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 E: Generalizability Theory

Language Policy Division, Strasbourg

Section E

Generalizability Theory

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

This report contains four sections. The first two sections give a non-technical introduction into generalizability theory (G.T.). In the third and fourth sections the same problems are treated in a somewhat more technical way.

It is interesting to notice that a very basic term of Classical Test Theory is not well defined. In explaining the concept of measurement error in the manual and in Section C, reference was made to repeated observations under ‘similar’ conditions, but ‘similar’ was not defined precisely. An often used example of a cause of (negative) measurement error is the noise in the testing environment. But suppose a student is only tested in his school. If the school is located in a very noisy environment, and if noise has indeed a negative impact on test performance, it will maintain this negative impact (because it is constant) on retesting or administration of a parallel test. In such a case the noise is to be considered systematic influence, and its impact cannot be conceived of as measurement error; it will lower the true score of the student. If one wants to have an idea about the magnitude of the negative impact of noise, one will have to conduct an experiment to find out. (A good experiment would be to administer the test to two equivalent samples in two different conditions - quiet and noisy - and to compute the differences between the average test scores in both conditions.)

An important way of controlling for such systematic effects is the standardization of the test administration, which, for example in the case of a listening test, could prescribe that headphones are to be used. It is, however, impossible to control for all possible sources of disturbance. A typical example occurs when the item scores have to be determined by means of ratings by some rater, e.g., by the teacher. Some teachers are more lenient than others, and if a candidate happens to get (always) a lenient teacher his true score will get higher than with a harsh teacher.

To find out whether differences in leniency of the raters make a lot of difference in the scores, one has to investigate this in a special study. Such an investigation can be supported by a psychometric theory that is able to quantify these differences. A theory which is especially created for this purpose is **Generalizability Theory** (G.T.), which was published in a series of articles in the 1960s, and as a book in 1972[1].

In this theory, measurements are described in terms of the conditions where they are observed. A set of conditions that belong together is called a **facet**. In this way, ‘items’ is a facet of the measurement procedure. The measurement object is usually the person who is tested, and the basic observations are usually collected by observing all persons in the sample with all items in the test, i.e., persons are crossed with a number of conditions (specific items) from the facet ‘items’, and such a set-up is called a single-facet crossed design. But sometimes more facets are involved: it is possible that the answers by persons to items are to be rated by a number of raters. If the answer of each person to each item is rated by each rater (from a well-defined group of raters), we have a crossed two-facet design: the facets are ‘items’ and ‘raters’. (At least, this is the description one usually finds in textbooks on G.T.; we will come back to this example in later sections.)

[1] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley. A more recent and more accessible book is: R.L. Brennan, (2001). *Generalizability Theory*. New-York: Springer-Verlag.

Two important views can be taken with respect to the conditions of the facets: they can be considered as **fixed** or as **random**. (The principle is the same as in the analysis of variance; in generalizability theory the concepts of analysis of variance are used throughout.). In the fixed case, the conditions are taken as they are: if items are considered as fixed, this means that we are interested in the very items that are part of the test. In the random case, the conditions are considered as a random sample from a much bigger collection of conditions. Such a collection is called a **universe**.

Two conditions will be considered in detail: the one-facet crossed design (persons by items) and the two-facet crossed design (persons by items by raters).

E.1. The persons by items design

Consider the Classical Test Theory as it was presented in the manual and in appendix C. The true score was defined as the average or expected score **on the very same test** (under repeated administrations). This means that the items are considered as fixed. But we could also draw a new random sample of (the same number of) items from the universe of items at each administration, and then compute the expected score of the person over these test administrations. This expectation is called the **universe score** of that person. It is clear from this that a particular observed score will deviate from the universe score not only because of measurement error but also because of the factual composition of the test that has been used: if the items in the test happen to be relatively easy, the observed score will probably be higher than in case the items in the (randomly composed) test happen to be relatively difficult. This means that in the random view, the difficulty of the items now has to be considered as an extra source of variance (of the observed scores).

But generalizability theory considers yet an extra source of variance. To see this, imagine that the basic observations are arranged in a two-way table, the rows associated with the persons and the columns with the items. A particular cell contains the observed item score for the person of that row on the item of that column. Then four sources of variability are (in principle) distinguished:
- the persons (having different universe scores);
- the items (having different difficulties);
- the interaction between persons and items (John is especially good on items 1 and 2, while Mary performs especially poorly on item 3 but especially well on item 17, etc.);
- the measurement error.

With each source of variability there is a corresponding variance, and the theory makes it clear that the total variance in the two-way table is the sum of those four variances. These four variances are called **variance components**. The main purpose of the analysis of the two-way table is to estimate (from a single sample) these variance components. But since there is only one observation per cell, it is impossible to estimate the interaction component and the error component separately (interaction and error are confounded); only their sum can be estimated. This sum is usually called the residual component. The variance components are usually estimated by techniques of analysis of variance.

If the variance components are known, then some interesting correlations can be predicted. In the case of a two-way design (one facet), two correlations are interesting:

1. The correlation between the actual scores of the persons and their scores on an independent replication **with the same items.** Notice that this correlation is the reliability of the test (see Section C). Unfortunately, to predict this correlation one has to know the interaction component and the error component, and since they cannot be estimated separately, one has to be satisfied with an approximation. The approximation used in G. T. happens to be identical to Cronbach’s alpha, and it can be shown that this approximation equals the true coefficient only when the interaction component is zero.
2. The correlation between the actual scores of the persons and their scores on another test (with the same number of items). This latter test has to be randomly drawn from the universe of items. It will be clear that in this case the items are considered as random.

These two correlation coefficients are called generalizability coefficients. More technical details and questions of interpretation are discussed in Section E.3.

E.2 The persons by items by raters design

In the one-facet design, it is usually not difficult to construct the two-way table needed for estimating the variance components, since the data (the responses to the items) are commonly collected on the calibration sample. Moreover, to get a stable estimate of the variance components one needs a reasonable number of persons and a reasonable number of items, but in the usual procedures of internal validation this is no problem (40 items is a reasonable number). If one uses a second facet (raters) in a crossed design, things become more complicated: for the analysis one needs a **three-way** table, which one can consider as a piling up of a number of two-way tables. Each two-way table (a layer in the pile) has the same structure as in the one-facet design, but corresponds to a single rater. To estimate the variance components, one needs at least two layers, but to have stable estimates, one needs more. Suppose that a test constructor can use ten raters. Usually, it is a lot of expensive work to have all raters rate the responses of all persons in the sample to all items in the test. Therefore, one uses only a subset of persons (drawn at random from the calibration sample), and (if there are many items) a subset of items. For this (these) subset(s), all available raters rate all answers in order to have a completely filled three-way table. (Incomplete three-way tables are very difficult to handle when estimating variance components[2].) This special data collection together with its analysis to estimate the variance components is called a G-study. It is good practice to carry out a G-study when using raters.

In the two-facet crossed design there are eight variance components: three components associated with main effects, three first order interactions, one second order interaction and one error component. The three main components are associated with persons, items and raters, respectively. The raters component refers to different degrees of leniency of the raters. The three first order interaction terms are listed below, together with a typical example to illustrate the ideas:
- person-item interaction: John is especially good on item 1;
- person-rater interaction: Rater A is especially lenient with Mary;
- item-rater interaction: Rater A is especially lenient when rating item 1.

A second-order interaction then occurs when rater A is especially harsh with John when rating item 1. Since in the three-way table, we have only one observation per cell, the second order interaction and the error are confounded, so that their variance components cannot be estimated separately; only their sum (the residual component) can.

At this point, however, a serious problem with respect to the correct interpretation of the variance components must be noted, because the three-way table (students by items by raters) may come about in two quite different ways, which we illustrate by the following example. A number of young musicians has to play a number of fragments from different composers, and each performance has to be scored by a number of jury members. The fragments play the role of items; the jury members act as raters. The whole contest may be arranged (at least conceptually) in two different ways. Firstly, it may be that each student plays each fragment only once in the presence of the whole jury (which is what usually will happen); but, secondly, it might well be that each student plays all fragments in turn for each jury member. In both cases the data collection will be arranged in a similar three-way table, and in both cases the analysis will be carried out in an identical way, but the interpretation of the variance components is different. In the former case the jury members all judge the very same performances, and it may happen that a single performance (of John, say, playing a fragment of Brahms) is incidentally quite poor, which means the judged performance may be infected by a negative measurement error, but this will lead probably to a low score given by all raters. This means, in more general terms, that the scores given by the raters will be correlated. Because of this dependence on the same measurement error in a single performance it is better to conceive of such a set-up as a **nested**

[2] Special software to estimate variance components in the two facet design with missing observations can be obtained on request from Ton.Heuvelmans@citogroep.nl

design (the raters are nested under the student-item combinations; even if for all student-item combinations the same set of raters has been used[3]). In the latter case, where each student plays each fragment (independently) for each rater, the measurement errors in the performances are assumed to be independent, and we have a genuine crossed design. Of course such a set-up will probably never occur in educational settings, and it is remarkable that the nested design (which is the usual way of data collection) has been treated in G.T. as if it were a truly crossed design. A more technical treatment of this problem will be given in Section E.4.

As an example, the results of a G-study are given for a number of countries which participated in the first cycle of PISA[4]. The items were reading items (in the Mother Tongue) used for a scale that was called Retrieving Information. The number of students participating in the G-study varied between 48 and 72 (depending on the country), the number of items is 15 and the number of raters is 4. See Table E.1. Notice that in this case the students answered each item only once. In a G-study, the numerical values of the variance components are not important, only their relative contributions to the total variance matters. Therefore, one usually reports the different components as a percentage of the total variance. This is done in Table E.1: the numbers in each row add up to 100.

Table E.1. Variance components in the first cycle of PISA for a reading scale (expressed as a percentage of the total variance)

	Students	Items	Raters	S x I	S x R	I x R	residual
Australia	22.40	19.01	-0.02	50.36	0.01	0.22	8.01
Denmark	13.24	24.56	0.01	54.22	0.16	0.25	7.56
England	14.79	22.14	0.00	59.71	0.01	0.00	3.35
Finland	18.97	18.30	0.02	55.93	-0.11	0.07	6.81
Norway	15.66	17.79	0.00	61.43	0.21	0.17	4.74

A number of interesting observations can be made fromTable E.1. An extensive discussion can be found in Section E.4. We make only three observations here:

1. Two numbers in the table are negative. Although variances cannot be negative, their estimates can, which usually indicates that the true variances are near zero. It is customary to treat small negative values as zero.
2. The three shaded columns involve the raters: one as a main effect and two in interaction with either students or items. We see that in all three columns the contributions to the total variance are very small, and for all practical purposes negligible. This result was the basis for the decision taken to let the items be rated by a single rater (for all students not involved in the G-study). In Section E.4, some critical remarks on this decision will be made. For now, it is important to realize that the three shaded columns point to the almost complete absence of systematic rater effects: there are no systematic overall differences in leniency (the main effect component is almost zero), and there are no systematic interactions of raters with students and items. The low student-rater-interaction component is to be expected, since the students came from a national sample in each country and were unknown to the raters; the low rater-item-interaction component means that there were no systematic differences in scoring some of the items, and this may be due to a large part to the careful construction of the rating rules, and to all kinds of measures taken in the PISA project to check that these rules were followed meticulously. But it does not necessarily mean that the agreement between raters was very high, because there might have been **unsystematic** differences between raters which were not taken into account in the PISA study. A detailed discussion of this problem can be found in Section E.4.

[3] The usual way of conceiving nesting is where all instances of one facet are specific to each instance of the other facet. A typical example in educational measurement is the facet schools and the facet students. One says that students are nested within schools, and of course, one assumes that each student belongs to only one school. This unique assignment, however, is not necessary to have a nested design.

[4] PISA stands for Program for International Student Assessment. An overview of the first cycle is given in Knowledge and Skills for Life (2001). More details can be found in PISA 2000, Technical Report (2002), Edited by R. Adams and M. Wu. Both volumes are published by the OECD (Paris).

3. Probably, the most puzzling result in Table E.1 is that the most important variance component is the interaction component between students and items, accounting in each country for more than 50% of the total variance. This result is especially remarkable if it is compared to the residual component which takes relatively modest values in the PISA study. This finding will be commented upon in detail in Section E.4.

As a final comment of this section, it must be emphasized that in collecting the data for a G-study, the raters must work independently of each other. Joint decisions by the raters may look attractive for a number of reasons, but they make the results of a G-study misleading and useless.

E.3. Generalizability Theory for the one-facet crossed design

Generalizability Theory is a statistical theory which is highly similar to Classical Test Theory, but it is more general. In every theory, the starting point consists of a number of assumptions. Because it is a mathematical theory, these assumptions are usually expressed by mathematical statements (as a formula). The whole of the assumptions is called a **model**. In Section E.3.1 the model will be introduced and some comments will be given on the estimation procedures, while section E.3.2 will be devoted to the use one can make of the results of the analysis.

E.3.1 The model

We start with the model for a one-facet crossed design (the facet being ‘items’). Variables will have one or two subscripts; the subscript *p* refers to a person (a test taker), and the subscript *i* to an item. The basic observed score is the score of person *p* on item *i*, and this score is denoted by Y_{pi} . In the model this score is considered as the sum of five parts, called effects: a general effect, a person effect, an item effect, an interaction effect (between person and item) and a measurement error. Symbolically, this is written as

$$Y_{pi} = \mu + \alpha_p + \beta_i + (\alpha\beta)_{pi} + \varepsilon^*_{pi} \tag{E.1}$$

1. The Greek letter μ symbolizes the general effect. It corresponds to the average item score, where the average is to be understood as the average in the population of persons and across all the items in the universe.
2. The person effect is α_p. It is an unknown number and every person in the population can be characterized by a person effect. So, generally speaking, the person effect is a **random variable**, which has some distribution in the population of persons. The population average of the person effects is set to zero. (This is a technical restriction, without which the model cannot ‘work’). The practical implication of this restriction is that person effects have to be considered as deviations from the mean: a positive person effect means an effect greater than the average, and a negative effect means an effect smaller than the average. The main problem in the analysis is to estimate the variance of the person effects. This variance will be symbolized as σ^2_α.
3. The item effect is β_i. It is a random variable in the universe of items, with mean equal to zero. Its interpretation is completely analogous to that of the person effect. The variance of the item effects is symbolized as σ^2_β.
4. The interaction effect is symbolized as $(\alpha\beta)_{pi}$. A double symbol is used to indicate this interaction; it is not to be understood as a product. (The subscripts *p* and *i* refer to the whole symbol, and therefore the symbol is placed between parentheses.) So, like person effects and item effects, $(\alpha\beta)_{pi}$ is an unknown number which applies to the particular combination of person *p* and item *i*. For every possible combination of a person from the population and an item from the item universe, there is such an interaction effect. The average of these effects is set to zero, and the problem to be faced is the estimation of the variance $\sigma^2_{\alpha\beta}$ of the interaction effects.

5. The measurement error is symbolized as ε^*_{pi}, which is also a random variable with mean zero. Its variance is $\sigma^2_{\varepsilon^*}$.
6. There is one important assumption to be added: it has to be assumed that all random variables in the right hand side of equation (E.1) are independent of each other. Using this assumption, a very useful result from statistics follows directly: the variance of the item scores (across the population of persons and across the universe of items) is just the **sum** of σ^2_α, σ^2_β, $\sigma^2_{\alpha\beta}$ and $\sigma^2_{\varepsilon^*}$.These four variances are called the **variance components**.

The main purpose of a so-called G-study is to estimate these four variance components. To do so, one needs to administer a **random sample** of items (from the universe) to a **random sample** of persons (from the population). One can store the item scores thus obtained in a rectangular table where the rows correspond to the persons and the columns to the items, and each cell contains the observed item score (obtained by the row person on the column item). If the items are administered only once to each person (as is commonly done), then, unfortunately, it is impossible to estimate the variance components of the interaction and the measurement error separately; only their sum can be estimated. (Technically one says that interaction effects and measurement error are confounded. This confounding can also be deduced from formula (E.1): the interaction effect and the error have the same pair of subscripts. If there were more than one observation for the same person-item combination, the error term (and only this one) would have an extra subscript indicating the replication.) Although we started with a model as detailed as reflected in equation (E.1), we will have to simplify it a little bit. We do so by defining

$$\varepsilon_{pi} = (\alpha\beta)_{pi} + \varepsilon^*_{pi} \qquad \text{(E.2)}$$

The random variable ε_{pi} is called the **residual effect**, and its variance is called the residual variance.

The main purpose of the analysis to be carried out on the data table is to estimate the person variance (σ^2_α), the item variance (σ^2_β) and the residual variance (σ^2_ε). The analysis can be carried out by standard software like SPSS. An important condition, however, is that the table is complete, i.e., there must not be any empty cells.

E.3.2 Generalizability coefficients

In the literature on Generalizability Theory, much attention is given to so called generalizability coefficients. These coefficients are in some sense generalizations of the reliability coefficient from classical test theory. The latter, however, can also be expressed as a correlation: the correlation between two series of test scores from parallel tests. In the same way, generalizability coefficients can be considered as correlations between two series of tests scores, but to understand them well, we need to be rather precise as to how both tests are defined.

We need some more notation here. We will indicate the number of items in the test by the capital letter *I*. Of course, when we take decisions on persons, these decisions are based on the test score, and not on individual item scores. To arrive at relatively simple formulae, we will work with mean test scores, and we will denote them by the symbol Y_p defined as

$$Y_p = \frac{1}{I}\sum_{i=1}^{I} Y_{pi}$$

Applying the model (E.1) to the mean test score in the one-facet design gives

$$Y_p = \mu + \alpha_p + \frac{1}{I}\sum_{i=1}^{I}\beta_i + \frac{1}{I}\sum_{i=1}^{I}(\alpha\beta)_{pi} + \frac{1}{I}\sum_{i=1}^{I}\varepsilon^*_{pi} \qquad \text{(E.3)}$$

Now we will distinguish three cases. In the first case we want to have an expression for the correlation between two series of test scores coming from administering the same test twice (and assuming that there are no memory effects, yielding scores on two parallel tests). In the second case two tests are used, one for the first administration and one for the second. The two tests have the same number of items, but are randomly drawn from the universe of items. In the third case, we want the correlation between two series of test scores in a rather peculiar situation where every person gets his/her own pair of tests. All the tests consist of *I* items, but for each person two independent tests of *I* items are drawn randomly from the universe of items.

The right hand side of equation (E.3) contains five terms whose sum is the mean score. Now we can ask for each term if it contributes to the variance of the mean scores and if it contributes to the covariance between the two mean scores. In the first case (same items for everybody) the general effect μ and the average item effect are the same for all persons and do not contribute to differences in mean scores. The person effect, the average interaction effect and the average measurement error may differ from person to person and will thus contribute to the variance. Terms which contribute to the covariance are those terms which are identical in both test administrations: this holds for the person effects and for the average interaction effect, but not for the measurement error which is assumed to be independent in each test administration. In general, terms which contribute to the covariance also contribute to the variance. So we can summarize the preceding discussion in a table, like in Table E.2, in the column labelled ‘one test’).

Table E.2. Contribution to variance and covariance (one-facet design)

	one test	two tests	2n tests
Constant	μ, β	μ, β	μ
Variance and covariance	$\alpha, (\alpha\beta)$	α	α
Variance only	ε^*	$(\alpha\beta), \varepsilon^*$	$\beta, (\alpha\beta), \varepsilon^*$

In the second case of two different tests, the only change is that the interaction effects will not contribute to the covariance, because the two tests are independently drawn from the universe. The two tests may be of unequal difficulty, but since the same test is used for all persons on each occasion, this difference in difficulty will not contribute to the variance within each test separately. In the third case, where everybody gets two independent tests, the item effects will contribute to the variance, because some persons will happen to get an easy test and some others will have a rather difficult test. The item effects and the interaction effects, however, will not contribute to the covariance, because they refer to two tests independently drawn from the universe.

To compute the correlation between the scores obtained in the two test administrations, we need the variances of the terms in the right hand side of (E.3). We take one term to illustrate how this variance comes about. To understand the result, we need two easy-to-prove but fundamental results from statistics, which we give here. Let X and Y represent two random variables and let *c* be a constant. Then

$$\text{Var}(cX) = c^2\text{Var}(X)$$

and

If X and Y are independent then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$

If we apply these rules to the variance of the mean item effect, we find that

$$\text{Var}\left[\frac{1}{I}\sum_{i=1}^{I}\beta_i\right] = \frac{1}{I^2}\text{Var}\left[\sum_{i=1}^{I}\beta_i\right] = \frac{1}{I^2}\sum_{i=1}^{I}\text{Var}(\beta_i) = \frac{1}{I^2}\sum_{i=1}^{I}\sigma_\beta^2 = \frac{\sigma_\beta^2}{I}$$

To find the expression for the correlation we have to take a ratio: the numerator consists of the sum of all variance terms contributing to the covariance, and the denominator is the sum of all variance terms. Referring to Table E.2, we find that the correlation in the first case (symbolized by ρ_1) is given by

$$\rho_1 = \frac{\sigma_\alpha^2 + \dfrac{\sigma_{\alpha\beta}^2}{I}}{\sigma_\alpha^2 + \dfrac{\sigma_{\alpha\beta}^2 + \sigma_{\varepsilon^*}^2}{I}} \qquad \text{(E.4)}$$

Similarly referring to Table E.2, we find for the second case that

$$\rho_2 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \dfrac{\sigma_{\alpha\beta}^2 + \sigma_{\varepsilon^*}^2}{I}} \tag{E.5}$$

and for the third case:

$$\rho_3 = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \dfrac{\sigma_\beta^2 + \sigma_{\alpha\beta}^2 + \sigma_{\varepsilon^*}^2}{I}} \tag{E.6}$$

There are a number of interesting observations to make about these three correlations:

1. If we know the variance components from a G-study (or have good estimates for them, which we substitute for the true but unknown values), we can compute the correlations for any value of the number of items. In all three cases it is true that the larger the number of items, the larger the correlation will be, and if the number of items is very large, all three correlations will become very close to one.
2. For any value of I, the three correlations are always in the same order:

$$\rho_3 \le \rho_2 \le \rho_1$$

3. Unfortunately, in a one-facet design ρ_1 cannot be computed, because we do not have separate estimates of the interaction component and the measurement error variance (measurement error and interaction effects are confounded). Therefore one uses ρ_2 instead, but from a comparison of formulae (E.4) and (E.5) we can easily see that both coefficients are equal if and only if the interaction component is zero; otherwise $\rho_2 < \rho_1$.
4. It has been shown mathematically that coefficient ρ_2 is equal to Cronbach's alpha, while we derived ρ_1 as the test-retest correlation (under the assumption of no memory effects). So ρ_1 is the reliability of the test in the sense of classical test theory. Cronbach's alpha will be smaller than the reliability unless the interaction term is zero.
5. Although one may regret that the coefficient ρ_1 is not available in one-facet designs, one should also be aware of the limitations of this coefficient, because it expresses the correlation between two test series based on exactly the same test. If interactions between students and items are really effective, the correlation ρ_1 will depend in a substantial way on the **specific** interaction effects in the test. If at the second administration the test is replaced by a parallel form, a quite different pattern of interaction effects may come about. One could think about this in very concrete terms: It is possible that John practiced hard last week, and he is lucky that some items in the test are very similar to the questions of his last-week exercises. So he profits from some coincidence. If, upon a second administration the very same items are used again, he will profit a second time, but in such a case the possibilities of generalization are quite narrow: we are in some sense only entitled to say that John is good at what the test measures if we stick to the very same set of items of which the test is composed. By dropping the item by person interaction term from the correlation formula (in the numerator), we just get rid of these coincidences, but that is precisely what is expressed by coefficient ρ_2. In Generalizability Theory ρ_2 is called the **generalizability coefficient for relative decisions**, because in principle it does not matter which items from the universe are chosen to compare (rank) different persons.
6. If one wants to know the level of proficiency in a more absolute way, of course it does matter which items are included in the test. A good example is a test of vocabulary. Suppose the test items ask for giving the meaning (e.g., by a translation) of 50 words. One might conceive the 50 items in the test as being randomly chosen from some lexicon or some corpus, the universe. The proportion of correctly answered items in the test is then to be seen as an estimate of the proportion of words mastered in the whole universe. This measure will not only show variation because of measurement error, but also because of sampling error in composing the test: scores will vary from test to test because of the varying difficulty of the included items and because of

interaction effects with the persons. The coefficient ρ_3 expresses the correlation between two series of test scores, based on randomly composed tests. In generalizability theory it is known as the **generalizability coefficient for absolute decisions**.

E.4 Generalizability Theory for the two-facet crossed design

As was noticed in Section E.2, data which are collected in a complete three-way table (students by items by raters) are usually treated as data in a two-facet crossed design, but we have distinguished between a genuine crossed design (unrealistic but conceivable), and a special case of a nested design where the student answers each item only once and each such response is rated by the same set of raters. This latter case is ubiquitous in educational measurement, and will be denoted here as the two-facet nested design.
In section E.4.1 the genuine crossed design will be treated; in Section E.4.2 the nested design will be discussed.

E.4.1 The genuine two-facet crossed design

For the two-facet (items and raters, say) crossed design, the model is a straightforward generalization of model (E.1). But now we have to use three subscripts, *p* for the person, *i* for the item and *r* for the rater. The model is given by

$$Y_{pir} = \mu + \alpha_p + \beta_i + \gamma_r + (\alpha\beta)_{pi} + (\alpha\gamma)_{pr} + (\beta\gamma)_{ir} + (\alpha\beta\gamma)_{pir} + \varepsilon^*_{pir} \qquad \text{(E.7)}$$

The three double symbols between parentheses indicate **first order** interactions. There are three of them: a person-item interaction, a person-rater interaction and an item-rater interaction. The triple symbol indicates the **second order** interaction. Examples of the meaning of such interaction terms are given in Section E.2. The typical data needed to estimate the variance components are now the answers of a sample of persons to a sample of items (from the universe of items) as rated (independently) by a random sample of raters (from the universe of raters). All these ratings can be arranged in a three-dimensional array, with as many layers as there are raters. Each layer is a rectangular table just as in the one-facet crossed design. Since each cell of this table contains just one observation (the rating by rater *r* of the answer of person *p* to item *i*), the second order interaction effect and the measurement error are confounded, and we need to take them together as a residual which is now defined as

$$\varepsilon_{pir} = (\alpha\beta\gamma)_{pir} + \varepsilon^*_{pir}$$

Notice that in this case it is perfectly possible to estimate variance components of the three first order interactions. But this is only possible in the genuine crossed design where the student answers as many times to each item as there are raters.

With techniques of the Analysis of Variance one can estimate seven variance components: three for the main effects ($\sigma^2_\alpha, \sigma^2_\beta$ and σ^2_γ), three for the first order interactions ($\sigma^2_{\alpha\beta}, \sigma^2_{\alpha\gamma}$ and $\sigma^2_{\beta\gamma}$) and one for the residual (σ^2_ε). For tabulation purposes it is suitable to convert all components to percentages, by dividing each component by the sum of all seven components (and multiplying by 100). If some components are in reality very close to zero, it may happen that their estimates are negative. Usually one sets such estimates equal to zero.

As to the generalizability coefficients, a large number of different correlations may be predicted, and one should be very careful in defining precisely the conditions of the two test administrations and/or ratings. We will consider four different cases, which are described hereafter. In all cases mean test scores are used, which are defined as

$$Y_p = \frac{1}{I \times R}\sum_i^I \sum_r^R Y_{pir}$$

i.e., the average score across items and raters. Notice that in the description of the four cases a test arrangement is described which would deliver the correlation wanted, but such an arrangement does not have to be carried out: the correlations can be predicted from the results of a G-study.

1. One test administration with the same set of *R* raters. This case is easy to implement: after a second rating the item answers are given a second time to the same set of raters, who are requested to give their ratings again. To warrant independent ratings, one usually will not tell the raters that they have rated the performances already. The correlation to be predicted is the correlation between the mean test scores for the two ratings.
2. One test administration where the performances are rated twice, each time by an independent sample of *R* raters. The data collection design consists in administering the test once to the student and to let these performances rated by two sets of *R* raters.
3. Two independent test administrations (to the same students with the same items) and each series of performances is rated by the same set of *R* raters.
4. Two independent test administrations (as in case 3) and each series is rated by a different set of *R* raters.

In all cases the needed set(s) of *R* raters are to be considered as a random sample from the universe of raters. In Table E.4 the nine effects (the nine terms in the right-hand side of (E.7)) are assigned to a constant term, the covariance between the two series or only the variance within each series. An extra row is added to indicate the confounded terms.

Table E.4. Contribution to variance and covariance (truly crossed two facet design)

Case	1	2	3	4
performances	Same		different	
sets of raters	1 set	2 sets	1 set	2 sets
Constant	$\mu, \beta, \gamma, (\beta\gamma)$	$\mu, \beta, \gamma, (\beta\gamma)$	$\mu, \beta, \gamma, (\beta\gamma)$	$\mu, \beta, \gamma, (\beta\gamma)$
Var. and cov.	$\alpha, (\alpha\beta), (\alpha\gamma), (\alpha\beta\gamma)$	$\alpha, (\alpha\beta)$	$\alpha, (\alpha\beta), (\alpha\gamma), (\alpha\beta\gamma)$	$\alpha, (\alpha\beta)$
Variance only		$(\alpha\gamma), (\alpha\beta\gamma)$	ε^*	$\varepsilon^*, (\alpha\gamma), (\alpha\beta\gamma)$
Confounded	ε^*	ε^*		

We comment on this table:

1. The constant terms are the same in all four cases. Notice that the rater effects and the rater-item interaction effect are constant also in the case of two different sets of raters, because these effects are the same within each series of ratings.
2. The interactions containing persons and raters contribute to the covariance in the case of a single set of raters because these effects are systematic. So when there is a positive effect between student John and rater one in the first series, this effect will also be present in the second series, because the combination John and rater 1 appear in both series. In the case of two different sets of raters these effects contribute only to the variance of the test scores.
3. The interaction between persons and items is always common in the two series, and therefore contribute to the covariance.
4. The most intriguing effect is the measurement error, which represents unsystematic effects which are associated with the triple combination student-item-rater. But such a combination comes about in two steps: the performance of the student on a particular item may be incidentally (in an unsystematic way) poor, for example, and this poor performance may then be incidentally rated as reasonably good by some particular rater. The total measurement error should be conceived as the sum of these two step effects, or to say it more correctly, the measurement error has two sources of variation: the student-item combination and an effect attributable to the rater. In the truly crossed design each cell represents an independent replication of a student-item-rater combination, but in the prediction of the correlations in the cases 1 and 2, the student-item combination is held constant, while only the part of the measurement error that is due to the raters is really needed. So

to be used, the variance of the measurement error should be split into two parts: one part going to the covariance row, and one part being measurement error due to the raters. But in a truly crossed G-study with only one observation in each cell of the data table, this splitting is impossible. So from such a design, the correlations in the cases 1 and 2 cannot be predicted.

5. In cases 3 and 4, where two independent test administrations are used, the two sources that influence the measurement error are active. Nevertheless, the correlation in case 3 cannot be computed, because the second-order interaction (αβγ) is needed separately for the covariances and the measurement error for the variance term. So, only case 4 is applicable.

This correlation, symbolized here as ρ_4, is given by

$$\rho_4 = \frac{\sigma_\alpha^2 + \frac{\sigma_{\alpha\beta}^2}{I}}{\sigma_\alpha^2 + \frac{\sigma_{\alpha\beta}^2}{I} + \frac{\sigma_{\alpha\gamma}^2}{R} + \frac{\sigma_\varepsilon^2}{I \times R}} \qquad \text{(E.8)}$$

where the last term in the denominator refers to the residual component, the sum of the measurement error and the second-order interaction.

It should be emphasized that the preceding formula is of little practical use because the genuine crossed design is almost never applied in educational settings with raters as the second facet. Applying formula (E.8) to the estimates given in Table E.1 (for the PISA study) does not make sense, since the G-studies to estimate the variance components were based on a special case of a nested design, where the students responded only once to each item. This case is discussed in the next section.

E.4.2 The special nested two-facets design

To model data from this design care must be taken to separate the two sources of variability in the measurement error. Therefore we will split the model in a two-step model: the first step models what happens when the student answers an item (with a given performance as the output), and the second step will model what happens when a rater rates such a performance. So the output of the first step will be the input of the second step, and the output of the second step is the observed item score given by rater *r*: Y_{pir}. The output of the first step will be conceived as a quantitative variable K_{pi} which is unobserved, but which will be treated as a kind of auxiliary variable.

To distinguish the present model from the model used in the crossed design, the symbols for the effects will be Roman letters instead of Greek letters For the first step (at the student level) upper case letters will be used, and for the second step, random variables will be denoted by lower case letters.

The first step of the model is identical to the one facet crossed design model:

$$K_{pi} = M + A_p + B_i + (AB)_{pi} + E_{pi}^* \qquad \text{(E.9)}$$

i.e., the unobserved output variable is the sum of a constant *M*, a main effect due to the person (A_p), a main effect due to the item (B_i), an interaction effect of person and item $(AB)_{pi}$ and a measurement error E^*_{pi}. The main effects, the interaction and the measurement error are conceived as independent random variables with a mean of zero and with variances σ_A^2, σ_B^2, σ_{AB}^2, and $\sigma_{E^*}^2$ respectively.

In the second step, one might conceive as if the output of the first step, K_{pi}, is amended by the rater to produce the observable rating Y_{pir}. Such amending may be influenced by a main effect of the raters, or an interaction effect between rater and person or between rater and item, or a second order effect (rater by item by person) and an unsystematic effect, a measurement error (at the rater level). Of course one can split all these effects into a mean effect (across raters, persons and items), and a deviation from the mean, and all the mean effects can be collected into a grand mean *m*. So we get as the second step

$$Y_{pir} = K_{pi} + m + b_i + c_r + (ac)_{pr} + (bc)_{ir} + (abc)_{pir} + e_{pir}^* \qquad \text{(E.10)}$$

The models (E.9) and (E.10) cannot be used separately, because the variable K_{pi} is not observed. So, both models have to be merged in some way. We do this by replacing K_{pi} in the right-hand side of

(E.10) by the right-hand side of equation (E.9), and be grouping all the terms with the same set of subscripts. The result is this (with brackets placed around sums with the same subscripts):

$$\begin{aligned} Y_{pir} &= [M+m] \\ &+A_p+[B_i+b_i]+c_r \\ &+[(AB)_{pi}+E^*_{pi}]+(ac)_{pr}+(bc)_{ir} \\ &+[(abc)_{pir}+e^*_{pir}] \end{aligned} \tag{E.11}$$

where M and m are constants, and all ten subscripted variables are random variables whose variances one might wish to estimate. But this is impossible: random variables with the same set of subscripts are confounded, and all one can achieve is to estimate the sum of their variances. We take $[B_i + b_i]$ as an example. B_i is a systematic item effect which influences the unobservable variable K_{pi} and which one might call the inherent difficulty of the item, while b_i is a systematic item effect which comes about during the rating of the performances, and which one might call the perceived item difficulty (by the raters). Confounding means that there is no way (in the nested design) to disentangle both effects, and that the only thing one can do is to estimate the variance of their sum. There are two other pairs of confounded variables. One is the second-order interaction effect and the measurement error at the rater level and the other is the confounding of the person-item interaction and the measurement error at the student level. Now, if we count the terms in the right-hand side of (E.11), counting bracketed terms as one single term, we see that we have one constant (first line), three main effects (second line), three first order interactions (third line) and a residual in the last line, which is just the same decomposition as in the genuine crossed design. This means that we can arrange the observed data in the nested design in a three way table which takes the same form as in the crossed design, and we can analyse this table in just the same way. The interpretation of the variance components, however, is different, as can be deduced from Table E.5

Table E.5 Correspondence between variance components in crossed and nested designs

Crossed design		Nested design	
Constant	μ	$[M+m]$	Constant
Persons	α_p	A_p	Persons
Items	β_i	$[B_i+b_i]$	Items
Raters	γ_r	c_r	Raters
Persons x items	$(\alpha\beta)_{pi}$	$[(AB)_{pi}+E^*_{pi}]$	Persons x items + error at person level
Persons x raters	$(\alpha\gamma)_{pr}$	$(ac)_{pr}$	Persons x raters
Items x raters	$(\beta\gamma)_{ir}$	$(bc)_{ir}$	Items x raters
Sec. order int. + error	$\varepsilon_{pir}=[(\alpha\beta\gamma)_{pir}+\varepsilon^*_{pir}]$	$e_{pir}=[(abc)_{pir}+e^*_{pir}]$	Sec. order int. + error at rater level

Now, we are ready to reconsider the four cases of generalizability coefficients that were discussed in the previous section. We reproduce Table E.4 here as Table E.6 but with the symbols used in the present section.

Table E.6. Contribution to variance and covariance (nested two facet design)

case	1	2	3	4
performances	same		different	
set of raters	1 set	2 sets	1 set	2 sets
Constant	$M,m,B,b,c,(bc)$	$M,m,B,b,c,(bc)$	$M,m,B,b,c,(bc)$	$M,m,B,b,c,(bc)$
Var. and cov.	$A,(AB),E^*,(ac),(abc)$	$A,(AB),E^*$	$A,(AB),(ac),(abc)$	$A,(AB)$
Variance only	e^*	$e^*,(ac),(abc)$	E^*, e^*	$E^*, e^*,(ac),(abc)$

Comparing Tables E.4 and E.6, we see that the row with confounded terms has disappeared in the nested design, but at the same time we see that not all four coefficients can be computed: case 1 is excluded because the components *(abc)* and e^* are needed separately, case 4 is excluded because the components *(AB)* and E^* are needed separately, and case 3 is excluded for both these reasons jointly. Therefore only the correlation for case 2 (same student performance rated by two sets of R raters) can be predicted from a G-study using a nested design.

This correlation, denoted here as ρ_5, is given by

$$\rho_5 = \frac{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB+E^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad \text{(E.12)}$$

As an example, we apply this formula to the case of Australia in the PISA study (see Table E.1), for I = 10 items and *R* = 1 rater (and replacing the negative variance component by zero), giving

$$\rho_5 = \frac{22.4 + \frac{50.36}{10}}{22.4 + \frac{50.36}{10} + \frac{8.01}{10 \times 1}} = 0.972$$

This is the prediction of the correlation one would find between two ratings (each by one rater) of the performances of a (random) sample of students on 10 items. However, one should be careful here, and not confuse this case with case 4, where the same sample of students takes the test twice, and each test performance is rated by an independent rater, which is case 4 with *R* = 1. In this case the correlation is given by

$$\rho_6 = \frac{\sigma_A^2 + \frac{\sigma^2_{AB}}{I}}{\sigma_A^2 + \frac{\sigma^2_{AB} + \sigma^2_{E^*}}{I} + \frac{\sigma^2_{ac}}{R} + \frac{\sigma^2_{abc+e^*}}{I \times R}} \qquad \text{(E.13)}$$

and it is immediately seen that the interaction component needed in the numerator is not available from the G-study. Nevertheless, we can make good use of (E.13) if we have a reasonable estimate of the person-by-item-interaction component. In the PISA study the Rasch model (see Section G) has been used as IRT model, and this model presupposes absence of interaction between persons and items[5]. So we might assume quite reasonably that the component 'person by item interaction plus error at the person level' is to be attributed (almost completely) to measurement error at the person level. Or, in other words, that the person by item component is zero. If we apply formula (E.13) with this assumption to the case of Australia with *I* = 10 item and *R* = 1 rater, we obtain

$$\rho_6 = \frac{22.4}{22.4 + \frac{50.36}{10} + \frac{8.01}{10 \times 1}} = 0.793$$

which is a marked difference with the previous result of 0.972[6,7].

[5] This absence of interaction is at the level of the latent variable, and does not preclude interaction at the level of the observed scores. Extensive simulation studies (with a crossed design) have shown, however, that the person-by-item-interaction component at the observed score level usually is below 5% of the total variance.

[6] In the Technical Report of Pisa 2000, a formula similar to (E.12) was used, but the result was erroneously interpreted as a correlation with two independent administrations, like formula (E.13). Moreover, the formula used in the Pisa report also contains an error, because the rater effect and the rater by item interaction were erroneously considered as contributing to the variance. But since the estimates of these effects were negligible, this latter error had no noticeable effect on the results.

[7] If the interaction component is set to 5% of the total variance (and consequently the error at the person level at 50.36% - 5% = 45.36%), the result for ρ_6 is 0.811

The use of the results of a G-study, however, is much more versatile than the preceding example suggests. On can use the formulae (E.12) and (E.13) (and many others) to predict the correlations for different values if *I* and *R*. One might, for example, investigate whether the correlation ρ_6 would increase more by doubling the number of items or by doubling the number of raters in a future application. Applying any of these strategies will lead to doubling the total amount of rating time and costs while the first strategy will lead to doubling of the test taking time. In Table E.7, formula (E.13) has been computed with the results of the G-studies displayed in Table E.1 for 10 and 20 items and for 1 and 2 raters, and using the assumption that the true person by item interaction component is zero throughout.

The results are very easy to interpret in this case: doubling the number of raters do increase the correlations marginally, while doubling the number of items leads to a much more impressive increase of the correlation. This is consistent with the order of magnitude of the residual components in Table E.1: the measurement error attributable to the students (given in the column 'student by item interaction') is much larger than the error attributable to the raters (the column 'residual' in Table E.1). To reduce the impact of the former, the number of items has to be increased (see the denominator in formula (E.13): the confounded student-level error and first order interaction component is divided by the number of items, and since this is the largest component, the impact of changing the number of items will be the most drastic. Changing the number of raters diminishes the impact of the student by rater interaction component, but since this component is negligibly small in all countries, the impact on the change of the correlation will be negligible as well. The residual term is influenced in an equal way by doubling either the number of items and the number of raters.

Table E.7 The coefficient ρ_6 for the results in Table E.1 (student by item interaction set to zero)

	I = 10		I = 20	
	R = 1	R = 2	R = 1	R = 2
Australia	0.793	0.805	0.884	0.892
Denmark	0.676	0.692	0.803	0.816
England	0.701	0.707	0.824	0.828
Finland	0.751	0.762	0.858	0.865
Norway	0.696	0.707	0.817	0.826

In conclusion we can summarize the results of the G-studies carried out in the PISA project as follows:

1. From Table E.1 we see that there are almost no systematic effects in the data due to the raters: rater main effect and first order interactions where raters are involved (the shaded columns) are negligible.
2. If the genuine student by item interaction component is assumed to be negligible, the big component in the column (S x I) has to be interpreted as measurement error at the student level, while the residual term is to be interpreted as a residual at the rater level (measurement error confounded with second order interaction). Although there is some confounding, it is reasonable to assume that the genuine interactions are much smaller than the measurement error.
3. This separation of two kinds of measurement errors (in the analysis of G-study data) is only possible in the special nested design (all raters judge on the same performances of the students), and not in the truly crossed design, where the two kinds of measurement errors are confounded.
4. Two different correlations, issuing from the nested design were studied. One (ρ_5, formula (E.12)) predicts the correlation between two series of independent ratings based on the very same student performances; the other (ρ_6, formula (E.13)) predicts the correlation between two series of independent ratings based on two independent test administrations. The former is an exact formula, the latter can only be used as an approximation, because one has to add an assumption about the student-item interaction component.
5. In the PISA study all ρ_5 correlations were very high (the present text gives only one example), while the ρ_6 correlations are substantially lower and also show substantial variation across countries. The reason why they are lower is due mainly to measurement error at the student level, which is much more important than the error at the rater level. In the light of this finding it would

have been of little use to let the performances of all students in the study to be rated by two (or more) raters. This can be clearly seen from Table E.7.

The example used in this Section may be atypical for many educational settings. In general one has to pay attention to a number of aspects when one carries out a G-study, using raters as one of the facets. We discuss these in turn.

1. The notion of random sampling in such studies is quite important. Especially the raters should be drawn randomly from the universe of raters which are possible candidates to do the rating work in large scale applications. Using only the best or most motivated raters for the G-study may invalidate the generalizability of the conclusions from such a study. Particularly, the use of only volunteers in the G-study may result in a non-representative sample. Moreover, the conditions for the rating work (allowed time, instructions, amount of training, etc.) should be the same in the G-study as in real applications.
2. In the PISA study the systematic effects associated with the raters were negligible, but this is not necessarily the case in G-studies.
    a. A substantial main effect component for the raters indicates differences in leniency. If in real applications of the test, the test score is to be compared with a pre-established standard (to succeed or to fail, for example), such differences may lead to incorrect decisions about the candidates.
    b. A substantial item-rater interaction component may be caused by different interpretation of the scoring rules by different raters. A more detailed search into the data (or an interview with the raters) may reveal that some rules are unclear or ambiguous. Although this interaction and the main effect do not appear in the formulae for ρ_5 and ρ_6, they may lower the reliability in other cases which are not discussed in detail in the present report. Here is an example. Suppose the work of 1000 students has to be rated (in an application), and one uses 10 raters to do the rating work, each rater rating 100 performances. If there are systematic differences between the raters, these will cause irrelevant (and therefore unreliable) variation in the test scores.
    c. A substantial student-rater interaction component is a serious problem. It may show up if some of the raters happen to know (and can identify) some of the students. This is important to remember when one tries to generalize the results of the G-study to future applications. It may be that in the G-study the students are anonymous to the raters and that no such interaction appears, but in future applications most of the rating may be done by the students´ own teacher. In such a case one cannot be sure that in the application this interaction will be absent.
3. The coefficient ρ_5 is the correlation between two independent ratings (each by R raters) of the same student performances. One can compute it for different values of R (usually the values 1, 2 and 3 will suffice). If this correlation is deemed too low if $R = 1$, but acceptable for $R = 2$, this means that in future applications one has to use two independent raters for each student, which can be very costly. Of course, one could also revise the scoring rules or provide better training or supervision of the raters, but one should realize that taking such measures does not automatically remove the problem. One can only be sure about this by doing a new G-study after these measures have been implemented.
4. It may be useful to compare ρ_5 to ρ_6 for different values of R and I. The coefficient ρ_6 can be interpreted as a test-retest correlation. We have seen that its departure from the ideal value of one is due partly to the students and partly to the raters. By comparing it to ρ_5, one gets an impression whose contribution is the most important, and one can take measures to improve the reliability either by increasing the number of raters or the number of items administered to the students. The construction of a table like Table E.7 may be helpful in such a case.