Tilburg University

Introduction to the measurement of psychological attributes

Sijtsma, K.

Published in:

Measurement. Journal of the International Measurement Confederation

DOI:

[10.1016/j.measurement.2011.03.019](https://doi.org/10.1016/j.measurement.2011.03.019)

Publication date:

2011

Document Version

Publisher's PDF, also known as Version of record

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):

Sijtsma, K. (2011). Introduction to the measurement of psychological attributes. *Measurement. Journal of the International Measurement Confederation*, 44(7), 1209-1219.
<https://doi.org/10.1016/j.measurement.2011.03.019>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Review

Introduction to the measurement of psychological attributes

Klaas Sijtsma*

Department of Methodology and Statistics, TiSSBeS, Tilburg University, PO Box 90153, 5000 LE, Tilburg, The Netherlands

ARTICLE INFO

Article history:

Received 14 September 2010

Received in revised form 24 November 2010

Accepted 16 March 2011

Available online 23 March 2011

Keywords:

Guttman model

Item response models

Psychological attributes

Psychological measurement

Rasch model

Tests and questionnaires

ABSTRACT

This article introduces the measurement of psychological attributes, such as intelligence and extraversion. Examples of measurement instruments are discussed, as well as a deterministic measurement model. Error sources that threaten measurement precision and validity are discussed, and also ways to control their detrimental influence. Statistical measurement models describe the random error component in empirical data and impose a structure that, if the model fits the data, implies particular measurement properties for the scale. The well-known Rasch model is discussed along with other models, and using a sample of data collected with 612 students who solved 13 arithmetic tasks it is demonstrated how a scale for arithmetic ability is calibrated. The difference between psychological measurement and physical measurement is briefly discussed.

© 2011 Elsevier Ltd. All rights reserved.

Contents

1. Introduction	1210
2. Measurement instruments for psychological attributes	1210
3. Basic concepts	1211
4. A deterministic measurement model	1212
5. Two error types	1213
5.1. Random measurement error and measurement precision	1213
5.2. Systematic measurement error and validity	1213
6. Probabilistic measurement models	1214
6.1. Models	1214
6.2. Estimation	1215
6.3. Goodness-of-fit research	1216
6.4. A Scale for arithmetic ability	1216
6.4.1. Goodness-of-fit investigation	1216
6.4.2. Calibrating the scale	1217
6.4.3. Practical use of the scale	1217
6.5. Differences with physical measurement	1217
7. Conclusions	1218
Acknowledgments	1218
References	1218

* Tel.: +31 13 4663222/31 13 4662544; fax: +31 13 4663002.

E-mail address: k.sijtsma@uvt.nl

1. Introduction

The goal of this article is to introduce the readers of *Measurement* to the basic ideas and principles underlying the measurement of psychological attributes, such as intelligence and extraversion. Several of the ideas and principles are also leading in educational assessment, health research, marketing, opinion research, policy research, political science, and sociology. Measurement instruments for psychological attributes are much different from measurement instruments used in physics, chemistry, biology and medicine, but several ideas and concepts are similar so that there are enough areas of recognition for *Measurement* readers.

In what follows, I first introduce psychological measurement instruments. Second, I discuss several important basic concepts. Third, I use a deterministic (mathematical) measurement model to explain how a formal model enables psychological measurement. Fourth, I discuss two error sources that threaten the quality of measurement instruments. Fifth, I discuss probabilistic (statistical) measurement models known as item response models. I provide an example of scale calibration, which shows how to use item response models and illustrates the logic of fitting models to data and drawing conclusions about scale calibration. Finally, I compare psychological and physical measurement.

2. Measurement instruments for psychological attributes

Measures of time, temperature, blood pressure and radioactivity have had a long history of unsuccessful attempts before instruments were constructed that provided precise measurement controlling as much as possible for disturbing influences. In psychology, attributes of interest are, for example, general intelligence [1]; specific abilities important in cognitive development during early childhood, such as conservation of quantities [2] and transitive reasoning [3]; personality traits, the five most important traits being extraversion, agreeableness, conscientiousness, emotional stability or neuroticism, and intellect, intellectual autonomy, or openness to experience [4, p. ix]; and attitudes, for example toward one's father or one's

body [5]. Attempts to measure psychological attributes started in the late nineteenth century [6–8] (also [9,10]) and have continued since, leading to the progressive improvement of the quality of the measurement instruments. Measurement instruments have become more precise—that is, providing repeatable measurement values by better controlling for random measurement error [11,12]—and valid—that is, better representing the attribute of interest by controlling for disturbing influences from other sources simultaneously influencing the measurement process [13–15].

Measurement instruments for psychological attributes appear different from clocks, thermometers, sphygmomanometers, and Geiger–Müller counters. I provide three examples. First, I consider the measurement of transitive reasoning [16] for which an instrument is used that consists of a set of problems, such as those shown in Fig. 1 [3,17]. The problem in the upper panel requires the child to deduct from two boxes containing differently colored sticks—the premise information—which of the two sticks on the right, which are partly hidden but identifiable by their color, is the longest. Each premise is presented on a pc screen while the other premise is invisible, and the final task must be solved in the absence of the premises. The problem in the lower panel is formally identical but involves the comparison of animals with respect to their age (told to the child by the experimenter), which deprives the child of visual information.

The choice of these and similar problems is based on the theory of transitive reasoning [3,16], and the problems operate as stimuli that invoke responses from the child that are informative about transitive reasoning. Transitive reasoning theory posits several properties that may be varied across different problems, so that several problems are presented in the measurement instrument eliciting a wide variety of relevant responses from the child. Properties may concern the logical relationships between the objects in a problem (inequalities, equalities, or a mixture of both; in Fig. 1, inequalities), the number of objects in a problem (defining the number of premises; in Fig. 1 there are 3 objects, hence two premises), and the mode of the problem (abstract or figural, as in Fig. 1). After recording incorrect responses as 0s and correct responses as 1s, a statistical

Fig. 1. Two transitive reasoning tasks.

Fig. 2. Four item response functions for the Guttman model.

measurement model [11,18] is used to analyze the children's 0/1 scores with the purpose of calibrating a scale for transitive reasoning.

Second, a measurement instrument for extraversion [19] may contain statements to which a person indicates the degree to which they apply to him/her, such as

I feel uncomfortable in the company of other people

Does not apply	☐	☐	☐	☐	☐	Applies
----------------	--------------------------	--------------------------	--------------------------	--------------------------	--------------------------	---------

The respondent is asked to rate one box. The five boxes together form a rating-scale. In different instruments, the number of boxes varies between 2 and 10 but 5 is the most frequently used number. Extraversion is a complex attribute, so that a measurement instrument typically consists of a large number of statements, each covering a different aspect of extraversion. For example, the most frequently used instrument for extraversion, the NEO-PI-R [19], uses 48 statements. The ratings for each of the statements are transformed to ordered scores, usually 0, 1, 2, 3, and 4, such that 0 stands for the lowest extraversion level for the aspect covered by the statement (here, the right-most box), and 4 for the highest level (the left-most box). A statistical measurement model is used to analyze the 0–4 scores with the purpose of calibrating a scale for extraversion.

Third, many school subjects are tested using sets of problems or questions. For example, in primary school pro-

ficiency in arithmetic may be tested using a set of arithmetic problems, one of which could be [20]:

A tower is 30 m high and casts a shadow 12 m long. The tree next to the tower casts a shadow 5 m long; how high is the tree?

Tests are often graded by assigning credit points (0 for an incorrect answer and 1 for a correct answer) to solutions given for individual problems and adding the credit points to obtain a total score, which may further be transformed to a scale well-known to the students and their parents. Intelligence measurement may also involve the ability to manipulate numbers, measured by a set of problems to which the student provides answers. The 0/1 scores are analyzed using a statistical measurement model leading to a calibrated scale. I give an example of scale calibration for arithmetic ability in the section on probabilistic measurement models.

The three examples clarify that psychological attributes are not directly observable through sensory detection but have to be inferred from the person's responses to a set of problems or statements or other kinds of stimuli not discussed here, such as building blocks, mazes and games (used in child intelligence measurement) and ranking and sorting tasks (for the measurement of attitudes and preferences). Psychological theories describe how the attributes manifest themselves in conjunction with environmental influences as observable behaviors, and posit the choice of the stimuli best suited for invoking the observable behaviors as responses.

3. Basic concepts

Psychological measurement instruments are called tests or questionnaires. A test requires maximum-performance—the person is instructed to do the best (s)he can. This is relevant in intelligence measurement, cognitive ability measurement, and educational testing. A questionnaire requires typical behavior; that is, the behavior a person usually exhibits when placed in a particular situation—the person is instructed to show who (s)he is. Typical

Fig. 3. Four item response functions for the Rasch model (left panel: δ s: $-1.5, -0.7, -0.3, 1.5$) and four item response functions for the 2-parameter logistic model (right panel: δ s: $-1.2, -0.7, 0.3, 1.5$; α s: $2, 1, 2.8, 1.5$).

behavior is required in the measurement of personality traits and attitudes.

The problems, statements, and other stimuli used in tests and questionnaires are called items. Sets of items replace the long lasting observation of a person in real life until (s)he spontaneously exhibits the behavior of interest, for example, typical of (non)intelligence. It simply would take too much time before enough evidence was collected. Thus, test and questionnaires are efficient, standardized means of collecting the relevant information.

Individuals that respond to items are often called testees (intelligence measurement), respondents (trait and attitude measurement, survey research), or examinees (educational assessment), or sometimes subjects or simply persons.

Measurement models define desiderata in terms of mathematical assumptions. For example, the different items must all invoke responses to the attribute of interest. In the measurement model, this is mathematically represented by one explanatory variable. When the formal structure of the measurement model corresponds well with the structure of the data—the 0/1 scores, or the ratings running from 0 to 4—one presumes that the model assumptions by implication hold for the scale defined by the set of items. Psychometrics is the branch of statistics that deals with the measurement of individual differences between persons on the attribute of interest, and includes such diverse topics as methods for test and questionnaire construction and validation, and statistical measurement models.

4. A deterministic measurement model

I consider an artificial example for the personality trait of introversion. Introversion (as opposed to extraversion, which is one of the so-called big-five personality traits [4]) is defined as “a keen interest in one’s own psyche, and often preferring to be alone” [4, p. 6]. In clinical contexts, interest is often with excessive introversion as part of pathologies, and in personnel selection interest is mostly with extraversion as a trait relevant to particular jobs (e.g., sales manager, teacher). I assume for the sake of simplicity that the next 4-item “questionnaire” can be used to measure introversion (the statements are different from statements used in the NEO-PI-R [19]; copy-right issues prohibit the use of items from this questionnaire):

	No	Yes
1. I like to be alone now and then	☐	☐
2. I prefer to spend New Years Eve with my closest friend	☐	☐
3. I feel uneasy in the company of other people	☐	☐
4. I will not say a word when I am among people I do not know very well	☐	☐

The respondent is asked to rate the box corresponding to the answer that best typifies what (s)he would do when finding himself in the situation described. Ratings in the

left-most box are scored 0 and ratings in the right-most box 1, reflecting that the higher score represents the higher level of introversion (it may be noted that with these four items the Yes answer is always the answer typical of the more-introvert person).

Assuming the items each reflect a different aspect of introversion as derived from the attribute’s theory, a measurement model can be used to represent the 0/1 scores obtained from a sample of persons as a single mathematical dimension. Once such a dimension has been extracted from the data it may serve as the scale on which people may be located with respect to their introversion levels.

From the respondent’s viewpoint, rating the Yes box of the four items requires increasingly higher levels of introversion as one moves from the first to the fourth item. Typical introvert people may be expected to endorse the first statement but this can also be expected from many non-introverts because almost everybody likes to be alone now and then. Even though many people like to spend New Years Eve in the company of other people, preferring to spend it with ones closest friend is not really a sign of an introvert personality but it is not as common as liking to be alone now and then. Feeling uneasy in the company of other people can happen occasionally for a large number of reasons but when this is the common reaction it may be a sign of maladjustment. This is even more likely for people who admit to keeping completely silent as a rule when being among other people.

The common view in psychological measurement is that answers to items are partly liable to uncertainty, reflecting that human behavior is partly unpredictable. Also, one particular behavior instance can have multiple causes. Thus, data also contain random error affecting measurement precision and systematic distortions affecting validity. For the moment, I assume that people respond to items without producing such irregularities, which results in perfect data. This provides an effective stepping-stone to statistical models that I discuss in the last section.

I assume that the items can be used to calibrate a scale for introversion. It is common to assume that the numbers on the scale are values of a so-called latent variable, which is denoted by θ . The items are ordered from low introversion level to high introversion level, corresponding to increasingly higher θ levels. I assume that the item levels or locations on the θ scale are represented by parameters $\delta_1, \delta_2, \delta_3$, and δ_4 . For the items in the example, I assume these values to be ordered $\delta_1 < \delta_2 < \delta_3 < \delta_4$. Generally, I use notation $\delta_j, j = 1, \dots, J$; J is the number of items in the questionnaire.

Guttman [21,22] formalized this simple idea as a measurement model. Let random variable X_j denote the scores on item j (here, $X_j = 0, 1$), and $P(A|B)$ the probability of event A given event B . The Guttman model is defined by two assumptions:

- If a person indexed v is located on the θ scale to the left of item j , (s)he is not introvert enough to rate yes; that is

$$\theta_v < \delta_j \iff P(X_j = 1 | \theta_v) = 0; \text{ and}$$

- If person v is located to the right of item j , (s)he is more-introvert than the level the item represents, and will rate yes; that is

$$\theta_v \geq \delta_j \iff P(X_j = 1 | \theta_v) = 1.$$

Fig. 2 provides a graphic representation of these two assumptions for the four introversion items. The unique feature of this model is that it prescribes that if a person answered no to, say, item 2, it is completely certain that (s)he also answered no to the items 3 and 4. Also, if we would know that another person answered yes to, say, item 3, it would follow that (s)he also answered yes to the items 1 and 2. Thus, under the Guttman model it must not happen that a particular person says yes to one particular item and no to another item that represents a lower introversion level.

The Guttman model is an extreme model for human behavior, because it assumes that behavior is completely predictable and perfectly consistent. Given that four ($J = 4$) items each with two answer options in principle can produce $2^J = 16$ different patterns of 0s and 1s, it is easily deduced that the Guttman model only allows $J + 1 = 5$ of these patterns: 0000, 1000, 1100, 1110, and 1111. This restriction means that given the number of yes answers, respondents can be located in intervals between two adjacent item location parameters. For example, if John answered yes three times, his θ value is located between δ_3 and δ_4 (Fig. 2). Because intervals can be ordered, the Guttman model defines an ordinal scale.

In contrast to what the Guttman model predicts, real data produced by a sample of respondents in principle contain all 16 item-score patterns. Probabilistic models can describe the frequencies in which such patterns are expected given the respondent attribute levels and the item parameters. If the expected frequencies are consistent with the observed frequencies obtained in a sample, the model fits the data and a scale can be calibrated. Significant deviations suggest misfit, but may also suggest possible ways of improving the instrument.

Before discussing probabilistic measurement models, I discuss two types of error that affect the quality of the measurement instrument. The first type is random measurement error. More random error impairs measurement precision. The second type is systematic error. A greater influence of this error type impairs measurement validity. I discuss how both errors can be controlled, but also notice that control is imperfect.

5. Two error types

5.1. Random measurement error and measurement precision

This error source reflects the random component in human behavior [11] that impairs the precision of an observable measurement value as an estimate of the person parameter θ . It is assumed that even under well-controlled measurement conditions, respondents give partly unpredictable responses, for example, due to variation in mood, concentration and attention, alertness due to their physical condition, and consistency of decision-making. Given this

variation, repeated measurement of the same person would produce a so-called propensity distribution of observable measurement values rather than one value [11, p. 30]. The smaller the variation of the measurement values, usually expressed in the standard error of the distribution, the more precise an observable measurement value. As these repeated measurements are impossible to obtain due to practice and memory effects, in practice only one measurement value for each respondent is available. Mellenbergh [12] discusses the two standard solutions to get a grip on measurement precision.

One solution is to use the data collected in the whole sample of respondents to estimate one standard error that is assumed to be useful for each respondent. This is the 'classical test theory' solution [11], which still is the approach that psychologists use most frequently to determine measurement precision. This popularity probably is due to the approach's simplicity even though it is at odds with the assumption that different persons may be measured with different precision. The other, more advanced solution is to statistically model the response process to the items such that a standard error is obtained that varies across different scale values, reflecting the amount of statistical information present in the item scores for different scale values. In psychometrics, the latter solution is considered superior.

Real tests and questionnaires do not have perfect measurement precision but the researcher can construct his/her instrument to have at least high precision by using two principles. First, a larger number of items usually increase measurement precision. Second, the item location parameters δ_j determine to a high degree where the scale measures precisely (i.e., with small standard error). For example, if a test uses a cut-score θ_0 to make a decision about passing or failing, the two principles jointly stipulate using many items with $\delta_j = \theta_0$. Item parameter estimates have to be obtained in prior research.

5.2. Systematic measurement error and validity

This error source reflects the problem that it is impossible to isolate a psychological attribute as the only source systematically influencing the responses to the items. In practice, responses to items often have multiple causes that are impossible to separate completely from the attribute of interest. A test or questionnaire that measures the intended attribute well is said to be valid [13]. The reduction of unwanted influences on the measurement process improves validity. Unlike measurement precision, validity is a controversial topic in psychometrics; for an anthology of different views, see [14].

Language skills provide an example of a disturbing influence that is active in nearly all psychological measurement [23]. In the examples I gave—verbally formulated introversion statements and arithmetic problems presented as little stories—language skills influence the response process and disturb the measurement of introversion and arithmetic ability. Some control over language skills can be realized by using simple words and sentences.

In maximum-performance measurement, badly chosen items may invoke cognitive skills different from the ones

that are really of interest, and thus pose another threat to validity. Recently, statistical models known as cognitive diagnosis models [24,25] have been proposed to study the skills active in solving particular problems, thus facilitating the identification of irrelevant skills and improving validity.

Response styles pose a threat to the validity of typical-behavior measurement. Examples are the tendency to avoid giving answers in the most extreme rating-scale categories [26] or tending to answer in the “safe” middle category [27]. Typically, these tendencies are independent of the item content. Another example is social desirability [28,29], which is the inclination to give answers the person expects to be acceptable for many people. For example, in response to the item “I will not say a word when I am among people I do not know very well” a person who indeed is silent and knows this but also is aware that this kind of behavior is not well accepted, might be tempted to answer no out of a desire to conform. Also, persons may be tempted to answer coherently, avoiding apparently inconsistent answers even if these inconsistent answers describe their typical behavior well [30,31]. These tendencies are sometimes measured using separate questionnaires, and the results are used to statistically correct the measures of interest.

In general, many of the disturbing influences can be controlled to some degree by carefully choosing items and preparing well-standardized testing conditions. Additional control is obtained by using a probabilistic measurement model. Such a model is based on assumptions that restrict the number of explanatory variables and exclude (or model) dependencies among items attributable to additional influences on responses. This is the topic of the next section.

6. Probabilistic measurement models

Probabilistic measurement models, such as item response models, overcome the problem of determinism typical of the Guttman model (e.g., [18]). I discuss several item response models and one model in particular, which is the simple yet popular Rasch model [32,33]. The discussion of the Rasch model illustrates several general principles of item response models. I also present the results from a data analysis using the Rasch model, and explain how the results lead to a calibrated scale for arithmetic ability [20].

6.1. Models

The Rasch model is defined for items with 0/1 scoring. For a latent variable θ and an item location parameter δ_j , the Rasch model defines the conditional probability of obtaining a 1 score as a continuous function of latent variable θ ,

$$P(X_j = 1|\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}. \quad (1)$$

This function is called the item response function for item j . Eq. (1) depends only on one item parameter, δ_j . As a result, the Rasch model is also called the 1-parameter logistic

model. Fig. 3 (left) shows item response functions for four items with different δ values. In contrast to the Guttman model (Fig. 2), in the Rasch model the function increases monotonically in θ . Let θ represent arithmetic ability, then the Rasch model assumes that as ability level increases, the probability of giving the correct answer also increases. Response probabilities are between 0 and 1, thus explicitly allowing for inconsistency. The location parameters correspond to scale values for which $P(X_j = 1|\theta) = .5$. Items located further to the right have higher δ values and response probabilities that are uniformly smaller for all θ values. Hence, the δ s are also interpreted as difficulty parameters.

Fig. 3 (right) shows four item response functions from the more general 2-parameter logistic model [18]. In this model, item response functions have different slopes at the inflexion point ($\delta_j, .5$). The parameter, which quantifies the slope in the inflexion point, is called the slope parameter or the discrimination parameter, and is denoted α_j . The item response function for the 2-parameter logistic model equals

$$P(X_j = 1|\theta) = \frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}. \quad (2)$$

The steeper the slope, the better the item separates relatively low θ values to the left of location δ_j from relatively high θ values to the right of location δ_j . The 3-parameter logistic model [18] is much used in educational measurement when guessing for the right answer, as with multiple-choice items, can be a problem. A third item parameter γ_j is added to the model in Eq. (2), which equals the probability that someone with an extremely low θ value gives the correct answer. The resulting item response function is

$$P(X_j = 1|\theta) = \gamma_j + \frac{(1 - \gamma_j) \exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}.$$

Several other models have been proposed for 0/1 scores. So-called explanatory item response models [34] may, for example, lay out the difficulty parameter δ_j into a sum of parameters v_m that represent the contributions of particular task features to the difficulty of the item, such that $\delta_j = \sum q_{jm} v_m$ (weight q_{jm} indicates, for example, whether task feature m is relevant to item j ; e.g., if so, $q_{jm} = 1$, else $q_{jm} = 0$). Inserting this sum in Eq. (1) yields the linear logistic test model [35],

$$P(X_j = 1|\theta) = \frac{\exp(\theta - \sum q_{jm} v_m)}{1 + \exp(\theta - \sum q_{jm} v_m)}.$$

For the transitive reasoning items in Fig. 1, I mentioned task features such as the logical relationships between the objects, the number of objects, and the mode of the problem. Explanatory item response models use a limited number of parameters that explain why, for example, tasks using equalities are easier than tasks using inequalities; and so on. Not only can such models produce calibrated scales but they also contribute to a better understanding of the response process; in addition see [24,25].

Several item response models have been proposed for rating-scale scores [18]. Such models are more complex than models for 0/1 scores because they define response

functions for each discrete item score, so that each item is characterized by several such functions. The linear common factor model, which is not an item response model, can be used for the analysis of continuous data [36], such as response times (i.e., the time it takes a respondent to solve a problem). Recently, Van der Linden [37] proposed an item response modeling approach to continuous response times.

In preference scaling, respondents are asked, for example, which beer brands they prefer with respect to bitterness. Beer brands are assumed to have a location on a bitterness scale ranging from mild to strong, and different respondents are assumed to have their own locations corresponding to their optimal preference. The closer one's location is to that of the beer brand, the higher the probability that one picks out that brand, and the further away one's location is on either side of the item location (the brand is either too bland or too bitter), the lower the probability. Preference, unfolding and ideal-point item response models for preference data thus have response functions that are single peaked [38].

Response functions usually are represented by parametric functions such as the logistic functions in the 1-, 2-, and 3-parameter logistic models, but it may be argued that parametric functions are unnecessarily restrictive and hamper the fit of a model to the data. Alternatively, nonparametric item response models only posit order restrictions on response functions while maintaining an ordinal person scale [39]. Nonparametric item response models thus are more flexible and fit more readily to data, and ordinal scales are often sufficient for the applications envisaged in psychology. See Post [40] for a nonparametric preference item response model.

Reckase [41] proposed multidimensional item response models that account for several attributes simultaneously influencing item responses. These models facilitate the inclusion of additional latent variables that may describe influences on test performance that are difficult to control, such as the previously mentioned language skills, response styles and social desirability. Latent class models can be used whenever the latent variable is discrete [42] rather than continuous. Discrete attributes are typically found in the clinical context, for example, when people can be classified into a group showing a pre-schizophrenic profile known as schizotypal personality disorder and a group that does not have this profile [43]. The first group is at risk of developing schizophrenia.

6.2. Estimation

I briefly explain parameter estimation for unidimensional item response models for 0/1 item scores, and the Rasch model in particular. Let $\mathbf{X}$ be the data matrix for the 0/1 item scores of N persons on J items (order $N \times J$). Let vector $\theta = (\theta_1, \dots, \theta_N)$ and let ω denote the vector containing all item parameters for a particular item response model. For example, for the Rasch model ω contains the J item location parameters and for the 2-parameter logistic model ω in addition contains the J discrimination parameters. The problem to be solved is which sets of parameters θ and ω most likely generated data matrix $\mathbf{X}$. In statistics,

this is the well-known maximum likelihood (ML) estimation problem for which several approaches have been proposed. Here, I discuss two of these approaches.

The likelihood of the data is denoted $L(\mathbf{X}|\theta, \omega)$. Item score x_{vj} denotes the 0/1 score of person v on item j . Scores of different persons are independently distributed. For example, different persons did not have any knowledge of one another's answers when providing their own answers. Also, item response models assume local independence, meaning the absence of additional influences on the responses to some items but not to others. Technically, local independence means that, given a fixed value of θ , the item scores are independent. For brevity, let $P_j(\theta) = P(X_j = 1|\theta)$. Under these assumptions, the likelihood equals

$$L(\mathbf{X}|\theta, \omega) = \prod_{v=1}^N \prod_{j=1}^J P_j(\theta_v)^{x_{vj}} [1 - P_j(\theta_v)]^{1-x_{vj}}.$$

Joint maximum likelihood estimation is a method that estimates the item and person parameters simultaneously, but it is known to result in inconsistent estimates; that is, the estimates do not approach the parameter values as sample size N grows [44]. Alternatively, marginal maximum likelihood (MML) estimates the item and person parameters in separate steps, and produces consistent estimates. MML works as follows.

In MML, the distribution of random variable Θ , denoted $f(\Theta)$, is often assumed normal with mean μ and variance σ^2 . MML is based on a likelihood, which is the average of $L(\mathbf{X}|\theta, \omega)$ across $f(\Theta)$ (the jargon is that Θ is integrated out of the likelihood), and which is defined as

$$L_M(\mathbf{X}|\omega, \mu, \sigma^2) = \prod_{v=1}^N \int_0^1 \prod_{j=1}^J P_j(\theta_v)^{x_{vj}} [1 - P_j(\theta_v)]^{1-x_{vj}} f(\Theta) d\Theta.$$

MML estimates or assumes parameters μ and σ^2 , and then maximizes the likelihood by estimating the item parameters in ω . The averaging of the likelihood across $f(\Theta)$ leaving only the item parameters is what produces the consistency property for the item parameter estimates. Let vector $\mathbf{x}_v = (x_{v1}, \dots, x_{vJ})$ contain the J item scores of person v , then the integral can be written as $P(\mathbf{x}_v|\omega, \mu, \sigma^2)$ and the likelihood as

$$L_M(\mathbf{X}|\omega, \mu, \sigma^2) = \prod_{v=1}^N P(\mathbf{x}_v|\omega, \mu, \sigma^2).$$

For example, given distribution parameters μ and σ^2 for the Rasch model the maximization of this function with respect to the item difficulty parameters in ω yields estimates that are the most likely given the data in $\mathbf{X}$ [45]. Once these estimates are available, in a second step Bayesian methods [44] are used to estimate the person parameters θ . The usefulness of the estimates depends on the fit of the Rasch model to the data in $\mathbf{X}$, which is the topic of the next subsection.

Because I focus on the Rasch model, I discuss another interesting ML method, which is only feasible for the Rasch model, being a member of the exponential family [35]. This is conditional maximum likelihood (CML) estimation. Typical of CML is that by conditioning on the total scores for

persons (i.e., the number of 1 scores on the test), one obtains equations that only contain the item parameters $\delta_1, \dots, \delta_J$ but not the person parameters θ . The item parameters are then estimated independently of θ and $f(\theta)$, which thus allows calibration independent of the particular group that took the test. Similarly, θ is estimated independent of $\delta_1, \dots, \delta_J$, and for a person different item sets thus produce the same measurement value. Hence, it is possible to disentangle the influence of the properties of the items and the attribute level of the tested person on the probability of giving correct answers.

Briefly, CML works as follows. Let $\xi = \exp(\theta)$ and $\varepsilon_j = \exp(-\delta_j)$, then Eq. (1) becomes

$$P_j(\xi) = \frac{\xi \varepsilon_j}{1 + \xi \varepsilon_j}.$$

Person parameter ξ has the same interpretation as θ , but ε_j is interpreted as item easiness (a higher value implies a higher response probability) rather than item difficulty. Let vectors $\xi = (\xi_1, \dots, \xi_N)$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_J)$. The total score for person v is the sum of his/her item scores, that is, $x_{v+} = \sum_{j=1}^J x_{vj}$. Similarly, the total score for item j is $x_{+j} = \sum_{v=1}^N x_{vj}$. The likelihood for the Rasch model can be written as

$$L(\mathbf{X}|\xi, \varepsilon) = \prod_{v=1}^N \prod_{j=1}^J P_j(\xi)^{x_{vj}} [1 - P_j(\xi)]^{1-x_{vj}} = \frac{\prod_{v=1}^N \xi^{x_{v+}} \prod_{j=1}^J \varepsilon_j^{x_{+j}}}{\prod_{v=1}^N \prod_{j=1}^J (1 + \xi_v \varepsilon_j)}.$$

From the right-hand side one can see that one does not need the complete data matrix $\mathbf{X}$ to estimate the model parameters but only the total scores for persons (x_{v+} , $v = 1, \dots, N$) and items (x_{+j} , $j = 1, \dots, J$). These total scores are *sufficient statistics* for the model parameters. Sufficiency is a feature of exponential models, and ascertains the estimates have certain desirable properties; see [46] for a discussion of these properties.

In the next step, the total scores for the N persons are collected in a vector $\mathbf{x}_{N+} = (x_{1+}, \dots, x_{N+})$, and the probability of the data matrix $\mathbf{X}$ given these person totals and the model parameters is considered, that is

$$P(\mathbf{X}|\mathbf{x}_{N+}; \xi, \varepsilon) = \frac{P(\mathbf{X}|\xi, \varepsilon)}{P(\mathbf{x}_{N+}|\xi, \varepsilon)}. \quad (3)$$

It can be shown [32,35] that Eq. (3) depends only on the item parameters ε and the sufficient statistics for ξ and ε but not on the person parameters ξ . The resulting equation is the conditional likelihood, which is solved for the item parameters ε . The CML item parameter estimates are consistent. Next, ML [47] is used to estimate ξ . ML estimation also yields the standard errors for the estimates of ξ . These standard errors are used to express the precision of the estimates and are scale-dependent. Thus, they provide the superior indicators for measurement precision that vary across the scale; see [12].

6.3. Goodness-of-fit research

The Rasch model predicts a particular structure in the data, and before estimated parameters can be interpreted, goodness-of-fit investigation must ascertain whether the model gives an adequate description of the data. Glas and Verhelst [48] provide a summary of goodness-of-fit methods. In the data example, I use the asymptotic χ^2 test statistics R_1 and R_2 . The R_1 statistic tests the null hypothesis that the J item response functions estimated from the data resemble parallel logistic functions as in Eq. (1). Rejection of the null hypothesis suggests that different item response functions have different slopes. The standard normal statistic U_j evaluates for each item whether the estimated slope is steeper than expected under the Rasch model (e.g., $U_j < -1.645$) or flatter (e.g., $U_j > 1.645$). Another way to assess slopes is to estimate the slope parameters α_j under the 2-parameter logistic model. The researcher may choose to delete deviant items from the test or to fit a more flexible model allowing varying slopes. The R_2 statistic tests whether local independence holds in the data. Rejection of the null hypothesis is taken as evidence of multidimensionality. The researcher may choose to split the test into subtests or to fit a multidimensional model [41] and provide persons with multiple scores corresponding to the different dimensions.

6.4. A Scale for arithmetic ability

The goal of this section is to show how the goodness-of-fit of the Rasch model to real data is investigated, and how the results are used for calibration. The example is only meant as an illustration. It does not result in a scale for use in real applications for measuring student's abilities.

The data were kindly made available by CITO National Institute of Educational Measurement (Arnhem, The Netherlands); also see [20]. The 13-item test measures arithmetic of proportions and ratios, using items like Item 5 mentioned previously: "A tower is 30 m high and casts a shadow 12 m long. The tree next to the tower casts a shadow 5 m long; how high is the tree? (Formal problem: $(30 : 12) \times 5 = ?$). A sample of 612 Dutch primary school students tried to solve the problems (0 = incorrect, 1 = correct). The Rasch model was used to calibrate a scale by means of CML estimation, using software package RSP [49]. First, I discuss the goodness-of-fit analysis, then the calibration of the scale, and finally I suggest directions for future research.

6.4.1. Goodness-of-fit investigation

The R_1 statistic indicated that the 13-item response functions were not parallel logistic curves ($R_1 = 64.20$, $df = 36$, $p = .003$) and that local independence did not hold ($R_2 = 128.07$, $df = 72$, $p = .000$). Table 1 shows that based on estimated relatively small slope parameters α_j ($< .73$), a subset of 8-items with nearly the same slopes could be selected ($0.90 < \alpha_j < 1.24$). For this 8-item subset, RSP analysis supported the hypothesis of equal slopes ($R_1 = 29.44$, $df = 21$, $p = .10$), which was corroborated by the U_j values ($|U_j| < 1.645$ for all j). Support was also found for local independence ($R_2 = 36.62$, $df = 24$, $p = .05$). Thus,

Table 1

Estimated Item Discrimination Parameters (α_j) and Standard Errors (SE) for 13 items, and Estimated Item Difficulty Parameters (δ_j), SE, and U_j Statistics for the final 8 items.

Item no.	Item text	α_j	SE	δ_j	SE	U_j
2		0.73	0.09	–	–	–
3	5:100 = 1:?	0.90	0.10	–0.38	0.10	0.36
4	60:40=?	0.92	0.10	–1.43	0.10	0.49
5	30:12 $\times$ 5=?	1.08	0.14	0.17	0.11	–0.41
6	2000:200 = 1500:?	1.00	0.14	–1.04	0.10	0.04
8	(15 $\times$ 800):(100:5)=?	1.24	0.16	0.27	0.11	–0.31
9		0.52	0.09	–	–	–
10		0.69	0.08	–	–	–
11	(4/3 $\times$ 60) $\times$ 4.5=?	1.11	0.17	1.11	0.13	–0.53
12	3:30=?	1.01	0.12	–0.44	0.10	–0.15
13	(5:2.5) $\times$ 100,000=?	1.01	0.15	1.74	0.15	0.39
14		0.72	0.09	–	–	–
15		0.52	0.08	–	–	–

I concluded that the Rasch model held for the 8-item subset. This result was used for scale calibration.

6.4.2. Calibrating the scale

The Rasch model does not fix the origin of the scale (adding a constant c to both θ and δ_j in Eq. (1) does not affect the response probability). This problem was routinely fixed during the estimation process by setting the sums of the estimated θ and δ values both equal to 0. Fig. 4 shows the calibrated scale for the 8 items. Item locations correspond to estimated δ s (Table 1). Item 4 is the easiest and Item 13 the most difficult. Based on the sufficient statistics (i.e., $X_+ = 0, 1, \dots, 8$), students can have one of 9 different estimated θ values, which are also displayed. For each θ value a standard error was estimated, which was used to estimate 80% confidence intervals expressing measurement precision as a function of the scale [12]. In Fig. 4, the interval for $\hat{\theta} = -0.03$ (standard error = 0.79, interval length = 2.03) is shorter than the interval for $\hat{\theta} = 2.02$ (standard error = 1.04, interval length = 2.67), because the

former $\hat{\theta}$ value better complies with the item difficulties than the latter (other results are not shown to keep the figure simple).

6.4.3. Practical use of the scale

The example served the purpose of illustrating how a scale is calibrated. In general, arithmetic scales contain more than 8 items so as to have higher precision. If the five deleted items had had higher discrimination than the 8 Rasch items I would have used the 2-parameter logistic model to calibrate the 13-item scale. However, the five deleted items had low discrimination and including them in the scale would only increase measurement precision marginally. A better research strategy is studying the causes why the 5 deleted items performed worse than the eight Rasch items, and using the resulting knowledge to construct and include items expected to have higher discrimination, so that a more precise scale results. Doing this is interesting but beyond the scope of this article. The specific application of the test determines the desired measurement precision. A diagnostic test, which is used to pinpoint difficulties students have with particular kinds of arithmetic problems, may use fewer items than a high-stakes test that is used for important pass-fail decisions with respect to an educational program.

6.5. Differences with physical measurement

Psychological scales do not have units comparable to meter, kelvin, joule, ohm, and becquerel. The cause for this absence is that psychology does not yet have theories about interesting attributes that are sufficiently precise to allow their experimental verification, justifying concatenation operations or other procedures logically leading to unit-based measurement [50]. Instead, the mathematical structure of an item response model defines the scale, and the goodness-of-fit of the model to the sample data implies that the scale can be used for the application at hand. For example, the Rasch model implies equal dis-

Fig. 4. Item locations (vertical lines), person locations (large bold dots), and 80% confidence intervals (horizontal lines printed above the scale) for $\hat{\theta} = -0.03$ and $\hat{\theta} = 2.02$.

tances between adjacent integer scale values but it does not fix the scale's origin. When estimating the parameters, setting the sums of the item and person parameters to 0 solves this problem. The units are expressed on a logit scale but there is no underlying theory that defines the units specifically for different attributes.

All of the above applies to the arithmetic scale I constructed: The fit of the Rasch model implies an equal-unit scale in terms of logits but without a meaningful zero point. If I had used different arithmetic items requiring different formal and cognitive operations for their solution, this might have produced another equal-unit scale but the units of the two scales might have been different. If different tests have items in common, this forms a basis for equating the units. The lack of an origin means that a measurement value does not represent an absolute arithmetic level. Instead, the content of the items is used to interpret test results. In psychology such scales prove useful for establishing the arithmetic level at which a student has arrived, for diagnosing the problems a student has and that might justify remedial teaching, and for determining whether a student should be admitted to a higher-level course.

Nonparametric item response models define ordinal scales, which not only better reflect the state of psychological theory development but also are sufficient for many applications. Examples are the selection of the highest-scoring applicant for a job and the selection of the 20 highest-scoring students for admittance to a specialized and expensive course. Interestingly, physics has had a profound influence on the thinking about psychological measurement [50,51] but the more primitive state of psychological attribute theories has moved psychological measurement in a different direction [50], as this article demonstrates.

7. Conclusions

Psychological measurement instruments suffer from problems that may be recognizable for measurement specialists in the exact sciences. Measurement precision—the degree to which measurements are repeatable under the same circumstances—and the construction of a calibrated scale are technical problems, which are mastered well. The validity problem of determining whether the instrument captures the psychological attribute of interest has raised much debate on preferred methodologies and philosophical viewpoints on psychological attributes [14].

The gap between psychometrics and the practice of test construction in psychology is noteworthy but I believe not uncommon in many other scientific areas. Theory development by definition is ahead of practical application and it may take some time for practitioners to catch up and give up on the older and more familiar methodologies. Nevertheless, the past few decades have shown a steady growth of the number of applications of item response models, and they may be expected to eventually replace the simpler and less effective psychometric methods such as classical test theory.

Three professional measurement organizations are the following. The Psychometric Society (<http://www.psychometrika.org/>) is an international nonprofit professional

organization devoted to the advancement of quantitative measurement practices in psychology, education, and the social sciences. The National Council on Measurement in Education (<http://www.ncme.org/>) is a nonprofit organization devoted to advance the science of measurement in the field of education. The International Test Commission (<http://www.intestcom.org/>) is an association of different organizations committed to promoting effective testing and assessment policies and to the proper development, evaluation and uses of educational and psychological instruments.

Acknowledgments

I thank Samantha Bouwmeester and Wilco H.M. Emons for providing the figures for this article, and Rob R. Meijer for his critical comments.

References

- [1] R.J. Sternberg, *Handbook of Intelligence*, Cambridge University Press, Cambridge, 2000.
- [2] G. Halford, An experimental test of Piaget's notions concerning the conservation of quantity in children, *J. Exp. Child Psychol.* 6 (1968) 33–43.
- [3] S. Bouwmeester, J.K. Vermunt, K. Sijtsma, Development and individual differences in transitive reasoning: a fuzzy trace theory approach, *Dev. Rev.* 27 (2007) 41–74.
- [4] B. De Raad, M. Perugini, *Big Five Assessment*, Hogrefe & Huber Publishers, Seattle, WA, 2002.
- [5] D.I. Ben-Tovim, M. K. Walker, The development of the Ben-Tovim Walker Body Attitudes Questionnaire (BAQ): A new measure of women's attitudes towards their own bodies, *Psychol. Med.* 21 (1991) 775–784.
- [6] J.M. Baldwin, J.M. Cattell, J. Jastrow, Physical and mental tests, *Psychol. Rev.* 5 (1898) 172–179.
- [7] J.McK. Cattell, Mental tests and measurements, *Mind* 15 (1890) 373–381.
- [8] F.Y. Edgeworth, The statistics of examinations, *J. Roy. Stat. Soc.* 51 (1888) 599–635.
- [9] A. Binet, Th.A. Simon, Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux, *Ann. Psychol.* 11 (1905) 191–244.
- [10] C. Spearman, "General intelligence", objectively determined and measured, *Am. J. Psychol.* 15 (1904) 201–293.
- [11] F.M. Lord, M.R. Novick, *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, MA, 1968.
- [12] G.J. Mellenbergh, Measurement precision in test score and item response models, *Psychol. Methods* 1 (1996) 293–299.
- [13] D. Borsboom, G.J. Mellenbergh, J. van Heerden, The concept of validity, *Psychol. Rev.* 111 (2004) 1061–1071.
- [14] R.W. Lissitz, *The Concept of Validity*, Information Age Publishing, Inc., Charlotte, NC, 2009.
- [15] H. Wainer, H.I. Braun, *Test Validity*, Erlbaum, Hillsdale, NJ, 1988.
- [16] C.J. Brainerd, V.F. Reyna, Fuzzy-trace theory and memory development, *Dev. Rev.* 24 (2004) 396–439.
- [17] S. Bouwmeester, K. Sijtsma, Measuring the ability of transitive reasoning, using product and strategy information, *Psychometrika* 69 (2004) 123–146.
- [18] W.J. Van der Linden, R.K. Hambleton, *Handbook of Modern Item Response Theory*, Springer-Verlag, New York, 1997.
- [19] P.T. Costa Jr., R.R. McCrae, Revised NEO Personality Inventory (NEO PI-RTM) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual, Psychological Assessment Resources, Odessa, FL, 1992.
- [20] K. Sijtsma, W.H.M. Emons, Statistical models for the development of psychological and educational tests, in: T. Rudas (Ed.), *Handbook of Probability: Theory and Applications*, Sage, Thousand Oaks, CA, 2008, pp. 257–275.
- [21] L. Guttman, A basis for scaling qualitative data, *Am. Sociol. Rev.* 9 (1944) 139–150.
- [22] L. Guttman, The basis for scalogram analysis, in: S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, J.A. Clausen (Eds.), *Measurement and Prediction*, Princeton University Press, Princeton NJ, 1950, pp. 60–90.

- [23] J.W. Berry, Y.H. Poortinga, M.H. Segall, P.R. Dasen, *Cross-Cultural Psychology: Research and Applications*, second ed., Cambridge University Press, Cambridge UK, 2002.
- [24] B.W. Junker, K. Sijtsma, Cognitive assessment models with few assumptions, and connections with nonparametric item response theory, *Appl. Psychol. Meas.* 25 (2001) 258–272.
- [25] J. Leighton, M. Gierl (Eds.), *Cognitive Diagnostic Assessment in Education: Theory and Applications*, Cambridge University Press, Cambridge UK, 2007.
- [26] E.A. Greenleaf, Measuring extreme response style, *Public Opin. Quart.* 56 (1992) 328–351.
- [27] G.F. Bishop, Experiments with the middle response alternative in survey questions, *Public Opin. Quart.* 51 (1987) 220–232.
- [28] A.C. Carle, Internal and external validity of scores on the Balanced Inventory of Desirable Responding and the Paulhus Deception Scales, *Educ. Psychol. Meas.* 67 (2007) 859–876.
- [29] R. Tourangeau, T. Yan, Sensitive questions in surveys, *Psychol. Bull.* 133 (2007) 859–883.
- [30] D.L. Paulhus, Two-component models of socially desirable responding, *J. Person. Soc. Psychol.* 46 (1984) 598–609.
- [31] D.L. Paulhus, Interpersonal and intrapsychic adaptiveness of trait self-enhancement: a mixed blessing?, *J. Person. Soc. Psychol.* 74 (1998) 1197–1208.
- [32] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, Nielsen & Lydiche, Copenhagen, 1960.
- [33] G.H. Fischer, I.W. Molenaar, *Rasch Models. Foundations, Recent Developments, and Applications*, Springer-Verlag, New York, 1995.
- [34] P. de Boeck, M. Wilson, *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*, Springer-Verlag, New York, 2004.
- [35] G.H. Fischer, *Einführung in die Theorie Psychologischer Tests (Introduction to the Theory of psychological Tests)*, Huber, Bern, Switzerland, 1974.
- [36] H.H. Harman, *Modern Factor Analysis*, University of Chicago Press, Chicago, IL, 1976.
- [37] W.J. van der Linden, A lognormal model for response times on test items, *J. Educ. Behav. Stat.* 31 (2006) 181–204.
- [38] J.P. Roberts, J.E. Laughlin, A unidimensional item response model for unfolding responses from a graded disagree-agree response scale, *Appl. Psychol. Meas.* 20 (1996) 231–255.
- [39] K. Sijtsma, I.W. Molenaar, *Introduction to Nonparametric Item Response Theory*, Sage Publications Inc., Thousand Oaks, CA, 2002.
- [40] W.J. Post, *Nonparametric Unfolding Models: A Latent Structure Approach*, DSWO Press, Leiden, The Netherlands, 1992.
- [41] M. Reckase, A linear logistic multidimensional model for dichotomous item response data, in: W.J. Van der Linden, R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*, Springer-Verlag, New York, 1997, pp. 271–286.
- [42] J.A. Hagenaars, A.L. McCutcheon, *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, UK, 2002.
- [43] N. Haslam, The dimensional view of personality disorders: a review of the taxometric evidence, *Clin. Psychol. Rev.* 23 (2003) 75–93.
- [44] F.B. Baker, S.-H. Kim, *Item response Theory. Parameter Estimation Techniques*, second ed., Marcel Dekker, New York, 2004.
- [45] D. Thissen, Marginal maximum likelihood estimation for the one-parameter logistic model, *Psychometrika* 47 (1982) 175–186.
- [46] I.W. Molenaar, Estimation of item parameters, in: G.H. Fischer, I.W. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments, and Applications*, Springer-Verlag, New York, 1995, pp. 39–51.
- [47] H. Hoijtink, A. Boomsma, On person parameter estimation in the dichotomous Rasch model, in: G.H. Fischer, I.W. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments, and Applications*, Springer-Verlag, New York, 1995, pp. 53–68.
- [48] C.A.W. Glas, N.D. Verhelst, Testing the Rasch model, in: G.H. Fischer, I.W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments, and Applications*, Springer-Verlag, New York, 1995, pp. 69–95.
- [49] C.A.W. Glas, J.L. Ellis, *User's Manual RSP: Rasch Scaling Program, iecProGAMMA*, Groningen, The Netherlands, 1993.
- [50] J. Michell, *Measurement in Psychology. A Critical History of a Methodological Concept*, Cambridge University Press, Cambridge, UK, 1999.
- [51] R.D. Luce, J.W. Tukey, Simultaneous conjoint measurement: a new type of fundamental measurement, *J. Math Psychol.* 1 (1964) 1–27.