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

Section B: Standard Setting

Language Policy Division, Strasbourg

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

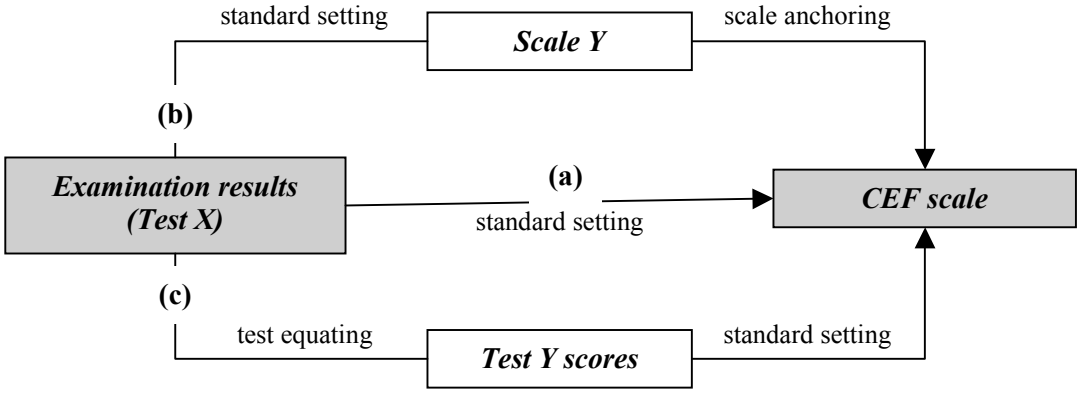
The linkage between language examinations and the Common European Framework for Language (CEF) means the establishment of a correspondence between examination results and CEF levels of language proficiency. This correspondence can be established in at least three different ways:

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.

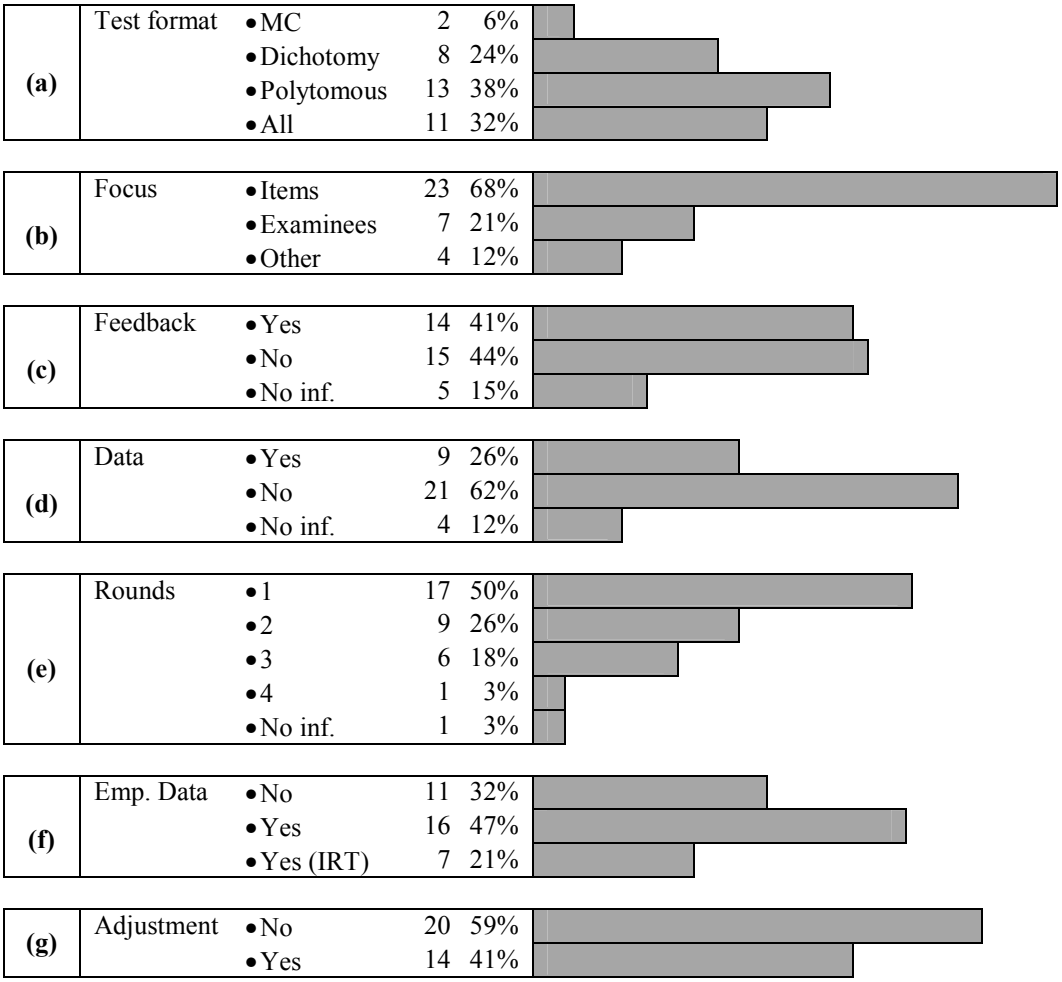
Column 11 (***Decision rule***) briefly describes the decision rule applied for cut-off score establishment. It should be mentioned that in many cases different decision rules can be applied to the same set of judgments and most likely different approaches will yield different cut-off scores. The adequacy 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 α 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** κ takes into account the probability of agreement by chance. Kappa (κ) 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 κ = **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 κ = **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 κ differs significantly from the chance agreement ($p < .05$) while in the second case κ 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 κ. 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.

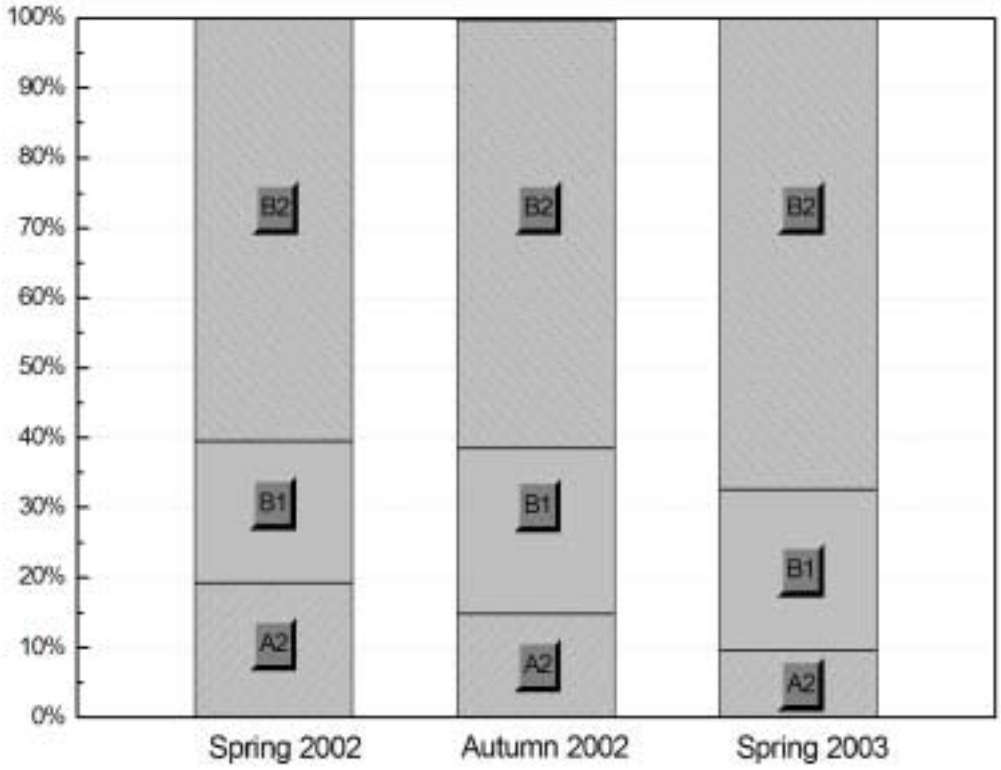
4.2.5. *Pass rate*

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: • Percent of partially correct • Typical score • Mean scores • 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 OE items	Items	• Item classification in two-way table (relevance-diffi-culty) • 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 • Mastery level	• Item classification • Level specific pro-bability of success	Yes	Yes	2	• Individual + Revision • Group Consensus	Median θ 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 • Mastery level	• Item domain cla-ssification • Probability of suc-cess	No	No	1	?	θ, corresponding to the established probability of success	Yes (IRT)	No
21.	Cognitive Components	Reckase, 2000	All	• Items • Cognitive componen ts	• Item decomposi-tion in cognitive components • 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 • Profiles • Examinee performance	Item classification Profile classification 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 Brandon, 2002 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 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 Linear regression	Yes	Yes
30.	Measuremen t Decision Theory	Rudner, 2003	Dichoto-mous items	• Population • Items	• Proportion at each level • Level specific item difficulty	No	No	1	Individual	Maximum a posteriori decision criterion	Yes	No
31.	Hofstee	Case & Swan-son, 1998 Huff, 2001	All	Score distribution	• Min & max failing rates • 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 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 • 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