not-linked designs, but one should be careful, because the technical feasibility of the estimation procedure does not necessarily guarantee valid results. We consider a number of cases:

   i) If the groups are statistically equivalent (they represent the same population), then a design like in the right-hand panel of Figure G.3 can be used: there are no common items, but the items in the two subsets are comparable because they are administered to comparable groups.

ii) If in the same design, the groups are not comparable, then it is unrealistic to assume that both groups come from the same population. In such a case, we could assume that there are two populations where the latent variable is normally distributed (and then we have to estimate two means and two variances). But in a non-linked design this is technically not feasible, and intuitively it should be clear why not: if group 2 obtains a higher average score on its test than group 1 on a completely different test, the difference could be explained by a difference in average proficiency or by a difference in difficulty of the two tests, and logically there is no way to distinguish between these two sources.

iii) If one uses non-linked designs, one is forced to apply MML (CML not being feasible) and to assume that the groups are equivalent. But what if they are not equivalent? Forming equivalent groups is a risky undertaking, and in principle there exists only one good method: randomization (e.g., tossing a coin to decide if John is going to group 1 or to group 2). But real randomization can be very impractical. Suppose one wants to administer a listening test with the stimulus text coming from loudspeakers. In an incomplete design with good randomization, this may mean that one half of a class has to listen to different sample texts than the other half, such that simultaneous testing is practically impossible. But serial testing may not be liked by the school. The practical solution in such a case – administer the same test to the whole class – will in all likelihood jeopardize the statistical equivalence of the two test groups (even if they 'look' comparable: randomization is a job for coins and dice, not for human judgment). If one proceeds anyway with MML, the estimates of the item parameters will by distorted in a systematic way: the difficulty of the items administered to the weakest group will be overestimated, and the difficulty of the other items will be underestimated, implying that the difference in the average difficulty of the two tests will contain a systematic error (called bias). This bias may be considerable. Therefore it is good practice to use linked designs as much as possible.

7. **The concept of information**. The discussion about test designs in the preceding paragraph might lead to overoptimistic ideas ("my design is linked, so nothing can happen to me"). A simple example will show this. Suppose a test consisting of items at C1 level is administered to A2 students. We will then probably observe very few correct answers, and the only valid conclusion we can draw from this observation is that the test is too difficult for the test takers. It will not be possible to estimate to an acceptable degree of accuracy the differences in difficulty between the items. This means that the answers obtained convey very little information about the items. In statistical theory the concept of information is defined rigorously, and it can be quantified. Technical details are discussed in Section G.7; here we discuss some features that are relevant for testing practice:
   a. The concept of information is related closely to the standard error of the estimates. The amount of information equals one divided by the square of the standard error. For example, if the standard error equals 0.4, the amount of information about the item parameter equals $1/0.4^2 = 6.25$.
   b. The amount of information provided by an answer is largest when the probability of a correct answer is 0.5. If the probability of a correct answer is near zero or near one, very little information is collected.
   c. In the Rasch model (when all discrimination parameters are equal to one), the maximum information coming from a single observation equals 0.25 (see also Section G.6).
   d. Information is additive. This means that the information provided by the answers of John may be added to the information provided by the answers of Mary. This holds only if the answers of John and Mary are independent of each other. (If John copies Mary's answers we have no new information).
   e. Combining a and d above shows that the standard error of the estimates will get smaller the larger the sample size is, but point b shows that not every person in the sample has an equal contribution to the total amount of information. This is important in planning the test design: to get accurate estimates of the item parameters, the difficulty of the items should correspond to the proficiency of the test takers. To accomplish this, the test constructor should have a

priori a rather good idea of the difficulty of the items and of the level of the intended calibration sample.

f. The relation between amount of information and the standard error of the estimates is an important one. If the sample size is doubled, the amount of information will (roughly) be doubled also, but the standard error of the estimates will not be halved, i.e. it will no be ½ of the original standard error, but only the square root of ½ (which is 0.7 approximately). To halve the standard errors, the sample size should be quadrupled. This relation is sometimes denoted as the square root rule.

g. The estimation of the difficulty parameter of an item is not possible if its observed *p*-value in the calibration sample equals zero or one.

8. **The concept of calibration**. If one buys a kilo of meat at the butcher's, the butcher places the meat on a balance and the customer can read the weight of the meat from a gauge. If the needle indicates one kilo, the customer trusts that the meat weighs really one kilo. This trust is based on the knowledge that the balance has been **calibrated** (in the old days by an inspector of weights and measures), i.e., it has been verified that the indicated weight corresponds to the real weight. The idea of calibrating a set of items has a similar meaning, but things are sometimes less evident than they seem to be, even at the butcher's. Two important concepts are discussed: unit and origin of the scale.

   a. **The unit of the scale**. In common social talk an utterance like "the weight of the meat I bought is one" is not acceptable, and will probably be followed by the question "one what?". But when one says that the difficulty parameter of an item equals 2, we should ask the same question: "2 what?", or more generally, what is the unit of measurement? This is not an easy question to answer. In principle the unit is arbitrary, and there is no internationally accepted standard, like for weights or lengths, and even stronger, there cannot be one, since the theory is built to measure concepts of different nature. It is a meaningless question to ask whether one unit in language proficiency is the same as one unit in attitude, just as it is meaningless to ask if one kilo is more or less than one meter. To interpret the unit of measurement, we need a comparison on the same scale. A good standard to compare with is the standard deviation of the underlying trait in the target population. Here is an example: suppose item one has a difficulty parameter of 1 and item two has a difficulty of 2. Suppose further that the measured proficiency in the target population has a mean of zero and a standard deviation of 0.8. Then we can say that the two items lie 1.25 ( = 1/0.8) standard deviations apart, or, equivalently, that the unit of measurement on the scale is 1.25 standard deviations of the target distribution.

   b. **The origin of the scale**. Weights and lengths are measured on a ratio scale, meaning that we can choose the unit of measurement arbitrarily, but not the origin: it is clear and unambiguous what we mean by a weight or length of zero, irrespective of the unit we use. But if we say that the temperature is zero degrees, we will have to add the specification of the scale used, because zero degrees Fahrenheit is a lot colder than zero degrees Celsius. Scales whose origin (the point or object or item which gets the number zero as its measure) is arbitrary (as well as the unit) are called interval scales. The scales that are constructed with IRT are interval scales, and therefore the origin can be chosen freely. Of course, to have meaningful communication, we have to fix in some way the origin and tell other people how we did choose the origin. The specific way in which the origin is chosen is called **normalization** (a confusing term, which has nothing to do with the normal distribution). Common ways to choose the normalization are: (i) defining the difficulty parameter of a specific item as being zero; (ii) defining the average difficulty of all the items in the test as zero and (iii) defining the mean proficiency of the target population as zero. Of course, only one of these definitions can be chosen.

### G.3 Check your narrative

One of the most attractive advantages of IRT is the possibility to carry out meaningful measurement in incomplete designs: it is possible to compare test takers with respect to some proficiency even if they did not all take the same test. The most pronounced case of this is Computer Adaptive Testing (CAT), where the items are selected during the process of test taking so as to fit optimally with the level of proficiency as currently estimated during test taking. To apply CAT or some more modest application

where incomplete designs are used, requires a lot of technical know-how. This is sometimes packed in nice looking software, and some users of this software may think that the problem is nothing more than technical know-how. This is however a naive way of thinking: the advantages of IRT are only available if the theoretical assumptions on which the theory is built are fulfilled. Therefore it is the responsibility of all users applying IRT to check as accurately as possible these assumptions.

In a deterministic model, a check is relatively easy. The model predicts exactly what can happen and what not. Finding a single case that is not predicted by the model is enough to reject it. In probabilistic theories, by contrast, the checking is more difficult. The models are built in such a way that almost everything is possible; for example, it is theoretically possible that a test taker with very low proficiency has all items of a difficult test correct, just as it is possible that a fair coin when tossed one thousand times lands 'heads' every time. Yet, if the latter event happens, we will not accept that the coin is fair (and the tossing was done without cheating), and we do so on statistical grounds: the observational outcome is so **unlikely** under the **hypothesis** (that the coin is fair and the tossing has been fair) that we reject the hypothesis. The checking of IRT models follows the same rationale, although the hypothesis is much more complex than the hypothesis in a coin tossing experiment. Before discussing statistical tests in some more detail, we give a small example of a statistical test as it is used in the program package OPLM. Although the result of a test is usually a number (a t-value or a chi-square value, possibly decorated with one or more stars to indicate the level of significance), in some cases it is possible to construct a graph which can be much more informative than a single number. Two such graphs are shown in Figure G.5, and will be commented upon.

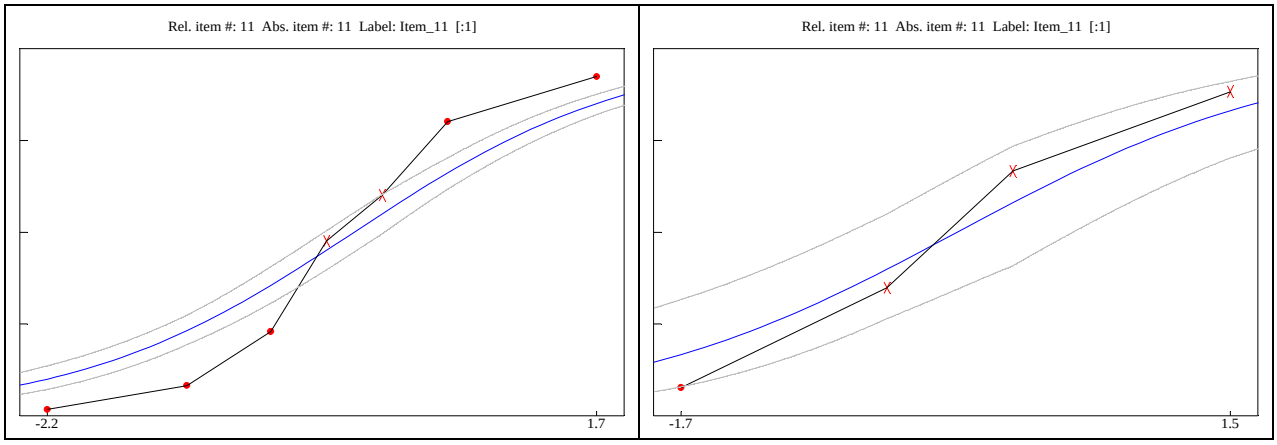


Figure G.5. Statistical tests for a single item

The graphs result from an analysis on an artificial data set, which has been constructed with the explicit purpose of showing several characteristics of statistical tests. The artificial tests contains 21 items, all equally difficult. Twenty items comply with the Rasch model; in particular this means that they all discriminate equally well. One item, however, discriminates better than the other twenty. So the 21 items taken jointly do not comply with the Rasch model. (The deviating item is number 11). Starting from known item parameters, artificial data may be created. For the example, 3000 artificial persons were submitted to the test (this is accomplished by running a rather simple computer program), such that as a result we have a data set with the answer of 3000 persons to 21 items. The next step is to analyze this data set without making use of the knowledge we have of the real parameters. Thus the data set was analyzed using the Rasch model; more formally we can say that we use the model as a hypothesis. It is important to realize that the estimation procedure in the software does 'not know' that the Rasch model is not valid; it is nothing else than a mechanical handling of numbers, designed to solve a set of (complicated) equations. If the program is (technically) successful, this means nothing else than that the equations are solved, but it does not follow in any way from this that the model is valid.

After the estimation, however, we can do something which is not possible in Classical Test Theory. If we know the item parameters of the Rasch model, then we can compute the probability that somebody

with a score of 15, say, will have a correct response on item 11, say. (This computation is rather complex, but the software takes care of this.). Suppose that this probability is 0.6. This means that we expect that in the group of students with a test score of 15, 60% will give a correct response to item 11. But this percentage is observable: we can find in the data set all students with a score of 15, and in this subgroup we can count the number of people with a correct response to item 11. Suppose that 96% of these students have item 11 correct, a lot more than predicted by the model. This means that the observations (the observed percentage) do not correspond closely to what we predict; so our prediction is wrong. But the prediction follows mechanically from the assumption of the Rasch model, and therefore the Rasch model must be wrong. In Classical Test Theory a similar procedure is not possible, because there is no way to predict how students with a score of 15 on the test should behave on item 11; the theory is so weak that it cannot make any such prediction.

The procedure described in the preceding paragraph can of course be applied also to the group with test scores 1, 2, 3 and so on up to the highest possible score. But if we do this for all scores, we construct a table with predicted and observed percentages correct, and from this table we can construct a graphical representation. This is essentially what is displayed in the left-hand panel of Figure G.5. But there are some more things to be said on this:

1. With 21 items, 22 different test scores can be obtained (0 to 21). But if your test score is zero, the probability that you have item 11 correct must also be zero, and it is impossible to find a person with a test score of zero and item 11 correct. So in this case, the predicted and observed percentages correct are zero by definition, and this case is uninformative. The same holds for the group with the maximum test score, where observed and predicted percentages correct must equal to one hundred. So these two scores can be discarded.
2. With the remaining scores, 20 groups can be formed, but in cases where the sample size is rather modest, some of these groups will contain very few test takers, with the consequence that the constructed graph may look quite erratic. Therefore, groups of scores are defined, much as in the technique of graphical item analysis – see Section C. The groups are formed in such a way that they contain (approximately) an equal number of test takers. In the example, seven such groups have been formed.
3. For each group the predicted percentage of correct answers on item 11 is computed. This percentage can be plotted against the group number. The plotted points can then be connected by lines. If the connecting lines are smoothed, one smooth line of predicted percentages will occur. In Figure G.5 this line is the middle one of the three smooth lines (blue if color is available).
4. In each of the seven score groups one can count the number of people with a correct response to item 11, and convert this number to a percentage. In Figure G.5 these percentages are plotted as crosses or bullets, and then connected by straight lines to give visual support. This curve with broken lines is sometimes referred to as the empirical item response curve. Notice that it is the same curve that is constructed when applying techniques of graphical item analysis.
5. Essentially, the test consists of a comparison of the empirical and the predicted curves. Clearly, in the left hand panel of Figure G.5, the two curves differ markedly from each other, meaning that the predictions are grossly wrong. But the problem is to have a clear definition of what we call ‘grossly wrong’. In the software package OPLM two tools are available which can be helpful in judging the discrepancies between predicted and observed percentages. These are discussed next.
6. Suppose there are 500 students in the sixth score group, and the predicted percentage of correct responses in this group is 80. If the model is correct, we expect 0.8 x 500 = 400 correct responses in this group, but this is not the same as requiring that **exactly** 400 correct responses should be observed. Everybody will agree that we should observe about 400 correct responses. But what do we mean by ‘about’? What one can do, for example, is to define a 95% confidence interval around the expected value of 80%, and require that the observed percentage falls within this interval. If such an interval is defined for all score groups and the upper and lower bounds are plotted and then connected by a smooth line, a kind of envelope around the theoretical curve results. In the left-hand panel of Figure G.5 the two outer smooth lines (gray in a colored figure) define this envelope, and now we see clearly that five of the seven observed percentages fall outside the envelope, indicating clearly that the behavior of item 11 is quite different from what the model predicts. (Observed percentages falling outside are plotted as bullets, those inside as crosses.)

7. The left-hand panel of Figure G.5, however, is an easy case: the difference between the two curves is so marked that it hits the eye, and a correct conclusion would also be drawn without the aid of the envelope. But things become more complicated if six of the seven observed percentages fall within the envelope and one lies (a little bit) outside. What we need in such a case is an answer to the question whether the difference between the predicted and the observed curves – both considered as a whole – can be attributed reasonably to random fluctuations, given that the Rasch model is the correct model. To do this we need a more formal criterion, which is provided by a statistical test. In the present case a quantity, labeled $S_{11}$ (because it is concerned with the $11^{th}$ item) is computed from the differences between the two curves. Its value is 180.3. It can be compared to a so-called critical value in the theoretical chi-square distribution (with 6 degrees of freedom). At the 5% level of significance this critical value is 18.55. Since the observed value is larger than the critical value, the hypothesis that the difference is due to random fluctuations is rejected.
8. The added value of a graph like Figure G.5 is that it does reveal that the Rasch model is not a valid model here, but it gives also information why this is so. The empirical curve is much steeper than the predicted one, indicating that the item discriminates better than predicted by the Rasch model.
9. The confidence envelope in the left-hand panel of Figure G.5 is quite narrow. The reason for this is that the number of test takers in each group is large (on the average 3000/7 = 429). The sample size has a definite influence on the outcome of the statistical test. To illustrate this, a random sample of 175 test takers was drawn from the original 3000 artificial test takers, and the responses of this small sample was analyzed in the same way as the original sample. The graphical outcome of the statistical test for item 11 is displayed in the right hand panel of Figure G.5. We see immediately that the confidence envelope is much broader now, and we also notice that the empirical curve falls within the envelope, with just one borderline group. The statistical test yields a non-significant result. The value of $S_{11}$ equals 4.89 while the critical chi-square value with 3 degrees of freedom is 12.84. (With such a small sample size only four score groups are formed; the number of degrees of freedom is the number of score groups minus one.) The important result here is that we do not have sufficient empirical evidence to reject the hypothesis that the Rasch model is valid, although we know it is not, because we work with artificial data which do not comply with the Rasch model.

We generalize this example somewhat and introduce at the same time some important theoretical concepts:

1. In statistical testing, we always test a hypothesis. This hypothesis is called the null hypothesis. In the present example this hypothesis is quite complex and may be worded as follows: "*The 21 items together comply with the Rasch model, and as a consequence the predicted and observed curves for item 11, as given in Figure G.5, will not differ more than can be explained by random fluctuations*."
2. Although random fluctuations may cause big differences, we will reject the null hypothesis if the difference is very big. The notion of 'very big' is formalized in statistical theory as follows: From the difference between the two curves, a certain quantity can be computed which we label here as $S_{11}$. **If the null hypothesis is true**, we know from statistical theory that there is a probability of 5% that the quantity will have a value which is larger than the critical value of 18.55 (when we use 7 score groups). We may take that risk of 5%, and decide that we will reject the null hypothesis if we observe indeed that $S_{11} > 18.55$. It is important to understand that this risk only applies if the null hypothesis is true indeed; but we do not know this in general. Moreover, the risk of 5% is widely accepted in the scientific community, but in principle it is arbitrary. This risk level is called the **level of significance**.
3. The computation of the quantity $S_{11}$ is technically quite complex (one cannot check it quickly on a piece of paper), and the mathematical proof that one can use the critical value of 18.55 (or more generally, that one can use the tables of the theoretical chi-square distribution) is quite complex, and will not be discussed here.
4. The preceding, however, tells only half of the story. It was used to find a decision rule, which is based roughly on the following rationale: "*If the null hypothesis is true we will (often) find a small*

*value for $S_{11}$, but if the null hypothesis is not true, it is more likely to find big values. So let us decide now that we reject the hypothesis if we find a big value and we do not reject if we find a small value.*" In the preceding paragraphs, it was admitted that we can find also big values if the hypothesis is true, but we have a calculated risk: we set the decision rule (the borderline point between 'small' and 'big') such that we make the wrong decision in only 5% of the cases if the hypothesis is true. But we still have to discuss the risk if the hypothesis is not true.

5. This is a much more complex situation: if the 21 items jointly do not comply with the Rasch model, this may be so for many reasons. In the example, it was told what the reason was: 20 items did comply with the Rasch model, and just one item discriminated better than the others. But even in this case, we are not fully informed: it may be that item 11 discriminates just a tiny little bit better than the other items, or it might discriminate much better. In the former case, it is not reasonable to expect that big values for the quantity $S_{11}$ are very likely, while in the latter case big differences will be much more likely. Suppose that in the former case there is a probability of 6% to find an $S_{11}$ quantity larger than 18.55, while in the latter the probability is as high as 88%. But this means that in the former case the false null hypothesis will be rejected in only 6% of the cases. This means that with our test we only have a probability of 6% to **detect** a deviation from the Rasch model, i.e., to reject a false null hypothesis, while in the latter case this probability is 88%. The technical term to denote the probability of rejecting a false null hypothesis is called the **power** of the test. It is important to realize that the power depends on the degree of deviation between the actual test and the model to describe it, i.e., the degree of deviation between the real world (what we really observe) and our narrative about the world.
6. But the degree of deviation is not the only factor which influences the power of a statistical test. In the example of Figure G.5 the reality for the left hand panel is just the same as the reality for the right hand panel. The fact that we found a significant result, i.e., really detected that the Rasch model was not valid, with a big sample, and not with a small sample is not a mere coincidence. It is a statistical law that the power of a statistical test increases with increasing sample size. This is the main tool by which a researcher can manipulate the power of the statistical tests he wants to use. We will come back to this point in later paragraphs.
7. **Sometimes one hopes to reject the null hypothesis.** Historically the first applications of statistical hypothesis testing were in agronomy. To show that a fertilizer is effective, a simple design like using no fertilizers on an number of plots and using a certain dosage of fertilizer on an equal number of plots, and comparing the crops (using a statistical test) under both conditions, may lead or not lead to the conclusion that using fertilizers is effective. In such a set-up it is hoped for that fertilizers are effective indeed – this is the research hypothesis. The statistical hypothesis, however, is the denial of this research hypothesis, and it was hoped that this hypothesis could be rejected. The technical name of such a complementary hypothesis is called **null hypothesis**, and the research hypothesis is often called the **alternative hypothesis**. In statistical testing it is always the null hypothesis which is tested, and in experimental science, it is usually hoped that it will be rejected. If it does not succeed (the test result does not yield significance), this is not to be taken as strong evidence that the null hypothesis is true, but as a lack of empirical evidence to demonstrate the truth of the research hypothesis. This can be understood by using the concept of power: it is possible that the effect of fertilizers is positive, but rather small (perhaps because the dose is too low). If at the same time the number of plots used in the experiment, i.e., the sample size, is rather modest, the test used may have little power, i.e., the probability of rejecting the null hypothesis may be very low.
8. **But sometimes one does not hope to reject the null hypothesis.** When one uses an IRT model, like the Rasch model, the model itself is the research hypothesis. Users of such a model may like it because it is parsimonious and gives a description of (part of) the reality in quite simple terms. But such a model is not valid just by positing it; it must be tested, just like a newly designed car must be tested. With probabilistic models, the tests are statistical, but the important difference with experimental research is that the model itself is the statistical null hypothesis, and thus it is in the interest of the proponents of the model **not** to reject the null hypothesis. Although the technical machinery (the formulae, the way of reasoning, the use of statistical tables, etc.) is just the same as with testing in experimental research, the general context is essentially different. Statistical tests used to show the adequacy of a probabilistic model borrow their strength by showing that the

observations, or some aspects of it, fit well with the predictions ensuing from the model. Therefore they are usually called **goodness-of-fit** tests. A non-significant result is often interpreted as evidence in favor of the model, but one should be very careful with such a reasoning. One could use a test with almost no power (for example by using a very small sample size), such that one is almost sure that no significance will be found. Of course this is not strong evidence in favor of the model, although sometimes it is presented as such.

9. There exist many different tests of goodness-of-fit for the Rasch model or other IRT-models. In the preceding example with the artificial data, the deviation between the (artificial) reality and the Rasch model concerned the equality of the discriminative power of all items. The $S_{11}$ quantity was designed especially to be sensitive for differences in discrimination of item 11 compared with the average discrimination of the other items. But of course, a similar quantity can be computed for the other items as well ($S_1$ for item 1 up to $S_{21}$ for item 21), and all these quantities can be used in a similar statistical test, which in general tests the validity of the Rasch model for the 21 items. But of the 21 tests (which all were carried out in the analysis with 3000 test takers), only $S_{11}$ yielded a significant result. If we repeat the whole procedure a thousand times, i.e., if we construct 1000 samples of 3000 artificial respondents, it is very probable (and indeed this has been done), that we will get a similar result in the majority of the cases: $S_{11}$ leading to a significant result and the others not or a very few times (in fact a little bit more than 5% of the cases for each of the other tests). This means that the test based on $S_1$, for example, has very little power to detect the deviation from the Rasch model, while the test based on $S_{11}$ has very much power.
10. Differences in discrimination, however, are not the only possible reason why the Rasch model may be invalid. An important assumption of the model is unidimensionality. This means that all items should be indicative jointly of just one underlying latent variable. Now suppose that a test for English is constructed which consists of 20 reading items and 20 listening items by a researcher who is convinced that the distinction between reading and listening is just a matter of convenience but has nothing to do with really different proficiencies, i.e., he is convinced that in the target population the proficiency for reading and for listening have a correlation equal to one. Notice that this is not a trivial problem, and the researcher's hypothesis cannot be refuted simply by showing that the correlation between reading and listening test scores (as observed in the sample) is less than one; see the discussion on attenuation in Appendix C. A possible approach, which in fact is used quite often in the social sciences, is 'to show' that the reading and listening items jointly comply with the Rasch model, or some other more complicated but still unidimensional IRT model. The demonstration is usually carried out by applying a series of statistical tests which happen to be available in one's favorite software package for IRT. If this package happens to be OPLM, there is little chance that the model will be rejected, even if in reality the correlation between reading and listening is substantially lower than one. The reason is that the tests implemented in OLPM have little power against multidimensionality. If this is combined with a moderate sample size, probably not a single test will lead to significance. But as a demonstration of the 'truth' of the researcher's hypothesis, the whole procedure is not convincing.
11. The preceding paragraph may look disappointing, and in some respects, it is. For many widely used statistical tests in IRT there is little or no insight into their power characteristics. This topic has been neglected widely, in research as well as in education. In some introductory statistics books the concept of power is not even introduced. And the technical complexity to carry out a statistical test probably leads to obscuring the necessity of power considerations. Yet, technicality and quality are not synonyms. Sometimes it is much more convincing to bring about evidence by simple means than by some highly sophisticated technique which is beside the point. The researcher referred to in the previous paragraph would be better off if he used a technique which is especially designed to uncover a multidimensional structure, such as factor analysis.

The main points of this section are summarized below.

| |
|---|
| 1. An IRT-model is a hypothesis about how the data come about. Its validity (appropriateness) must be demonstrated. |

2. Since most IRT models are probabilistic, the test of the model will be mainly based on statistical tests.
3. Formally the model and specific consequences following from it have the role of null hypothesis in the statistical test.
4. Most tests try to demonstrate that predictions following from the model are in good correspondence with the data. If they are, this can be taken as evidence in favor of the model.
5. An important concept in statistical testing is power, the probability that one can demonstrate (by a significant result) that the model is not valid. The most important tool to manipulate the power is the sample size: the larger it is, the more power.
6. Since the model is complex, it may be defective in several ways. Particular tests are sensitive to some defects but not to others. It is good practice to apply all statistical tests available in the software one uses. Professional assistance may be needed for a correct interpretation of the results.

**G.4 Go and measure**

The preceding sections on estimation and statistical testing are concerned with the construction of the measurement instrument, and the demonstration that the theory underlying the model is valid for describing the test behavior of test takers from the target population. If the evidence is strong enough to justify the conclusion that the model is trustworthy, then one can proceed to use the test as an instrumental tool. In terms of the model, this means that the answers of a test taker are used to make an estimate of his position on the underlying continuum, i.e., to make an estimate of the person's theta value. This estimate is usually computed by the same software that is used for doing the calibration. In section G.6 some technical details on these estimates are discussed. In the present section we will treat some topics of a more conceptual nature.

1. The estimate of a person's theta value is not equal to the real theta value. The estimate is based on the response pattern of the test taker. The theta value itself is considered as a stable characteristic of the person, but if the test is administered twice (assuming in-between 'brain-washing') it is not very likely that we will observe twice the same response pattern, and therefore we will probably end up with two different estimates of the same theta value. The accuracy of the estimate is expressed by its standard error. Usually the standard error is larger for response patterns with an extreme high or an extreme low score than for response patterns in the middle of the score range. This has to do with the concept of information: if a test is too difficult for John, he will probably end up with a low score, but the amount of information collected by the responses is low. So, essentially, what we learn is not much more than that the test is indeed too hard, but we cannot infer with high precision the location of John's position on the underlying continuum, and this is reflected in a (relatively) high standard error. In Section G.7 it will be explained how this information can be computed.
2. In the section on estimation, it was explained that the amount of information we collect on an item parameter will increase as the sample size increases, because every test taker answering a particular item adds to the information about the item. A similar reasoning holds for the estimation of theta, but we do not collect information on John's theta by the answers of Mary. So, the information on John's theta must come from the answers of John himself, and the only way to get more answers is to make the test longer: The standard error of the estimate of theta depends highly on the test length, but also here does the square root rule apply: to halve the standard error requires four times as many items.
3. To compute the estimate of theta, one needs to know the value of the item parameters, but these values are not known exactly. What is used in the computation are the estimates of the item parameters as they become available in the calibration phase. But these estimates also contain an error, and this error is usually ignored in computing the standard error of the theta estimate. So in fact, the standard error of the theta estimate is larger than reported by the software. If the calibration sample is large, this extra error is not too important, but if the calibration is done on a small sample, the extra error may be considerable.

4. In the Rasch model, all test takers with the same raw score (number of items correct) will have the same estimate of theta; in the two-parameter model, all test takers with the same weighted score have the same theta-estimate.
5. The correlation between the theta estimate and the score is usually very high (even over 0.99). This observation makes many researchers say that using IRT instead of classical test theory has no added value. There is a theoretical and a practical reply to this:
    a. In Classical Test Theory we can learn something about the characteristics of test scores, e.g. their reliability in some population, but the theory by itself does not offer a criterion to judge the meaningfulness of including a particular item from a set of items in the test. For example, it cannot be deduced from Classical Test Theory whether listening and reading items can be combined meaningfully in the same test (yielding a single number as test score), or not. In IRT, this is quite possible, and even essential, because the theoretical construct one wants to measure is at the center of the theory itself. If listening and reading are really two different concepts, then listening and reading items together will not comply with a unidimensional IRT-model. So, in this sense, using a unidimensional IRT model (and demonstrating convincingly its validity) can be considered as the justification to summarize the test performance by a single number. If this number is the test score or the theta estimate is not important, at least if everybody takes the very same test.
    b. The most important practical advantage of using IRT is that one can meaningfully compare performances on different tests. Suppose John takes a reading test consisting of 30 items and obtains a raw score of 22; Mary takes another reading test, consisting of 35 items and gets a score of 24. In the framework of Classical Test Theory there is no rational way to infer from these two observations whether Mary's reading proficiency is higher or lower than John's. In IRT, however, this is very well possible, on the condition that the items of both tests have been calibrated jointly. The comparison usually takes place by comparing John's and Mary's estimated theta. It is precisely this practical advantage that forms the basis for computer adaptive testing.
6. It may be good to end this section with a caveat to overoptimistic proponents of IRT: using an IRT-model does not convert a bad test into a good one. A careless construction process cannot be compensated by a use of the Rasch model; on the contrary, the more careless the test is composed, the greater the risk that a thorough testing of the model assumptions will reveal the bad quality of the test. In this respect, it is important to reconsider the very definition of IRT models: the model says that there is a particular relation between the latent variable and the response probabilities, meaning that somebody with a high theta value has a higher probability of a correct response than someone with a low theta value. But this is a conditional statement: " if somebody with a high theta value takes the item or the test, then etc..". It does not follow from this statement that there actually exists somebody with a high theta and another somebody with a low theta value. To see the implications of this, suppose that in some population the Rasch model is valid for three items, with difficulty parameters of –1, 0 and +0.5 respectively. Suppose further that in this population everybody has a theta value between –0.1 and +0.1. The situation is displayed graphically in the left-hand panel of Figure G.6; the place where the members of the population are situated is marked by a bold piece of the x-axis. In the right hand panel of Figure G.6, we have zoomed in on the first display, just to show what will happen in this particular population, and the remarkable thing is that for the theta-values in this small range, the three item response curves are almost flat. This means that every member of this population has almost the same probability of answering correct each of the three items, but this means the same thing as saying that the expected score on the three items together will be almost the same for everybody. Remembering that expected score is the same as true score in the terminology of Classical Test Theory, this means that the true variance will be very near zero, and thus that in this population the reliability of the test will also be near zero

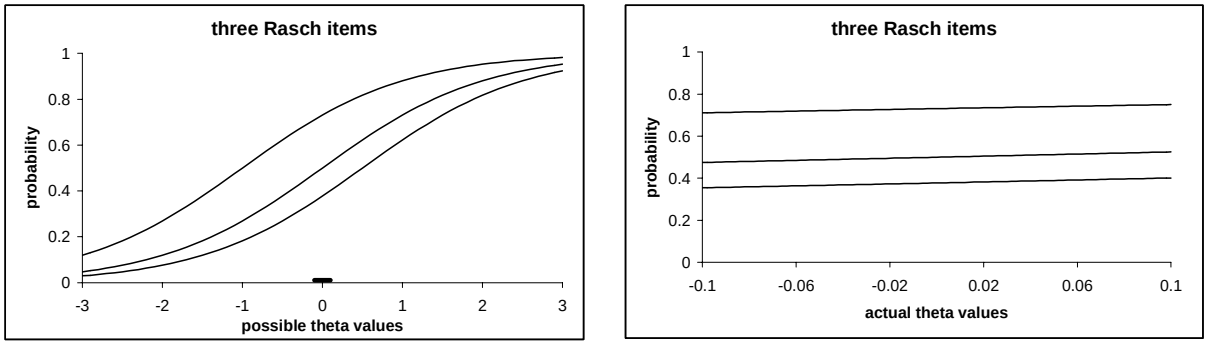


Figure G.6. The Rasch model with different ranges of theta

The important thing to learn from Figure G.6 is that the Rasch model may be valid in a population even if the response curves are almost flat over the range of theta values which are present in this population. But if this is the case the reliability of the test will be very low, and make the test practically useless for individual measurement. The practical consequence is that a separate assessment of the test reliability is needed; it cannot be inferred from statistical tests of goodness-of-fit.

## G.5 The basic equations

The logistic function is a mathematical function which has a very special form. If *x* is the argument of the function, the function rule of the logistic function is given by

$$f(x) = \frac{e^x}{1+e^x} \tag{G.1}$$

where *e* is a mathematical constant which equals 2.71828... (*e* is a very important number in mathematics, so important that it has received its own symbol, the letter *e*.) Notice that in the function rule, *x* is an exponent of the number *e*. Because sometimes the exponent of *e* is not a simple symbol, but a quite long expression, using the notation as above may lead to confusion (we do not see any more that the whole expression is an exponent). Therefore, another way of writing down the very same thing is more convenient, and used quite commonly. Here it is:

$$f(x) = \frac{\exp(x)}{1+\exp(x)} \tag{G.2}$$

The formulae (G.1) and (G.2) are identical, and are said to be the standard form of the logistic function. Notice that it is important to recognize the logistic function. It is the "exp of something divided by one plus the exp of the same something".

In the Rasch model the item response functions are all logistic functions of the latent variable $\theta$. Here is the function rule for these functions

$$f_i(\theta) = \frac{\exp(\theta-\beta_i)}{1+\exp(\theta-\beta_i)} \tag{G.3}$$

We comment on this function rule:

1. The right hand side of (G.3) is the logistic function. The "something", however, is not just $\theta$, but $\theta-\beta_i$. So the logistic function is not in its standard form.
2. The function symbol *f* has a subscript *i* (referring to the item). This means that the function rule for each item can be written as a logistic function. So, (G.3) does not define a single function, but a family of functions.
3. If we look at the rule itself (the right hand side of (G.3)), we see that there is only one entity which depends on the item, i.e., there is only one symbol which has the subscript *i*, namely, $\beta_i$. This is a number, which we leave unspecified here (and therefore it is a parameter). If we choose a value

for this parameter, then we can compute the value of the function for every possible value of $\theta$. If we plot these function values against $\theta$, we get a curve like in the right-hand panel of Figure G.1.

In the two parameter logistic model, the function rule is given by

$$f_i(\theta) = \frac{\exp\left[a_i(\theta - \beta_i)\right]}{1 + \exp\left[a_i(\theta - \beta_i)\right]} \qquad \text{(G.4)}$$

and here we see that the function rule has two entities with subscript *i*, i.e., the function rule defines a family of functions with two parameters. The parameter $a_i$ is the discrimination parameter. It must be positive. If it is very near zero, the curve of the function is almost flat (at a value of 0.5); if it is very big, the curve looks very much like a Guttman item (see the left hand panel of Figure G.1): it increases very steeply for values of $\theta$ which are very close to $\beta_i$. For smaller values it is very near zero, and for larger values it is very near one.

OPLM uses also the function rule (G.4), but in its use it is assumed that the discrimination parameters $a_i$ are known, and do not have to be estimated from the data.

There exists also a model with three parameters which is commonly denoted as the three parameter logistic model. Its function rule is given by

$$f_i(\theta) = c_i + (1 - c_i)\frac{\exp\left[a_i(\theta - \beta_i)\right]}{1 + \exp\left[a_i(\theta - \beta_i)\right]} \qquad \text{(G.5)}$$

Here are some comments:

1. The parameter $c_i$ is a number between zero and one, and is usually called the guessing parameter (or the pseudo-guessing parameter). It can be understood as follows: suppose that $c_i = 0.25$. If the value of $\theta$ is very low (say, -100), then the fraction in the right hand side of (G.5) will be very close to zero, but the function value itself will be very close to 0.25. This may be useful when using multiple choice items. If there are four alternatives in the item, and if the ability is very low, there is still a probability of 0.25 of getting the item right by pure guessing.
2. The function rule of (G.5) is **not** the logistic function. So, designating the model as a logistic model is not justified, but it is often referred to with that name.
3. The model is very popular in the U.S.A. but far less in, e.g., Europe and Australia. An important reason for such reservations is that it is very difficult to estimate the parameters in this model, and that often the estimation procedure fails unless one has very big samples (and this is more common in the U.S. than in Europe or Australia.) There are, however, also more subtle mathematical and philosophical reasons at the base of this 'global' disagreement.

### G.6 The information function of a test

In section G.2, the concept of information was discussed in relation to the estimation of item parameters. It is quite hard to explain this concept further – even graphically- because it concerns the information about many parameters at the same time. Once the item parameters are known (or fixed at their estimated values) and we turn to the estimation of theta, the problem becomes a bit simpler, because in such a case we have only one unknown quantity, namely, theta itself.

Without discussing the mathematical background of the information concept, it may be instructive to look at the formula for the item information in the two-parameter logistic model. Here it is:

$$I_i(\theta) = a_i^2 f_i(\theta)[1 - f_i(\theta)] \qquad \text{(G.6)}$$

and we comment on it:

1 The function symbol is I (for information). It is a function of theta, and every item has its own function, hence the subscript *i*.

2 The function $f_i$ is the item response function as defined by formula (G.3), and $a_i$ is the discrimination parameter of item *i*. The formula is also valid for the Rasch model, because this

model is a special case of the two-parameter model, where all the discrimination parameters are equal to one.

3 The information function is always positive, whatever the value of theta, but it is not constant: it reaches its maximal value in the Rasch model and the two-parameter model if $f_i(\theta) = 0.5$ and this happens if $\theta = \beta_i$. In the Rasch model (where $a_i$ = 1) the maximal information of an item is 0.5 x (1-0.5) = 0.25.

Because of the assumption of statistical independence of the item responses, the information functions for several items may simply be added. Therefore the information function of a test is the sum of the information functions of the items, which, with a formula, can be written as

$$I_t(\theta) = \sum_i I_i(\theta) = \sum_i a_i^2 f_i(\theta)[1 - f_i(\theta)] \qquad \text{(G.7)}$$

where the subscript *t* refers to the whole test. As an illustration, the information functions of the four items in an example test are plotted separately in the left hand panel of Figure G.7. Their sum is plotted in the right-hand panel. The items comply with the Rasch model, and their difficulty parameters are: $\beta_1 = -1$, $\beta_2 = -0.9$, $\beta_3 = 0.8$ and $\beta_4 = 1.1$.

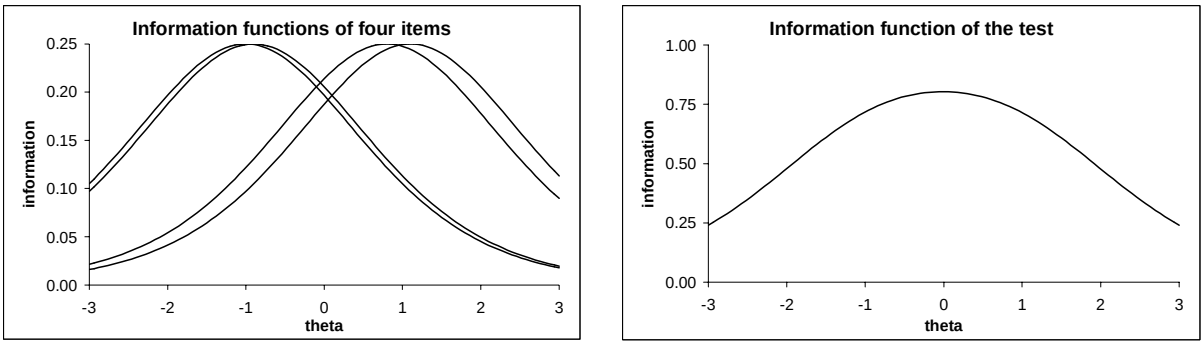



Figure G.7. Information functions of items and tests

We comment on these figures:

1. In the left-hand panel, the four curves reach their maximal value at the value of the item parameters (-1, -0.9, 0.8 and 1.1 respectively). The information value at these points is 0.25 since we are using the Rasch model. We see that the two easy items do convey very little information for high values of theta, and the difficult items have low information for low values of theta.
2. The right hand panel displays the sum of the four curves from the left-hand panel (notice the different scales used for the y-axes in both panels). Its maximum value (about 0.75) is at a theta value near to zero. This is an important observation: none of the four items has its maximal value near zero, but the sum has. We also observe that the curve on the right hand side is flatter than any of the curves in the left-hand panel, meaning that the different contributions of the four test items are spread out along the latent continuum.
3. This finding may be a little bit counterintuitive. Sometimes the argument is heard that, in order to have a good spread of the information the item parameters must be spread evenly. We investigate this a bit more deeply. The preceding example is a test with two (small) clusters of items. In Figure G.8 (left panel) the information function of this test is displayed together with the information function of a four item test with difficulty parameters equal to –1, -0.33, +0.33 and +1 respectively. In the right-hand panel, the information functions for the example test and a four item test with all item parameters equal to zero is displayed. (The curve for the example test is in black, the others are in red and have thicker lines.)

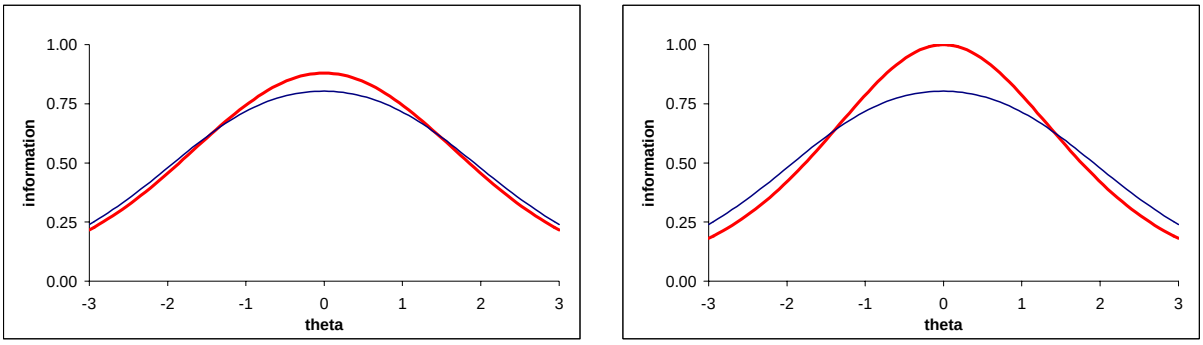



Figure G.8 Comparison of test information functions

4. From the left-hand panel, we see that the information function of the example test, with two clusters of items results in a flatter information function than the test with evenly spread item parameters. In the right-hand panel the curve is fairly peaked at the value of the single common difficulty parameter (zero), while further away the information decreases rather fast.
5. In designing a test, it is useful to construct graphs of the information functions of several tests, and to keep in mind the main use of the test. If the main purpose of a test is selection (such as a decision who failed and who passed in an examination), then the test is best composed of items having their difficulty in the neighbourhood of the cut-off theta value. Suppose one decides that a candidate has succeeded an exam or is accepted for a job if his theta value is larger than zero. Then the best test in the framework of IRT is one with all difficulty parameters equal to zero, because this maximizes the information at that theta value. This means that candidates with a theta value near zero will have their theta estimated with the smallest standard error. For candidates further away from the cutting point, the standard error will be larger, but this is not very important, because for an apt candidate (say with a theta value of 1.5), it does not matter very much if we end up with an estimate of one or two; he will (with very high probability) be accepted anyway.
6. If on the other hand, it is the purpose to estimate the theta value of every candidate as accurately as possible, one is better off with a very flat information function. In the left-hand panel of Figure G.9, a reasonably flat information curve is constructed with 18 Rasch items. The amount of information is at least two (which corresponds to eight maximally informative items) in the range (-2.5, +2.5). If this test were applied in a population where theta is normally distributed with a mean of zero and a SD of one, about 99% of the population members could be measured with about equal accuracy (corresponding to eight to ten optimally adapted items). This may look as an admirable accomplishment, but there is a serious drawback. In the right-hand panel of Figure G.9, the frequency distribution of the difficulty parameters is displayed, showing that 14 of the 18 items are either difficult or easy, and only a minority of four items has medium difficulty. This is what always will happen if one tries to construct flat information functions: the item parameters will contain a cluster of difficult and a cluster of easy items, the items of medium difficulty being a minority.

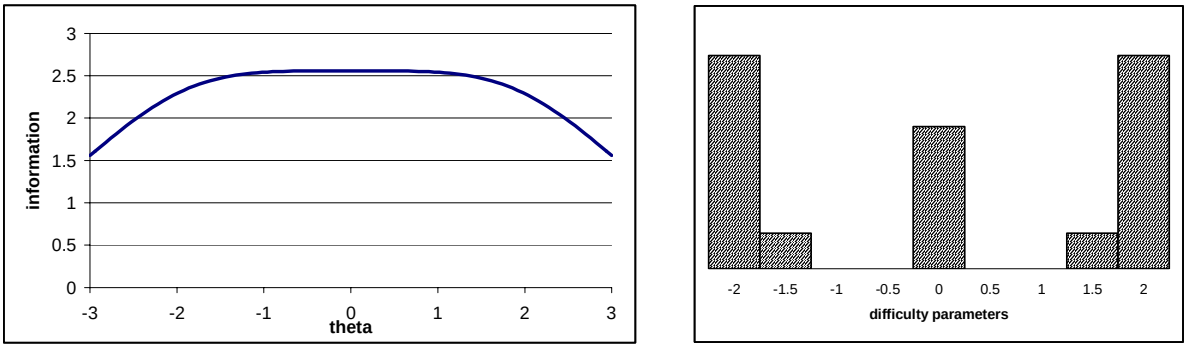



Figure G.9. A flat information function and the distribution of parameters

7. But what does this mean in a practical application? The weak students will be frustrated by the cluster of difficult items and the good students will be bored by the easy items, while in both cases the extreme items – either the easy ones or the difficult ones - will provide very little information. So, it may turn out profitable if we try to construct tests which are more adapted to the level of the test taker. With the foregoing example we might construct an easy test, consisting, for example, of the easy and medium items, and a difficult test consisting of the medium and difficult items. In the left-hand panel of Figure G.10, the information curves for the two tests (each having 11 items) are displayed.

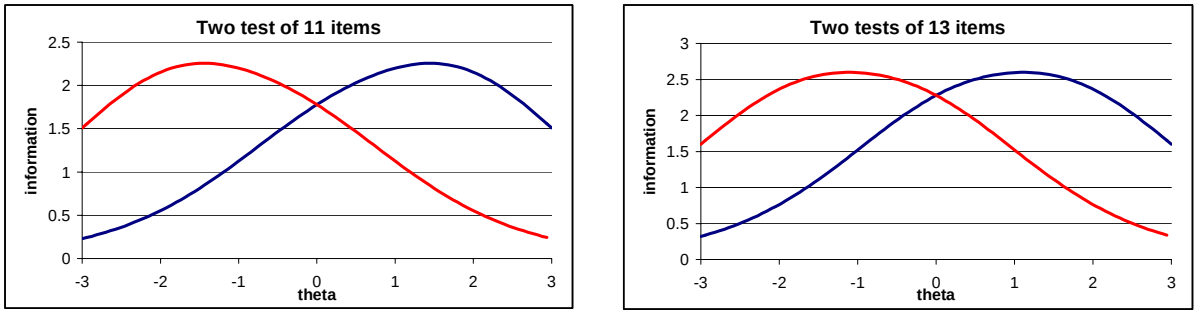



Figure G.10 Information curves for an easy and a difficult test

8. The tests thus composed do not reach the previous level of at least two units of information in a small range around zero. We may repair this by adding one or two items of medium difficulty to each test. The result with two items added is displayed in the right hand panel of Figure G.10.
9. Summarizing then
    a. We have constructed two tests of 13 items each. Both tests have six items in common and seven unique items, giving a total of 20 items.
    b. The easy tests yields information values of at least 2 in the interval (-2.50, +0.42) and the difficult test reaches this value in the interval (-0.42, 2.5).
    c. In the interval (-0.42, +0.42), both tests reach an information value of at least 2, and in a sense, they are exchangeable
    d. If the theta values in the population are normally distributed with mean zero and SD equal to one, about 99% of the theta values falls in the range (-2.5, +2.5). The percentage of people falling in the range (-0.42, +0.42) is 32, about one third of the population.
    e. Of course, we only gain considerably if we succeed in administering the easy test to the weak students and the difficult test to the good students. This means that we need a kind of pretesting to assign students to the easy or difficult test. Because of the safe buffer zone comprising about one third of the population, where it does not matter very much which test is used, things only go wrong if a student belonging to the weakest third of the population is given the difficult test, or the other way around. So, the pretest does not have to be too accurate. In many cases the judgment by the teacher will suffice.
    f. Notice that with these two shorter tests, the estimated theta values from both tests lie on the same scale, and are comparable. Of course this is only possible if the items of both tests were calibrated together.
    g. It may seem that there was something arbitrary in the preceding example, namely, the assumption that the population mean is zero and the SD equal to one. This is true for the example, but in practice it is fairly easy to make a quite accurate estimation of mean and SD using MML in the calibration, and the procedure of the example can easily be adapted to the results. The only assumption that remains arbitrary is the assumption of the normal distribution, but for this application, this is not very important.
10. All the figures in this appendix have been constructed with the program EXCEL, including all the computational work with the formulae. If one masters the basic operations in EXCEL, this goes very quickly. Therefore, it is strongly advised to construct graphs of item response functions and information functions as much as possible, and to experiment with them to see the consequences

of test construction and possible changes in it. For the inexperienced reader, the construction of figures like Figure G.10 will be explained step by step in Section G.8.

## G.7 Estimation of the latent variable $\theta$

Once the calibration phase is successfully finished the item parameters of the items are considered to be known to a sufficient degree of accuracy, and one can say that the measurement instrument is now ready to be used in the field. But the basic observations we collect when administering a test are the answers of a test taker to a number of items, and these answers are converted into item scores. We will stick here to the simplest case of binary scores: the test taker gets a score of '1' for each correct answer and a score of '0' for an incorrect answer. If there are 30 items in the test, our observation consists of a string of 30 zeros and ones, and this string (called the response pattern) must be converted into an estimate of the test taker's latent value $\theta$. The purpose of the present section is to show in some detail how this works.

The problem is not very simple. In fact, there exists several ways of estimating theta values from the observed responses, each having advantages and disadvantages. We will consider three important ways of estimating theta:

1. The maximum likelihood estimator, dicussed in Section G.7.1. In this section the concept of likelihood and of maximimum likelihood (ML) estimation will be discussed in some detail.
2. In Section G.7.2 the concept of bias of the ML-estimator will be explained, and another estimator (the so-called Warm –estimator) which has far less bias will be introduced.
3. In Section G.7.3, at last, an estimator which uses more information than contained in a specific response pattern will be discussed. This estimator fits nicely in a branch of statistics know as Bayesian statistics.

### G.7.1 Maximum likelihood estimation

To use as few formulae as possible, we will use the same example of a four-item test as in section G.6: the test complies with the Rasch model and the item parameters are: $\beta_1 = -1$, $\beta_2 = -0.9$, $\beta_3 = 0.8$ and $\beta_4 = 1.1$. Of course, we do not know the 'true' value of the item parameters, but in practice one uses the estimates of the item parameters as issued in the calibration phase, and treats them as if they were the true values.

Two response patterns will be studied, John's and Mary's. Both have two correct answers and two errors. John's pattern is (0,0,1,1) and Mary's is (1,1,0,0). Mary's pattern looks more like what we would expect; she gave a correct answer to the two easiest items, and could not solve the two hardest. In John's pattern we see just the opposite: he failed on the two easy items, but got the two hard ones correct. So, one might expect that John's response pattern is evidence of higher ability, and that therefore the estimate of John's theta should be larger than Mary's. We will see that this is not the case.

We will investigate the likelihood of John's response pattern. Using formula (G.3) of Section G.5, and substituting the unknown item parameter value by the value we know from the calibration ($\beta_1 = -1$), we find

$$P(\text{item 1 correct}) = \frac{\exp[\theta - (-1)]}{1 + \exp[\theta - (-1)]} \qquad \text{(G.8)}$$

and of course, the probability of an incorrect response is one minus the probability of a correct response:

$$P(\text{item 1 incorrect}) = 1 - \frac{\exp[\theta-(-1)]}{1+\exp[\theta-(-1)]} = \frac{1}{1+\exp[\theta-(-1)]} \qquad \text{(G.9)}$$

We cannot compute from (G.8) or (G.9) the probability that John will have the item correct or incorrect, because we do not know the value of John's theta: the right hand sides of (G.8) and (G.9) are functions of theta. But we can substitute the symbol $\theta$ in these formulae by an arbitrary number and compute the value of the probability. Suppose we use zero for the value of theta, then we find for the probability of a correct answer 0.731 (and 1 – 0.731 = 0.269 as the probability of an incorrect response). So, what we could say then is: if John's theta value were zero, the probability of observing what we did observe (namely, an incorrect response to item 1) is 0.269. We can compute this probability for other values of theta as well, and we can repeat the whole procedure for the other items. This has been done for three values of theta, and the results are displayed in Table G.3, where each row corresponds to an item. Observe that the first column is precisely John's response pattern.

Table G.3 Probabilities and likelihood

| observed resp. | $\theta = -1$ | $\theta = 0$ | $\theta = 1$ |
|---|---|---|---|
| 0 | 0.500 | 0.269 | 0.119 |
| 0 | 0.525 | 0.289 | 0.130 |
| 1 | 0.142 | 0.310 | 0.550 |
| 1 | 0.109 | 0.250 | 0.475 |
| likelihood | 0.004063 | 0.006025 | 0.004042 |

In the preceding paragraph it was explained how to determine the probability of an observed response for a single item. But there remains to determine the probability of a whole response pattern, i.e., the probability of the four observed responses **jointly**. To do this in general is not an easy problem, unless a special assumption is introduced. This assumption is the assumption of **statistical independence**. In the present context it says that once the value of theta is given, the probability of a correct response on some item does not depend on the responses given to the other items. More concretely: suppose John's theta value equals –1, then the probability that he will have the fourth item correct is 0.109, whatever his responses have been on the other items. This assumption is omnipresent in IRT (and in many other models as well), and if it is fulfilled, then we have a very simple but powerful rule: the probability of a response pattern is just the **product** of the probabilities of the item responses. These products are displayed in the last line of table G.3. They are called the likelihood of the observed response pattern.

In Table G.3 the likelihood is displayed for three different values of theta. We see that the likelihood values are small numbers, but this is not important; the important thing is that the likelihood values change as theta changes. This means that the likelihood is a function of theta. If we compute the likelihood for many values of theta, we can display the function graphically. This is done in the left-hand panel of Figure G11 for John's response pattern. In the right-hand panel, the likelihood function for Mary's response pattern is displayed

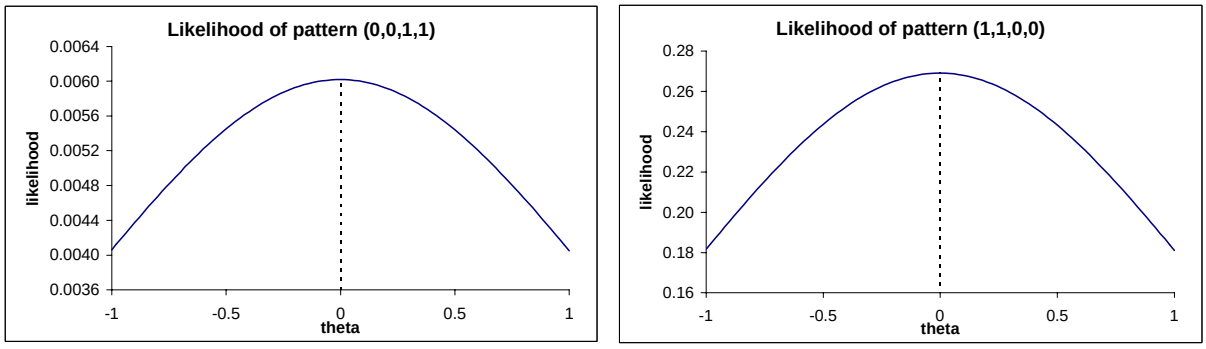



Figure G.11. Likelihood functions for two response patterns

We comment on this figure:

1 If one moves from left to right along the x-axis, the likelihood function of John's response pattern first increases and then decreases; it reaches its maximum at a theta value of about zero (a more fine grained computation reveals that the maximum is at –0.0022). Therefore –0.0022 is the **maximum likelihood** estimate of theta for this response pattern.

2 In IRT-software where the maximum likelihood estimate is computed, special mathematical techniques are used to find the estimate quickly (also in the case of many items). It is not necessary, however, to master these techniques to understand what a maximum likelihood estimate means.

3 The right-hand panel of Figure G.11 is the likelihood function for Mary's response pattern. The curve has exactly the same form as the curve for John. Therefore the maximum likelihood estimate of Mary's theta is also –0.0022, the same as John's.

4 The equality of John's and Mary's estimates is not a coincidence. In the Rasch model it holds that all response patterns with an equal number of correct responses get the same maximum likelihood estimate. This means that in the Rasch model (i.e., when the Rasch model is valid), all information about a person's theta value is contained in the raw score, and that, consequently, no rational consequences can be drawn from the observation that John got the two most difficult items correct and Mary the two easiest ones.

5 One should be careful, however, not to turn the argument around and to say that all possible response patterns with the same raw score are equally likely. This can be seen from a careful comparison of the two panels in Figure G.11. The **form** of both figures is the same, but the likelihood values are quite different. For a theta value of 0.5, for example, the likelihood of Mary's pattern is 0.24324, while for John we get a value of 0.00544. (Compare the numbers written next to the y-axes in both panels of Figure G.11.) The ratio of these two values is 44.7, meaning that the pattern (1,1,0,0) is 44.7 times as probable as the pattern (0,0,1,1). This holds at a theta value of 0.5, but it holds also at all other theta values. If the Rasch model is valid in a population with the $\beta$-values as given above, and we draw a huge sample of response patterns from this population, we should observe that the pattern (1,1,0,0) occurs about 44.7 times as often as the pattern (0,0,1,1). If these two patterns were about equally frequent in the sample, this would be evidence that the Rasch model is not valid.

6 A comparison like in the preceding paragraph may be useful in some applications. If one takes a test, and gets about half of the items right, then it seems reasonable that the correct answers will be given on the easier items and the wrong answers on the hardest ones. With such a reasoning, John's response pattern may look a bit strange or even suspicious. But we should be careful here, and keep in mind that only a very simple example is discussed. With four items, there are only six possible response patterns with a raw score of two (and we discussed only two of these). With 20 items there are more than 180,000 ways of getting half of the items correct, and with 40 items one can obtain a raw score of 20 in more than one hundred billion ways. So, since it is practically impossible to list the likelihood for all these response patterns, there results a double problem:

   a We need a definition of a 'strange' pattern, such that we can decide for every observed pattern in a sample if it is strange or not. There exists a rather rapid expanding literature on how to define and find 'strange' response patterns. (One such a procedure is implemented in the program package OPLM.)

   b But the most difficult problem is how to draw conclusions from the occurrence of strange response patterns. In high stakes applications (like examinations), cheating behaviour may be an explanation, but one should be careful with such accusations, because sometimes a more trivial (and innocent) reason is the cause of 'strange' response patterns. Here is an example. Suppose a test consists of 60 multiple choice questions, which are arranged (approximately) in increasing order of difficulty. The answers are to be marked by the test takers on two optical reading forms, one form for the items 1 to 30, to be answered before the break, and the second for the items 31 to 60, to be answered after the break. The answer forms have a standard lay-out, leaving room for 40 answers, say, per sheet. John is a bright student but a bit careless. At item 3, he skips a row on his form and marks his answer for item 3 on the place for item 4, and continues to shift a row for the remaining items of the first part. After the break, he starts the second form and makes no mistakes any more. As standard software for reading optical forms

does not check for such skipping of lines (which would be rather difficult in general), John's response pattern will look quite strange, having many errors in the first (easy) part of the test, and few (since John is bright) in the second.

7 In the Rasch model, equal raw scores lead to the same maximum likelihood estimate for theta. In the two parameter logistic model, a similar result holds but now for the weighted score. The weight to be used is the discrimination parameter of the item. In the three parameter model, there is no such thing as a score, and as a rule, every response pattern leads to a different maximum likelihood estimate of theta.

8 From Figure G.11 (and from Table G.3) something can be said about the accuracy of the theta estimate for John and Mary. The estimate contains an error, and the (average) magnitude of the error will depend on the amount of information we collected on John's and Mary's theta. This amount depends on the true value of theta (which we do not know), but it depends also on the number of items, which is small in the example. For a theta equal to zero (which is very close to the maximum likelihood estimate), the likelihood of John's response pattern is about 0.006 (see Table G.3), while at –1 or +1 it is about 0.004. The ratio of these two values is about 1.5, meaning that for a theta value of zero the observed response pattern is 1.5 times as probable than at a theta value of –1 or +1. This ratio is not very impressive. It also means that, when theta moves away from the maximum likelihood estimate (in either direction), the curve drops but not very fast. The rate at which the curve drops when departing from the maximum is an indication of the accuracy of the estimate. To see this more clearly, two likelihood functions are displayed in Figure G.12. The flat one in the left-hand panel is the same as in Figure G.11, the steep one in the right-hand panel comes from a test which has 20 items with the same parameters as the short one, i.e., each difficulty parameter of the short test occurs five times in the long test. The score on the long test is 10. (Notice that the y-values of both curves are in a different unit; the theta-values, however, are common so that the differences in steepness are correctly represented; the ratio of the likelihood at zero and at one in the steeper curve is 7.1. Notice also that the curve of the likelihood function for the long test is very similar to the curve of the normal distribution (and the similarity gets more striking as the length of the test increases). It is this similarity (which is a mathematical necessity) which is used to compute the standard error of the theta estimate in IRT-software.

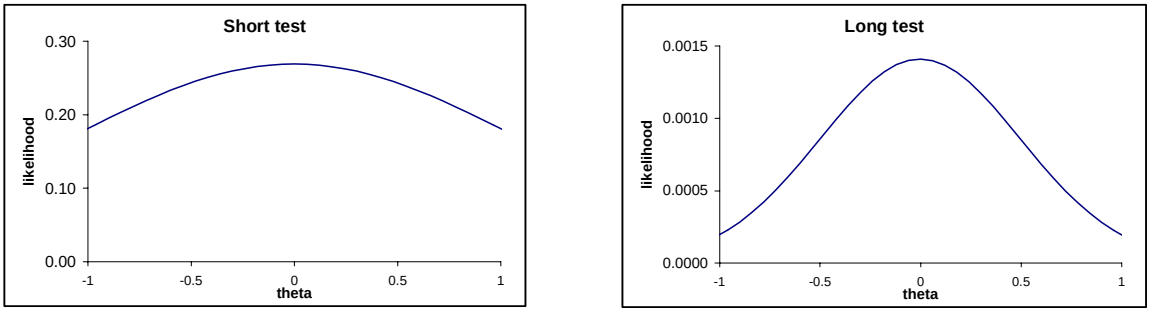



Figure G.12 Likelihood functions for a short and a long test

9 In the left-hand panel of Figure G.13 the likelihood functions are plotted for the response pattern (1,0,0,0) with a score of 1 and the response pattern (1,1,1,0) with a score of 3; their maxima are located at (approximately) –1.33 and +1.33 respectively. In the right-hand panel the likelihood functions for the scores of zero and four are plotted, and here we see that the curves do not have a maximum in the range (-2,+2), but if we make a plot in the range (-10, +10) we will not find a maximum either. This means that these two curves do not have a maximum, or, more generally, for a score of zero and for the maximum score in a test, the maximum likelihood estimates do not exist. The same is true for the two parameter and the three parameter model. Sometimes it is said that the maximum likelihood estimates for zero and perfect scores are at minus and plus infinity respectively, but infinity is not a number. This may cause problems if one wants to compare average theta estimates in two different groups. Each perfect or zero score gives an estimate of plus or minus infinity and these cannot be used in computing the average. Replacing these by a large number or discarding these response patterns are both bad practice. It is better to use other measures in such a case, like the median estimate. But for such comparisons, it is more efficient to

use the MML-estimation method for the item parameters, because it is possible then to estimate at the same time the average theta in the groups.

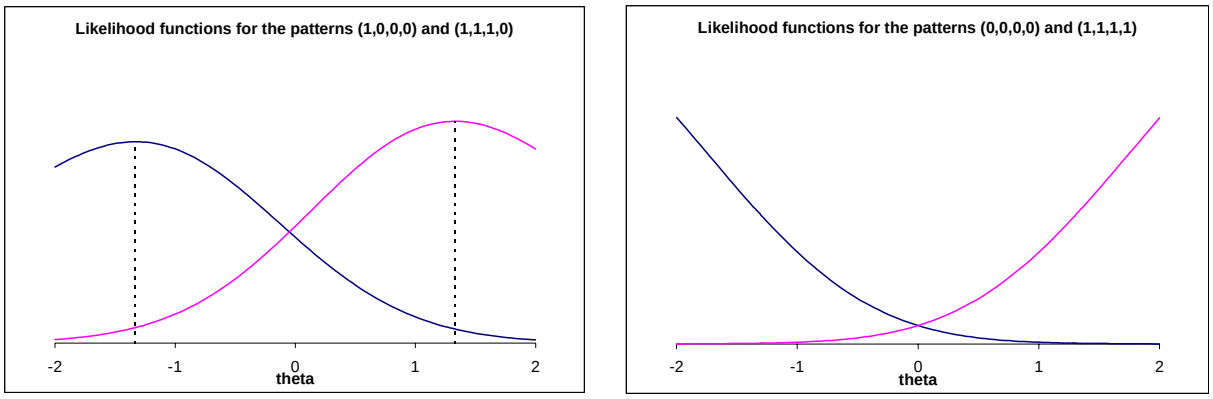



Figure G.13. More likelihood functions

### G.7.2 The bias of the ML-estimator of theta

The maximum likelihood (ML) estimator[1] of theta has two serious drawbacks:

- It does not exist for zero and perfect scores;
- It is seriously biased.

We first explain what is meant by bias in this context. Suppose John's theta value equals +1. He takes a test consisting of five Rasch items. Since the model can only predict the probability of the item responses, and not the responses themselves, it follows that the model cannot predict without error the score on the test. So, with a fixed value of theta, all possible scores (in the example from zero to five) are possible, although not all with the same probability. If the item parameters are known, then it is possible to compute the probability of each score. (The computations are a bit complicated and will not be explained here). In Table G.4 a small example is given, for the case where all five item parameters equal zero. From this table we can infer that there is a probability of 0.384 that John will obtain a score of 4 on this test, but we see also that there is a very small probability that he will fail on all items.

Table G.4 A (fictitious) distribution of test scores for a theta value of +1

| score | P(score) | ML-estimate | Warm-estimate |
|---|---|---|---|
| 0 | 0.001 | (-5) | -2.402 |
| 1 | 0.019 | -1.389 | -1.101 |
| 2 | 0.104 | -0.406 | -0.337 |
| 3 | 0.283 | +0.406 | 0.337 |
| 4 | 0.384 | +1.389 | 1.101 |
| 5 | 0.209 | (+5) | 2.402 |

Notice that the first two columns together constitute the 'private' distribution of John's observed scores as discussed in Appendix C. We can compute John's true score, which is the average value of this distribution. It is computed as

[1] In statistics there is a difference between the terms 'estimator' and 'estimate'. The term '**estimator**' refers to the procedure to be followed to estimate a certain population quantity. The '**estimate**' is the numerical outcome of this procedure in a particular case. So we say that the sample average is an **estimator** of the population mean. If in a particular sample the average is 25, we say that the **estimate** of the population mean is 25.

$$0\times 0.001+1\times 0.019+\cdots+5\times 0.209=3.657$$

But in the framework of IRT, we are not interested in the true score, but in the estimate of John's theta value. As we have seen above, a score on the test results in a certain estimate of theta: if John upon a single test administration would happen to obtain a score of 3, then the estimate of his theta will be 0.406. For a score of zero or five, there is no estimate, but we filled in an arbitrary number of –5 and +5 respectively as theta estimates. Now the two columns of Table G.4, labelled P(score) and 'ML-estimate' constitute the distribution of the ML-estimated theta's: we see, for example, that John's estimated theta will be +0.409 with a probability of 0.283. So we can compute the average ML-theta estimate, or, what amounts to the same thing but is more common to say, his **expected** theta-estimate. This expected value equals

$$(-5)\times 0.001+(-1.389)\times 0.019+\cdots+5\times 0.209=1.62,$$

**which is quite far from the real theta value of 1**. The difference between the expected estimate and the true value of theta is called the **bias**[2]. In this example, the bias is rather serious. Later on we will see in a more realistic example, that, in general, the bias of the ML-estimator remains serious.

In 1989, Th. Warm developed an alternative estimator, which, for reasonably long tests, is as accurate as the ML-estimator, but which is less biased. Commonly, this estimator is referred to as the Warm-estimator or as the weighted maximum likelihood estimator[3]. It has moreover the attractive property that it is defined for zero and perfect scores as well. The Warm estimates for the small example are displayed in the rightmost column of Table G.4. The expected value of the Warm-estimates is 0.96, which, compared to the true value of 1, results in a small negative bias.

We now consider a more realistic example with a 20-item test, complying with the Rasch model. The item parameters range from –1.05 through 1.7 with an average value of +0.5. In Figure G.14 the bias for the ML-estimator and the Warm-estimator are displayed. We comment on this figure:

1. The bias has been computed for 101 values of theta, put at equal distances from –3 to +3. The symbols for the same estimator form a reasonably smooth graph of a function, which is the bias function: the bias changes with the value of theta.
2. The graph running from the upper left, and staying stable at the zero line over a broad range and then decreasing further (dark blue diamonds) is the bias function for the Warm estimator. It is clearly seen that the bias is very near zero in the interval ranging from –1.5 to +2.5, and that even in a broader interval the bias is rather small: at +3 the bias is –0.022.
3. The interval where the bias is very small is not symmetric around zero. We will come back to that point later on.
4. The two other curves are the bias function for the ML-estimator. Since the ML-estimate does not exist for zero and perfect scores, we have a problem here. If we want to compute expected values (i.e., averages), we must have numbers, so that in the case of zero and perfect scores we have to fill in some number, which should be reasonable in some respect, but will always be arbitrary to some extent. This arbitrariness will influence the result, and the figure is constructed in such a way that we can see the consequences of this arbitrary decision.

[2] The bias found here is influenced by the arbitrary estimates plugged in for zero and perfect scores. This problem will be addressed in the sequel.

[3] The Warm estimate is defined (in the Rasch model and the two-parameter logistic model) as that value of theta for which a product of two functions is maximal. One function is the likelihood function, the other is the square root of the information function. The latter is considered as a weight for the former, hence the name 'weighted likelihood'.

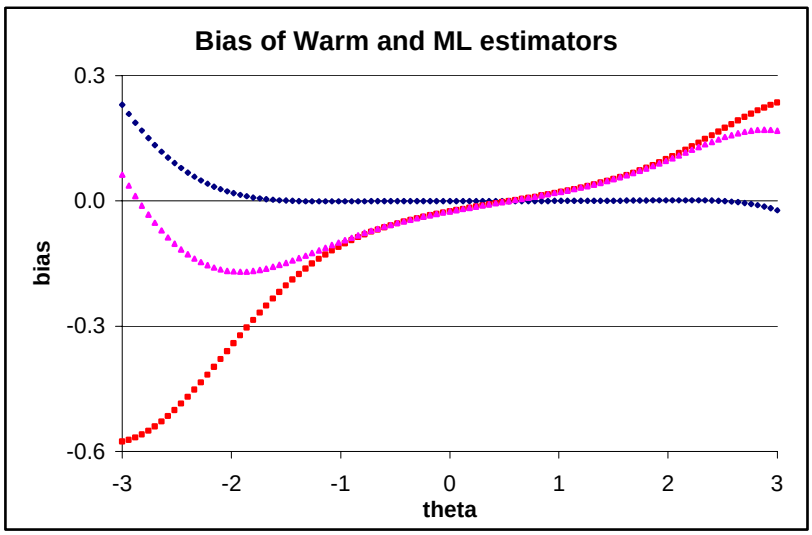


Figure G.14 Bias of theta estimators

5. The graph running from the lower left-hand corner of the display and ending in the upper right-hand corner (with red squares) is the bias function using –5 and +5 as estimates for zero and perfect scores respectively.
6. The third graph (with purple triangles) is the bias function of the ML-estimator, where the Warm-estimates for the zero and perfect scores have been used. These values are –3.56 and 4.50 respectively. We see that both bias functions coincide a great deal, roughly for theta values in the interval (-1,+2), while they differ outside this interval. This is caused by the fact that inside this interval, the probability of obtaining a zero or perfect score is so small that the precise value of their two theta estimates scarcely has any influence. For theta values to the left of the interval, the probability of a zero score is more substantial, and this probability is multiplied by –5 for the red curve and by –3.56 for the purple curve. That is why they go apart, as theta gets smaller: the smaller theta, the larger the probability of obtaining a zero score. A similar reason holds for values to the right of the interval.
7. The three curves cross at the same point, and at this point they have zero bias. In the example, this point corresponds to a theta value of about +0.5, and this corresponds with the **theta value where the test has its maximal information**. For the blue (Warm) and the red (ML, with plugged-in values of -5 and +5) curves in Figure G.14, the relation between information and bias is displayed graphically in Figure G.15. For the ML-estimator, we see that the bias is only zero if the information is maximal (which is about 4.4 in this example), and that when we move to the left along the x-axis, the bias increases in absolute value. For the Warm estimator, the bias remains very close to zero, even for information values lower than 2.
8. It appears in Figure G.15 that the red line (which has the appearance of a bird's beak) is symmetric around the horizontal zero line, but it is not completely so. This means that there is a close relation between bias and information, but one cannot be predicted exactly from the other. The precise relation is not known and this is a pity, because it restricts the generality of the conclusions we will draw from this small study.
9. Another interesting aspect in relation to the Warm estimator is the following observation: it appears from Figure G.15 that this estimator shows noticeable bias if the information drops under a value of two approximately. It would be interesting to know if this is also the case with other tests of a different length, with other item parameters, even with another model (like the two parameter logistic model with different item discriminations). If this were the case, we would have a quite valuable result, because from the information function we could then determine the range of theta values which will yield (approximately) unbiased Warm estimates.

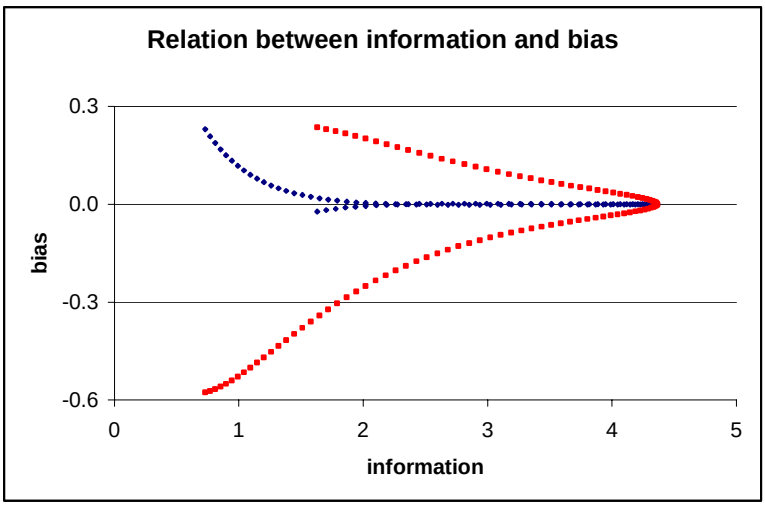



Figure G.15. Bias and information

10. To shed some light on this problem, the bias function for the Warm estimator and the information function for a test of 40 items were constructed. The item parameters used are the same as in the 20 item test; but they occurred twice as often. The maximal information value in this 40-item test is therefore exactly the double of the maximum in the 20 item test (its value is about 8.8). In Figure G.16 the relation between the bias of the Warm estimator and information is displayed. (The blue diamonds refer to the 40-item test; the red squares to the 20-item test). Although the value where the bias tends to depart from zero is about 2 in both cases, it is also clear that the departure from zero holds for larger values in the long test than in the short one. But for practical purposes, a value of 2 seems to be fairly useful for practical applications. (Notice that in Figure G.16 the unit for the y-axis is different from the unit in Figure G.15).

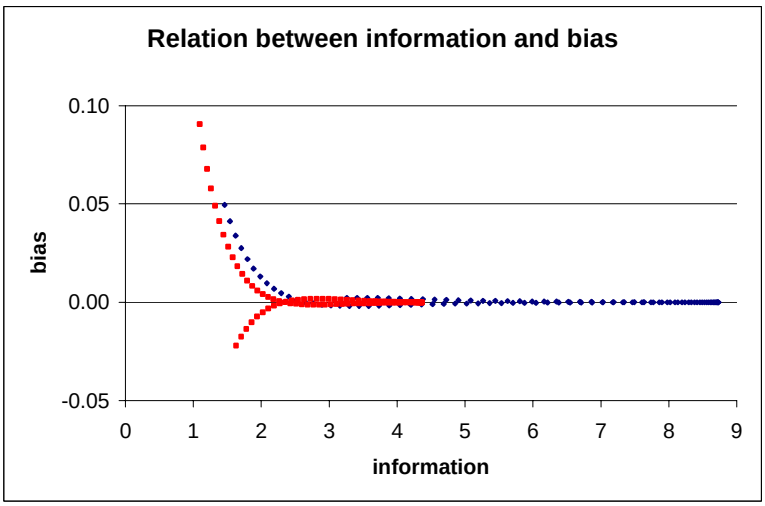



Figure G.16. Bias of the Warm estimator and information

Now we are ready to summarize the results on the estimation of theta:

1. Two estimators of theta can be used after calibration: the ML-estimator and the Warm estimator. Both have (approximately) the same standard error.
2. For both estimators it holds that the theta estimate depends only on the score of the test, not on the specific response pattern. This is true in the Rasch model and in the two parameter logistic model (2PLM). But it does not hold in the three parameter model.
3. The ML-estimate does not exist for zero and perfect scores but the Warm estimate does exist for all scores.

4. The ML-estimator is biased. For theta values larger than the point of maximal information, this bias is positive, meaning that on the average the estimate will be larger than the true value; for theta values smaller than the point of maximal information the bias is negative. If one takes these two effects jointly, this means that the ML-estimates will tend to have a larger variance than the real theta values.
5. The Warm estimator shows only a small (and negligible) bias in a large interval around the point of maximal information. Outside this interval it shows a bias which is in the opposite direction from the bias in the ML-estimator: for small values of theta the bias is positive, for large values it is negative. The effect of this bias is that the variance of the Warm estimates will tend to be smaller than the variance of the real thetas. This effect is known as shrinkage.
6. A small study suggests that with the Warm estimator, bias begins to be serious for theta values where the test information is smaller than 2. This result, however, is provisional and should be corroborated by more evidence. It is important to notice that this result was found for the Rasch model. It might be different for the 2PLM.

### G.7.3 EAP-estimates

The ML-estimator and the Warm-estimator are based exclusively on the test score, i.e., all the information that these two estimators use is provided by the test taker, and no other sources of information are used. There exist, however, also estimation procedures that use other information in a systematic way.

Suppose John will take a test. We know that he has followed a course of English for four years, and from other research, we happen to know that in the population of students who have studied four years of English, the mean theta value is 1.1 and the standard deviation is 0.7. We also happen to know that the distribution of theta in this population is approximately normal. Since John also belongs to this population, we could say that in some sense we have some information on John's ability. We are fairly sure, for example, that John's ability will not be larger than 2.5 on the theta scale (because 2.5 is two standard deviations above the mean), and if we should make a systematic guess, the population mean would be a good one. In fact, this guess is the best one we can make in many respects. But formally speaking, this guess is an estimate based on all the information we have about John before he takes the test. This information is called the **prior** information, and we take as the estimate the mean or expected value of the distribution of the theta values we happen to have information about.

After the test taking, we have collected more information about John, and suppose that he obtained a score of 18 on a 20-item test, a fairly good result. Then we could ask a very nice question: suppose that we happen to know the theta value of all the members of the population, and suppose further that we administer the test to everybody. So we have, for all population members, their theta value and a test score. Now we collect all people having obtained a test score of 18 (the same as John's), and we make a histogram of their theta values. What would that histogram look like? Notice that this question is different from a problem we studied in the section about bias: there we were looking for the distribution of test scores given the value of theta (see Table G.4 for an example); here we have the reverse problem: **what is the distribution of theta given the test score**. This distribution is called the **posterior** or **a posteriori** distribution (as opposed to the distribution we knew before the collection of test scores, which is called the **prior distribution**.)

Since John has obtained a score of 18, it seems wise to base our estimate of John's theta on the posterior distribution rather than on the prior distribution, because we then take into account the extra information John has delivered. And indeed, this is exactly what is done: the estimate of John's theta value is the mean or expected value of the posterior distribution. Hence the acronym EAP: **E**xpected **A** **P**osteriori. As an indication of the accuracy, one can take the standard deviation of the posterior distribution.

Here are some comments on this method:

1. In the Rasch model there is a different posterior distribution for each score. Once the score is given, the posterior distribution of theta does not depend on the specific response pattern. For example, in a four-item test the posterior distribution given the response pattern (0,0,1,1) is the same as that given the response pattern (1,1,0,0), because the two response patterns have the same score. In the two parameter logistic model there is a different posterior distribution for each value of the weighted score.
2. The imaginary situation described above (knowing everybody's theta value etc.) only served a didactic purpose, and cannot be realized. But if the prior distribution is known (e.g. we know that it is normal with a given mean and SD), and if the item parameters are known, then the exact form of the posterior distribution for each possible score can be computed. In Section G.8, it will be shown how the two distributions in Figure G.17 (see below) can be constructed with the program EXCEL.
3. If the prior distribution is normal (as it usually is in most applications), then the posterior distributions are not normal. For extreme scores the posterior distribution may be skewed. In Figure G.17 an example is given. The left-hand distribution is a normal prior with a mean of 1.1 and a SD of 0.7. The test consists of 15 items, all having the same difficulty of +1. The right-hand distribution is the posterior distribution for a score of 14. The right-hand tail is a bit more stretched than the left. The expected value of this distribution is 2.28 and its standard deviation is 0.47, a value markedly smaller than the prior standard deviation of 0.7. So in general, the posterior distribution, as graphed in the figure, reflects precisely what we can learn from such a score: the whole graph of the posterior is situated quite far to the right of the prior distribution, implying that people getting a score as high as 14 on this test in general have a quite high theta value But at the same time we have still a substantial SD in the posterior, so all we can say about John is that he belongs to this posterior population, but we cannot locate him more precisely with the information we got from him. (One should not draw conclusions from the fact that the posterior distribution's graph has a higher 'top' than the prior: both figures are scaled in such a way that the total surface under the graph is equal for both figures.)

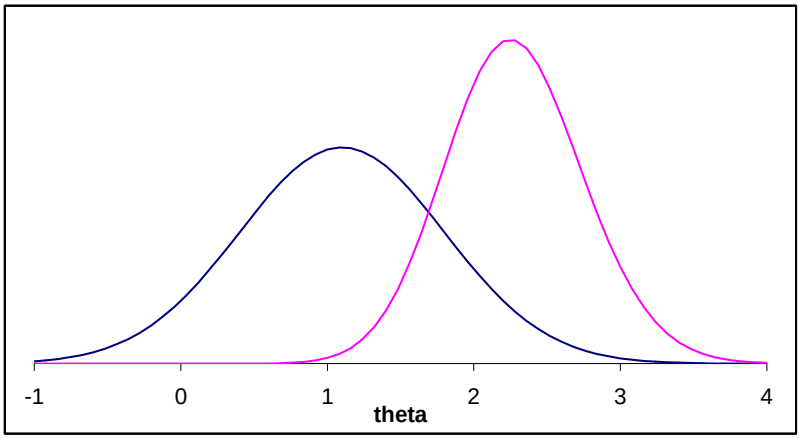



Figure G.17. Prior and posterior distributions

It may seem that the use of the EAP estimator is very attractive, since it uses all the available information one has. But one should be careful with such an approach, especially when decisions about individual persons are based on their estimated theta-value. The form and the location of the posterior distribution depend to some extent on the prior distribution, such that the mean of the posterior can be seen as a kind of compromise between the prior information we have (John comes from a population with a mean theta of 1.1) and the information we have from an individual test performance (John got a score of 14 out of 15 items). Now suppose the prior information that we had related only to male students having received four years of instruction in English, but that we also have prior information for the female population, and suppose further that in the female population the mean is 1.6 with an SD of 0.7. Mary belongs to this population and she happens to obtain also a score of 14 items correct, the same as John's. But for Mary the EAP-estimate will be higher than for John,

because it is a compromise between a larger prior mean and the same test score. Upon computation we find that Mary's EAP-estimate is 2.51, while John got 2.28 for the **very same test performance**. So in some way, John is punished for being male, and in situations where decisions are based on a test score, this may be conceived as unfair.

## G.8 Producing graphs with EXCEL

In the present section a step by step instruction will be given how to compute the function values for a number of interesting functions in an IRT-framework. It will be seen that the amount of formula typing and entering values is really modest while the result –an illuminating graph- is sometimes worth a thousand words.

The Section is arranged in four subsections:

1. In Section G.8.1 some general principles of handling a spreadsheet in EXCEL will be explained by constructing, step by step, the formulae and procedures to plot a number of item response curves
2. In Section G.8.2 the information function of a test will be built;
3. In Section G.8.3 a graphical method for the ML and the Warm-estimator will be developed
4. In Section G.8.4 posterior distributions of the theta values will be constructed.

Graphs G.14, G.15 and G.16 related to the previous section (on bias) are also produced with EXCEL, but the computation of the values is quite complicated, and has to be done with special software.

The whole section should be read and studied cumulatively: in later sections concepts and techniques explained in earlier sections will be used without further exposition. At the same time the results will be a bit more general than in sections G5 through G7, because we will use the two parameter logistic model instead of the Rasch model.

The section is not a beginner's introduction to EXCEL. If the concepts and techniques which are introduced here are not understood, it may be wise to consult an introductory tutorial in EXCEL. Sometimes, built-in functions from EXCEL will be used (like SUM). The name or acronym for these functions stems from an English version of EXCEL. If the language of the program is not English, these names may be different. Some functions, however, are so universally used, that they only have a single name across languages. An example is the function EXP.

### G.8.1. General principles of EXCEL

When EXCEL is opened from scratch, a sheet, containing cells is displayed on the screen. For our purposes, it is enough to work on a single sheet. The cells of the sheet (displayed as rectangles) are referred to by an **address**, which consists of a **column** letter (or pair of letters, to be understood as a single symbol), and a **row** number. These letters and numbers are displayed automatically by EXCEL. (See Figure G.18).

When we do computations for IRT we will need theta values and the values of the parameters. In what follows, the theta values will be stored in column A, starting at row 3, the discrimination parameters will be stored in row 1, and the difficulty parameters in row 2, both starting at column B.

In IRT, theta is a continuous variable which can assume any number. But one cannot type all numbers, so we will have to make a selection. Let us assume that we are only interested in theta values in the interval (-3,+3), and in this interval we will only use about 100 different theta values at equal distances from their neighbours. Since 3 – (-3) = 6, each value from the second on will be 6/100 = 0.06 units larger than its predecessor. The nice thing about EXCEL is that we only have to type two different numbers, and the other numbers can be generated by a simple technique of selecting and dragging. The whole process is exemplified in Figure G.18

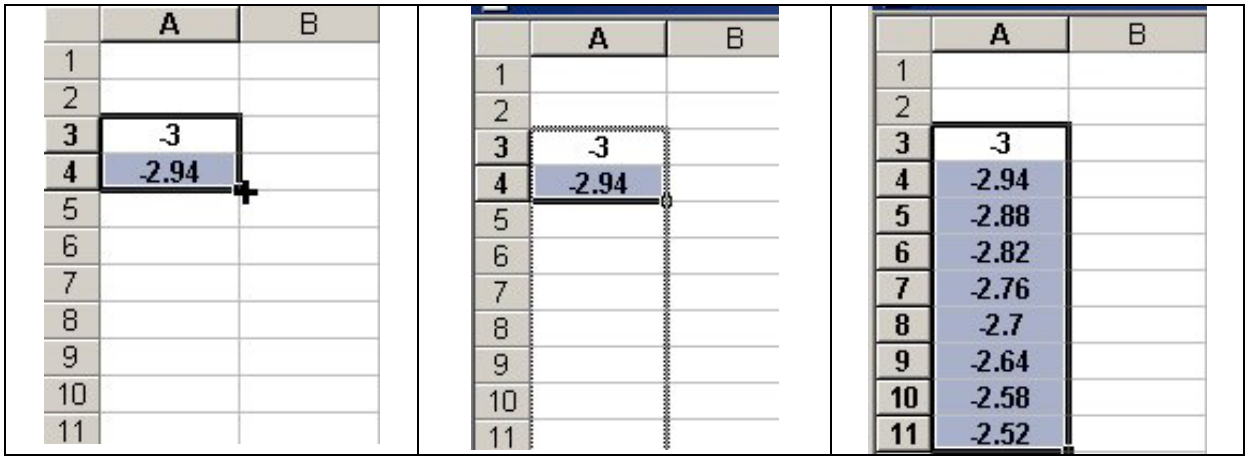


Figure G.18. Creating an equidistant series

In the left hand panel, the situation is depicted after having typed two values. The two values are selected jointly, and the cursor is placed at the lower right-hand corner of the black rectangle (at the place of the small black square). Put the cursor in such a way that a black +-sign appears, not a hollow one. Drag this +-sign downwards holding the left-hand button of the mouse down (see middle panel), and upon releasing the mouse button the equidistant values are filled in the black rectangle (which is selected as a whole; see right-hand panel). Clicking in any cell of the spreadsheet will undo the selection. If the mouse is dragged until row 103, we will have 101 equidistant theta values in the range (-3,+3).

It is good practice to distinguish between values that are typed (or dragged as in the example) and values which are the result of a formula application. This can be done by very simple lay-out functions. In the example (left panel) the two numbers are centered in their cells and made bold. This lay-out is automatically inherited by the cells defined by dragging. Dragging can also be applied starting from a selection of a single cell. In that case, the value of the cell is repeated in all cells attained.

In the left-hand panel of Figure G.19 the discrimination parameters for four items (row 1) and the difficulty parameters (row 2) are filled in, and the cursor is placed in cell B3, ready to accept a value or a formula. Notice that in top of the spreadsheet, the active cell is identified (B3) and that to the right of this, there is an empty box, preceded by the '='-sign. To type a formula one can just type with the cursor in cell B3, or one can place the cursor in the formula box. To edit an existing formula, however, one must place the cursor in the formula box.

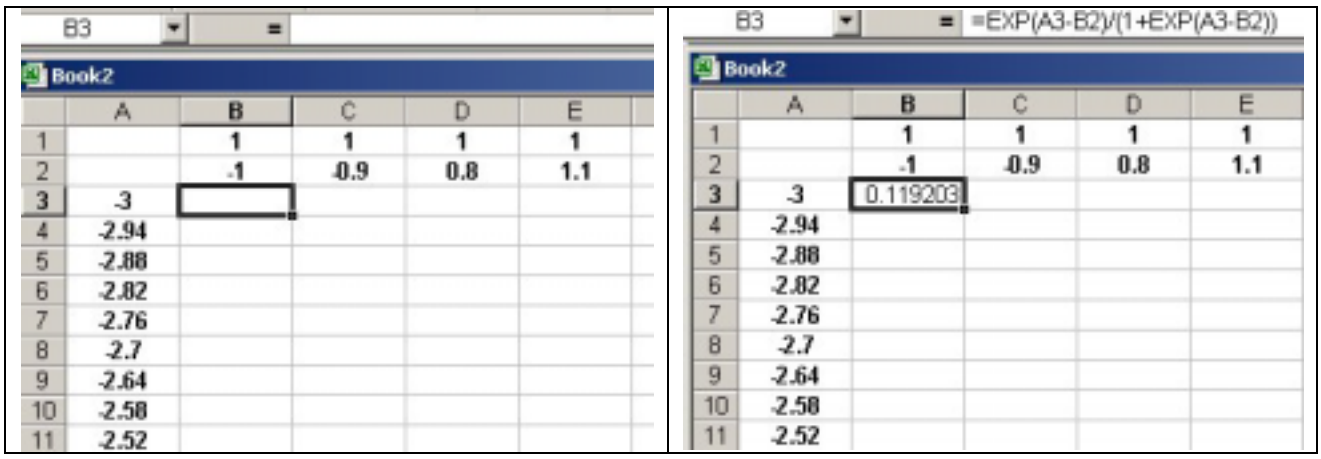


Figure G.19. Specifying a formula

To specify a formula one can almost use literally the mathematical formula as given in textbooks. The only difference is that for the variable (theta) we must specify the cell where the value of theta can be found and for the value of the parameter we must type in a specific numeric value or refer to a cell where that value can be found. For cell B3, it is natural to choose the theta value in cell A3 and the difficulty parameter from cell B2. So if we want to use formula (G.3) from section G.5 we could type:

=exp(a3-b2)/(1+exp(a3-b2))

and after typing the 'enter' key, the formula is evaluated, the cursor makes another cell active, but if we come back to cell B3 (by clicking on it), we see the spreadsheet as displayed in the right-hand panel of Figure G.19. Notice that:

- Typing a formula must begin with the '='-sign. If '=' is omitted, the formula itself will be displayed in the cell.
- The use of uppercase or lowercase symbols is arbitrary. EXCEL turns all used letters to uppercase.
- The function 'exp' is a built-in function in EXCEL.
- Addition and subtraction are symbolized by '+' and '-' respectively; multiplication and division by '*' and '/'. The multiplication must be mentioned explicitly: for example, 3*A2 (multiply the value in cell A2 by 3). Typing '3A2' is not understood by EXCEL and will lead to an error.

**Absolute and relative addresses**

A great advantage of EXCEL is that not only values can be copied from one cell to another but formulae as well. To understand properly what happens, we need to know what an address is. Suppose we make cell B3 active, i.e., we select it, and we type the formula

=2*a3

then the formula does not mean to multiply the number 2 by the number a3, which is not possible, since a3 is not a number. What is meant is to perform the multiplication of the number 2 with the number that can be found in cell 'a3'. The cell identification is called the address.

But addresses can be read in two different ways: absolutely and relatively. Since the active cell is B3, the address A3 can be read as

1. the preceding column, same row (relative to the current position B3)
2. the address in column A, row 3, whatever the current position: this is absolute addressing.

If we use the relative address A5 while being in cell B3, then A5 is to be understood as the cell in the preceding column, two rows below the current one.

EXCEL allows for both modes, relative and absolute, for the row and column indication separately, leading to four modes of addressing. Absolute addressing needs the '$'-sign; relative addressing is the default (no special sign involved). Now, still being in cell B3, we can write the above formula in four different ways:

1. row and column relative to the current position: =2*a3
2. row relative and column absolute: =2*$a3
3. row absolute and column relative: =2*a$3
4. row and column absolute: =2*$a$3

For each way of writing the formula we will get the same result. But things will change if we copy this formula to the clipboard, and then paste it in some other cell, C5, say. For the four cases listed above, we will find in the formula box the following formulae when C5 is made active:

1. =2*B5 (same row, preceding column);
2. =2*$A5 (same row, but column A, absolutely);
3. =2*B$3 (third row, absolutely, preceding column);
4. =2*$A$3 (third row and column A, both absolutely).

If we want the probability for a correct response to four items and for 101 different values of theta, it would be silly to type the formula 404 times. Using a clever mixture of relative and absolute addressing we only need to type the formula once. Here it is for cell B3 (and we generalize

immediately to the two parameter logistic model; compare to the mathematical formula (G.4) in Section G.5):

=exp(b$1*($a3-b$2))/(1+exp(b$1*($a3-b$2)))

Here are some comments:

- The reference to the discrimination parameter is b$1: the column address is relative (same column), because we need the discrimination parameter of the current item. If the formula is copied to column C, we will need the discrimination parameter of the next item; hence the column address is relative. But the row address is absolute: the discrimination parameter is in the first row, whichever row we are in. Relative addressing would mean 'two rows above the current one'. A similar reasoning applies to the difficulty parameter.
- The reference to the theta value is $a3. The column address is always column A, not just the preceding column. The row address, however, is relative: we want the current theta value. If the formula is copied to cell B4, we want to use the theta value in A4, not the one in A3.
- To copy the formula to all 404 cells (101 theta values and four items), we apply the same technique as for creating a series of values:
    - type the formula in cell B3, make cell B3 active, and put the cursor at the right-hand lower corner such that the black '+' appears.
    - Drag the black '+' horizontally to cell E3. Upon releasing the mouse button, the formula is copied in cells B3, C3, D3 and E3, and these four cells are selected, i.e., enclosed in a black rectangle.
    - Put the cursor at the right-hand lower corner of the rectangle such that the black '+' appears, and drag is downwards to cell E103. Upon releasing the mouse button, the formula is copied to all 404 cells, and the computations are done.

In Figure G.20 the situation is depicted after this copying, while cell D5 is the active cell. Notice the formula in the formula box.

D5 = =EXP(D$1*($A5-D$2))/(1+EXP(D$1*($A5-D$2)))

Book2

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | **1** | **1** | **1** | **1** | | |
| 2 | | **-1** | **-0.9** | **0.8** | **1.1** | | |
| 3 | **-3** | 0.119203 | 0.109097 | 0.021881 | 0.016302 | | |
| 4 | **-2.94** | 0.125648 | 0.115067 | 0.023203 | 0.017293 | | |
| 5 | **-2.88** | 0.132389 | 0.121319 | 0.024602 | 0.018343 | | |
| 6 | **-2.82** | 0.139434 | 0.127862 | 0.026084 | 0.019455 | | |
| 7 | **-2.76** | 0.14679 | 0.134703 | 0.027652 | 0.020633 | | |
| 8 | **-2.7** | 0.154465 | 0.141851 | 0.029312 | 0.021881 | | |
| 9 | **-2.64** | 0.162465 | 0.149313 | 0.031068 | 0.023203 | | |
| 10 | **-2.58** | 0.170795 | 0.157095 | 0.032926 | 0.024602 | | |
| 11 | **-2.52** | 0.179462 | 0.165205 | 0.034891 | 0.026084 | | |

Figure G.20. Copying formulae

**The power of a spreadsheet**

Once we have the probabilities of a correct answer for a few items, we can easily extend these formulae to new items. If we want a fifth item (in column F, say), we simply copy one of the other columns into column F, and the formulae of all the cells in this new column are automatically adapted.

If one wants other item parameters for this new item, all one has to do is to change the values for these parameters in cells F1 and F2. As soon as a change is made in some cell, say F1 (and this cell is left by

making another cell active), all formulae where reference is made to F1 are computed again and the result is displayed. If a graphical display is constructed, using the values in column F, the graph will be automatically adapted as well.

**Drawing a graph**

Here is some information on how to draw a graph quickly in EXCEL. We will draw a graph of the item response functions in columns B to E of the preceding example. In drawing a graph we need to provide the coordinates for a number of points. These points are then plotted in a plane and (optionally) connected by a line. It is also possible to plot only the connecting lines, without a special symbol for the points themselves. We will choose that latter option.

- Choose the button for the 'Chart Wizard' from the toolbar. It looks like this:

(If it is not visible, activate the standard toolbar: in the menu View, choose 'Toolbars', and click on 'Standard')

- The first step of the Wizard is displayed as in Figure G.21. Make the selection 'XY (Scatter)' from the list of Chart types and select the sub-type as indicated in the figure. Then, press the 'next' button. (It is also possible to work with 'Line' as chart type, but in our experience, it is easier to work with the scatter chart.)

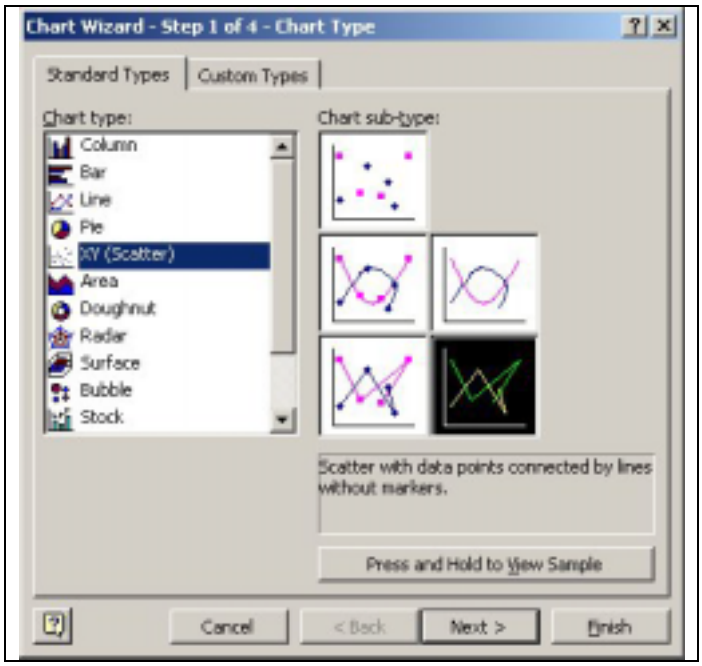



Figure G.21. Chart Wizard, step 1

- In the second step of the wizard, choose the tab 'Series' (see Figure G.22). It may happen that some graphs are defined already (it will not happen if the wizard is started while an empty cell is selected). To start from scratch, existing graphs can be removed with the 'Remove' button.

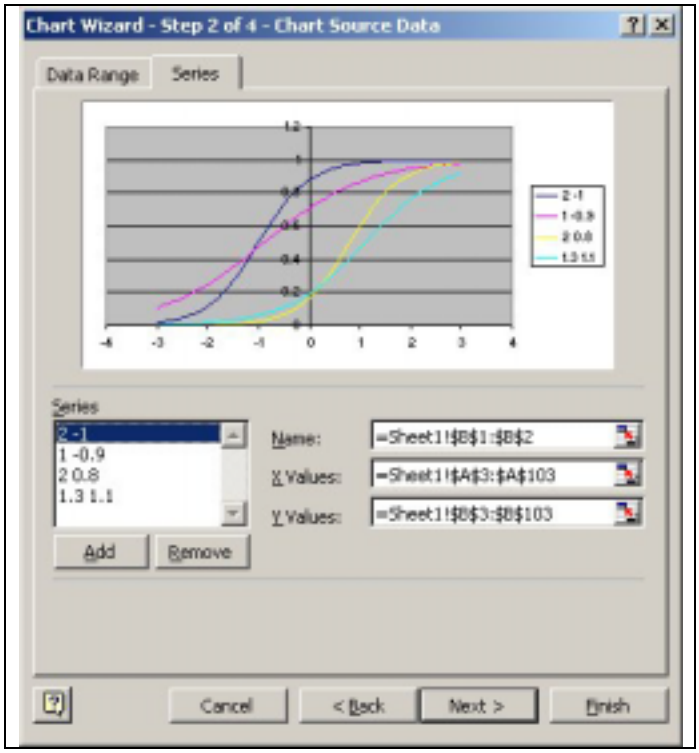


Figure G.22 Chart wizard, step 2

- To add a graph, use the ‘Add’ button. Upon pressing ‘Add’ the three boxes at the right become empty, and can be filled. The ‘Name’ box can be filled with the name of the graph (or with a reference of the cell(s) where the name is to be found). This name will appear in the legend accompanying the resulting graph. The other two boxes are used to specify the cells where the x- and y-coordinates are to be found. One can type these references as shown in the example in Figure G.22, but one can also use the button (red, blue and white at the right end of the box). This button is called the ‘Collapse Dialog’ button, and upon pressing it, the following happens:
    - The dialog as displayed in Figure G.22 disappears (provisionally);
    - The value box alone appears on the screen;
    - The values needed can be selected using the mouse, from the active sheet but also from another sheet. (The selected values are surrounded by a dashed rectangle.)
    - Upon pressing the ‘Collapse Dialog’ button in the box again, the dialog reappears and the selected cells are filled in the correct format in the value box.
- Choosing ‘Next’ brings the user to the third step where a number of choices can be made concerning the lay-out. These choices are self-evident. The last step (choosing ‘Next’ again) leaves the choice for the location of the graph: in the active sheet or on another sheet. Pressing the ‘Finish’ button brings one back to the EXCEL sheet with the constructed figure displayed on it. The ‘Finish’ button may be pressed after each step. In the example to be discussed next, the ‘Finish’ button was used after the second step.
- A figure thus constructed may be edited in all respects at all times. A figure consists of a number of objects which may be edited separately. These objects are: the chart area (indicated by a selection of the outer frame of the figure), the plot area (the rectangular area formed by x- and y-axes), the legend, the x-axis, the y-axis, each graph and each title. To edit an object in the figure, select it, click the right mouse button, after which a menu appears, and make a choice from that menu. In the left-hand panel of Figure G.23 the figure with the four item response curves is displayed using the default options for lay-out from EXCEL. The right-hand panel is the lay-out that is used mostly in the figures of the present section. We comment on how to proceed to get this lay-out.
    - *Remove the legend*: select the legend, click the right mouse button, choose ‘Clear’.

- *Remove the gray background*: select the plot area, click the right mouse button, choose 'Clear'. (To create another background: choose the option 'Format Plot Area', and choose whatever you like.)
- *Add titles*: select the chart area, click the right mouse button, choose the option 'Chart Options...', and go to the tab 'Titles'. Titles are written in a default font with a default size. To change these, select the title in the figure (and not while being in a title box of a dialog), click the right mouse button and select the option 'Format Title'. After adding or editing titles, it may happen that the plot area has become rather flat. To change its area, select it, put the cursor on one of the black squares (it changes into a single or double arrow) and drag the plot area to display the form and area you wish. (Notice that the text of a title cannot be edited after selecting the title itself; one should select the chart area, and choose the 'Chart Options...'.)
- *One of the curves has to be removed*: select it, click the right mouse button, choose 'Clear'.
- *Change the color of a curve*: select it, click the right mouse button, choose 'Format Data Series' and a dialog is opened. Select the tab 'Patterns' and change the 'Color' of the 'Line'.
- *The x-axis should be restricted to the interval (-3,+3), and, moreover, the y-axis should cross the x-axis at –3 and not at zero as in the left-hand panel of Figure G.23*. Select the x-axis, click the right mouse button, choose 'Format Axis...'. A dialog appears; choose the tab 'Scale' and specify the boxes 'Minimum:' (-3), 'Maximum:' (3) and 'Value (Y) axis crosses at:' (-3). Notice that once these options are used, they remain in effect until changed actively.
- *The y-axis should be restricted to the interval (0,1), we want numbers and gridlines displayed at a distance of 0.25, and not of 0.2 as in the default lay-out and, finally, all displayed numbers should have the same number (2) of decimals*. To restrict the maximum value, proceed as with the x-axis. To control the distance between gridlines and the displayed axis values, specify 0.25 in the box 'Major Unit:' of the same dialog. To control the number of decimals, select the tab 'Number' in the dialog, select 'Number' in the box 'Category:', and then select the wanted number of decimals in the box 'Decimal places:'.
- *To add a new graph to the figure*, select the plot area or the chart area, click the right mouse button and choose 'Source Data...', whereupon the dialog as displayed in Figure G.22 will appear. A new graph can be added.

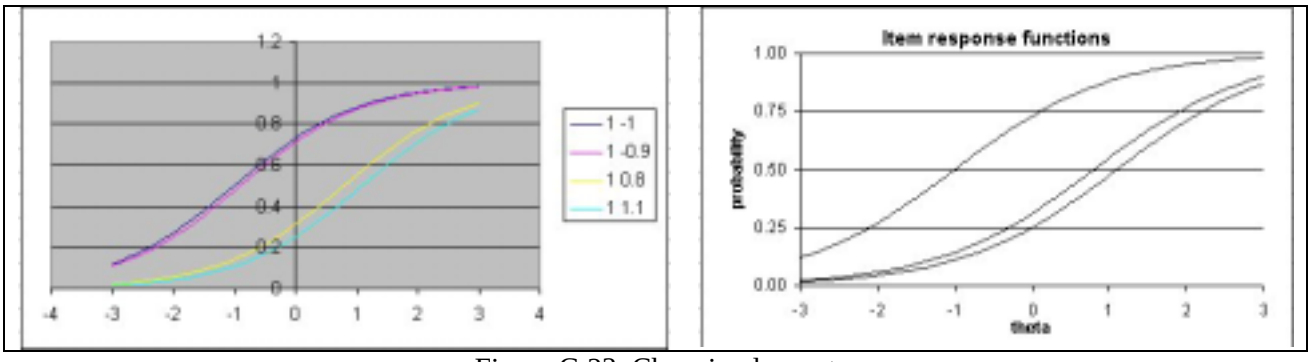


Figure G.23. Changing lay-out

### G.8.2 Computing the information function

The formula for the information function (given as formula (g.6)) is repeated here for convenience:

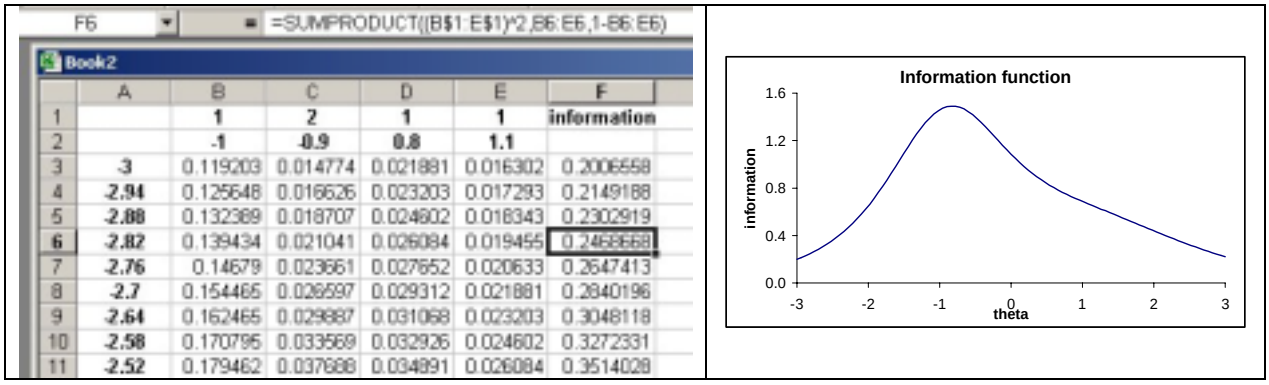
$$I_t(\theta) = \sum_i a_i^2 f_i(\theta)[1 - f_i(\theta)]$$

The formula is a **sum** across items and each term of the sum consists of a **product** of three quantities: the square of the discrimination parameter, the value of the item response function for some value of theta and one minus the value of the item response function for the same value of theta. So, for a

specified value of theta, the information function is a **sum** of **products**, and we can compute it directly in EXCEL by the very powerful built-in function SUMPRODUCT. We first give the formula and then comment on it. Refer to Figure G.20, and assume that cell F3 is active. The formula to be typed is:

=SUMPRODUCT(B$1:E$1^2,B3:E3,1-B3:E3)

- The function SUMPRODUCT has three arguments, placed between parentheses and separated by commas (in some languages the semi-colon has to be used to separate arguments). The second argument, for example, is written as B3:E3, and denotes the array of cells starting at B3 and ending at E3. Notice that the addresses are relative to the current active cell F3: the row indication '3' should be read as 'current row', and the column indication 'E', as the preceding column. (The function SUMPRODUCT can have as many as 30 arguments.)
- The third argument is '1-B3:E3'. It means that the values of the array B3:E3 must be subtracted from one, cell by cell, before they can be used. So we refer to an array which was not defined explicitly in the spreadsheet, but which will be created implicitly by the function SUMPRODUCT.
- The first argument is B$1:E$1^2. The caret ('^') denotes exponentiation, and since the exponent is 2, we want squares of all the values in the array B$1:E$1. Notice that we use absolute addressing for the rows, because the discrimination parameters are listed in row 1 and not in general two rows above the current row (true for cell F3, but not for F4).
- The result in F3 is the value of the information for the theta value stored in cell A3. The formula can be copied by dragging it downwards until cell F103, and the column F can be used to plot the information function. In Figure G.24 (left panel), part of the spreadsheet is displayed after these computations, but notice that the discrimination parameter of item two (cell C1) has been changed from one to two. In the formula box, the array indication B$1:E$1 is put between parentheses; this is allowed but not compulsory. In the right-hand panel, the information function is displayed graphically to show that it is not always nicely symmetric.




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | 1 | 2 | 1 | 1 | information |
| 2 | | -1 | -0.9 | 0.8 | 1.1 | |
| 3 | -3 | 0.119203 | 0.014774 | 0.021881 | 0.016302 | 0.2006558 |
| 4 | -2.94 | 0.125648 | 0.016626 | 0.023203 | 0.017293 | 0.2149188 |
| 5 | -2.88 | 0.132389 | 0.018707 | 0.024602 | 0.018343 | 0.2302919 |
| 6 | -2.82 | 0.139434 | 0.021041 | 0.026084 | 0.019455 | 0.2468668 |
| 7 | -2.76 | 0.14679 | 0.023661 | 0.027652 | 0.020633 | 0.2647413 |
| 8 | -2.7 | 0.154465 | 0.026597 | 0.029312 | 0.021881 | 0.2840196 |
| 9 | -2.64 | 0.162465 | 0.029887 | 0.031068 | 0.023203 | 0.3048118 |
| 10 | -2.58 | 0.170795 | 0.033569 | 0.032926 | 0.024602 | 0.3272331 |
| 11 | -2.52 | 0.179462 | 0.037688 | 0.034891 | 0.026084 | 0.3514028 |

Figure G.24. Information function

If the function SUMPRODUCT happens to have another name in another language, it can be found as follows: click on the button $f_x$ in the standard toolbar of EXCEL, and search in the function category 'Math & Trig' (mathematics and trigonometry). Placing the cursor at any of the displayed function names will give explanations on the chosen function. Double clicking on the selected function name will start a wizard which can be helpful in writing the correct format, although some extra editing may be necessary. Make sure to select the correct cell (where the formula has to apply) **before** starting the wizard.

### G.8.3 ML- and Warm-estimates

Usually software for IRT produces ML- or Warm-estimates for all possible test scores. Nonetheless, it may be instructive to produce some graphs of the likelihood function (for ML) or the weighted likelihood (Warm). Once the item response function has been evaluated (in columns from B through E) and the information function (column F) is computed, the required computations for the likelihood

and the weighted likelihood are simple. But we should keep in mind that the likelihood function (in general) is different for each response pattern: even if the score of two response patterns is the same, the likelihood function in general will be different. (See Figure G.11 for an example.)

We will use column G for the likelihood function of the response pattern (1,1,0,0), and column H for the weighted likelihood function. The formula to be typed in cell G3 is then

=B3*C3*(1-D3)*(1-E3)

and this formula can be copied in all relevant cells by dragging. Once this is done, the formula for the weighted likelihood is even simpler: it is the product of the likelihood and the square root of the information function. So, making the cell H3 active, we only type

=G3*SQRT(F3)

Plotting both functions in the same graph usually will not result in an elegant picture, because the units of both functions may be quite different. Even plotting two likelihood functions in the same graph may not be satisfying because of the (sometimes grossly) different scales. But since the (weighted) likelihood function will be mostly needed to find the theta value where it reaches its maximum, one can rescale one or both of the functions such that they can nicely be displayed together in the same graph. This can be done as follows:

- After having applied the two formulae above, we look up columns G and H to find the largest value. In column G the largest value happens to be 0.3247, and in column H 0.3506. We can also use the function MAX to find the maximum. Choose some empty cell and enter the formula =MAX(G3:G103)
- Next we recompute columns G and H, but we divide the former function values by their maximum values. So in cell G3 we specify the formula

  =B3*C3*(1-D3)*(1-E3)/0.3247

  and in cell H3 we specify

  =G3*SQRT(F3)*0.3247/0.3506

  (Notice that in the latter formula we have to multiply first by 0.3247 because we use a **new** G3 value which is the old one divided by 0.3247.)
- The new formulae are copied to the whole of columns G and H.
- Now the maximal value in both columns will be equal to one. Notice that in columns G and H we now do not find any longer the (weighted) likelihood, but the (weighted) likelihood multiplied by some constant (different for the two columns). But the important thing to understand is that by multiplying the function values by a constant, the **form** of the graph will not change, and in particular, the theta value at which the functions reach their maximum will not change. The standard way of expressing this is to say that the values in column G are now **proportional** to the likelihood. In Figure G.25 both proportional functions are displayed, and we see that the maximum likelihood estimate is larger than the Warm estimate. The y-axis has been deleted because the values to be displayed have a different meaning for the two curves.

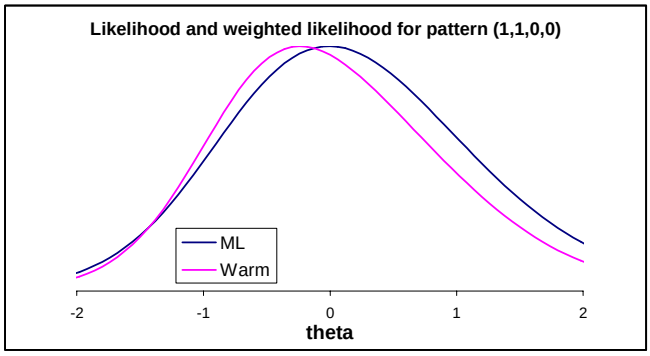



Figure G.25 Likelihood and weighted likelihood functions (proportional)

### G.8.4 Posterior distributions

Before we start with technical explanations, something has to be said about the graph of a distribution of a continuous variable. As an example we will take the prior distribution of the example used in Section G.7: it is a normal distribution with a mean of 1.1 and a standard deviation of 0.7. The graph of the distribution we are acquainted with is a bell shaped curve. The x-axis represents the values the variable can assume (in our case: theta). In the normal distribution these values run from minus infinity to plus infinity, but in drawing a graph we usually restrict the range of values to about three standard deviations at either side of the mean. To plot the curve, we need to know also the y-coordinate at each point (the y-value), and here there arise two questions: how does one compute these y-values and what do they mean? To compute the y-values for a given value of theta, we need a rule, the function rule of the normal distribution. Here it is:

$$y(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left[ -\frac{(\theta-\mu)^2}{2\sigma^2} \right] \qquad \text{(G.10)}$$

- $y(\theta)$ is the value of the function for a given value of theta.
- $\sigma$ is the value of standard deviation (in our case 0.7) and $\mu$ is the value of the mean (in our case 1.1). The symbol $\pi$ represents the number 3.14159..., well known from trigonometry.
- We see that in the right-hand side of (G.10) the symbol theta also appears. If we substitute a number for this symbol, we can compute the value of the y-coordinate at that number, and for different numbers used we will get different results (in general). So, formula (G.10) is a function rule. If we compute it for a number of theta values and make a plot, we will get that famous bell shaped curve. But we can make the computations a bit simpler.
- The right-hand side of formula (G.10) contains two factors (indicated explicitly by the multiplication sign); the first factor does not contain theta, the second one does. So one might ask why this first factor is there. The reason is that in a probability distribution the total area under the curve must be equal to one, and we need the first factor to make sure that this will be the case. Therefore this first factor is called a **normalizing constant**. (It is constant because it does not depend on the variable theta.)
- But what do we mean by an area of one? one what? If we make a plot of the function on paper, we could measure the area under the curve and find that the area is 1.3 square inches. But if we make a reduced photo copy of the plot, we might find that the area on the copy is now 0.8 square inch, but nobody will think that the figures on the original and the copied plot represent something different. So for plotting purposes we do not need this normalizing constant, and we may replace the rule (G.10) by a simpler rule:

  $$y(\theta) \text{ is proportional to } \exp\left[ -\frac{(\theta-\mu)^2}{2\sigma^2} \right] \qquad \text{(g.11)}$$

  and this is all we need to compute in the spreadsheet. Continuing the example of the preceding section, we will define a formula in cell I3 and then copy it to the whole column I (by dragging). The formula is

  =exp((a3-1.1)^2/(-2*0.7^2))

  where the numerical values of 1.1 for the mean and 0.7 for the standard deviation are used.
- In Figure G.26 the distribution is plotted in three different ways. In all three panels the interval used for the theta values and the length of the x-axes are exactly the same; yet, the three plots look quite different. The reason is that the y-axis is scaled differently in the three cases. There is no mathematical reason why one should prefer any one of the three graphs. Usually, the middle one will be preferred, but this is only for aesthetic reasons (usually, the ratio of the length of the y-axis to the length of the x-axis is about 3:4). It is useful to realize this when constructing or judging plots. The plot in the left-hand panel might suggest a distribution with a large standard deviation and the one in the right-hand panel a small standard deviation, but all three plots represent the same distribution; only the lay-out of the pictures differ.

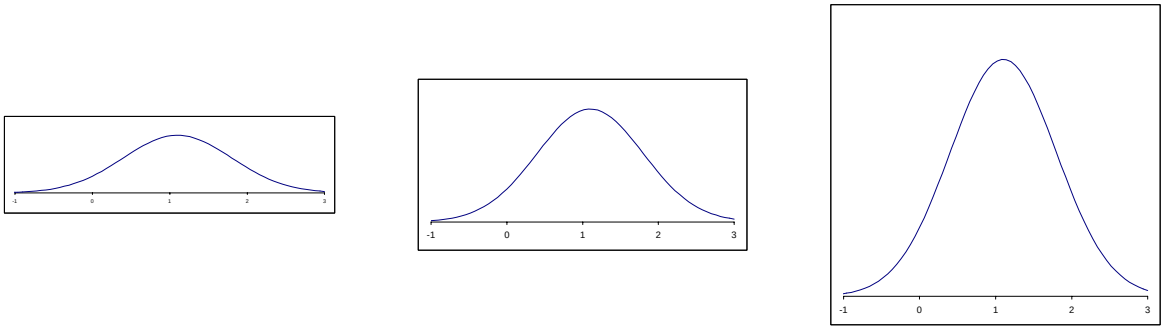



Figure G.26. Three times the same normal distribution

- What is the meaning of $y(\theta)$, the y-value of the function rule (G.10)? It is certainly not a frequency or a proportion or a probability. We know that in a normal distribution most values are concentrated around the mean (where the y-value is largest), and less so further away from the mean (where the y-values are small). Another term for concentration is density, and the name of the y-values is called the **probability density** (or sometimes density for short) and the function rule (G.10 for the normal distribution) is called the **probability density function**. In a graph of the normal distribution, probabilities are represented by areas. The whole area equals one, and the area under the curve for theta values running from minus infinity up to the mean equals one half, meaning that there is a probability of 0.5 to observe a value smaller than the mean upon a random draw from the distribution.

Now we are ready to discuss the posterior distribution. It is also a distribution of the values of theta, which is a continuous variable, and just as with the normal distribution (the prior), we will need a rule (a probability density function) for the posterior. In applications of IRT, this posterior distribution is generally not the normal distribution, and we should realize that for each response pattern there is another posterior distribution. There exists a very famous rule which is the result of a celebrated theorem by Thomas Bayes (after whom an important branch in statistics is named: Bayesian Statistics; the theorem was proved in 1763):

**The posterior density is proportional to the product of the prior density and the likelihood.**

The application to our spreadsheet example is now very simple: in column G the likelihood for the response pattern (1,1,0,0) was computed (and later on multiplied with a constant: see Section G.8.3) and in column I the prior densities are stored, but also multiplied by a constant because we left out the normalizing constant. If we make cell J3 the active cell, we can apply the formula:

=g3*i3

and then drag it down to cell J103. Notice that in column J we did not compute densities, but values which are proportional to the wanted density. To have the real densities we should multiply the values in column J with some number, but this number is generally very difficult to determine exactly. If we plot a single posterior distribution, this number is not important, because EXCEL will scale x- and y-axes to produce a rather good looking graph.

A problem, however, may crop up if we want to make a graph of the prior and the posterior distributions in the same picture. The problem has to do with the concept of proportionality. We explain it with an example. Suppose we have computed prior and posterior densities correctly (using the correct normalizing constant), but then we multiply the column of the prior densities with 1,000 and divide the posterior densities by 1,000. The result will be that the transformed priors will be approximately 1,000,000 times as large as the transformed posterior densities, and if we plot both distributions within the same frame of axes, the posterior distribution will not be visible (unless the length of the y-axis is about ten kilometers). More generally, this means that we must make the y-values of both distributions comparable.

The total area under the graph of a distribution equals one (undefined unit of area). But this also means that if we plot two distributions, their areas should be **equal to each other**. There is a simple way to compare plotted areas of a distribution: we could plot the distribution also as a histogram, a collection of rectangles (101 in the example), all having the same base, but heights equal to (or proportional to) the values listed in the relevant column of the spreadsheet. The total area of these rectangles will be very close to the total area under the graph of the function. To find this total area under the histogram, all we have to do is to take the **sum** of the density values we use.

A convenient way to compute and store the sum of the values in a column is to use the built-in function SUM in the cell just under the last value computed. For the prior densities this will be cell I104 and for the posterior densities cell J104. Making cell I104 active and typing the formula

=SUM(I3:I103)

will display the sum of the prior densities. In the example used up to now, this gives a value of 29.16. The sum of the posterior densities is 16.74 (computed with the SUM function in cell J104). If we plot prior and posterior with the values as stored, the area under the graph of the prior will be 29.16/16.74 = 1.74 times as large as the area under the graph of the posterior. To make them equal, we should multiply the posterior densities by a factor 1.74. So we can recompute column J, by defining in cell J3 the formula

=g3*i3*1.74

and dragging until cell J103. The sum will be automatically adapted in cell J104, and should be equal (up to rounding error) to the number displayed in cell I103. It is with this technique that Figure G.17 has been constructed. Notice that the y-axis has been deleted, because it has a different meaning for the two curves.