Chapter 11
Ensuring Validity in International Comparisons Using State-of-the-Art Psychometric Methodologies

Matthias Von Davier, Eugenio Gonzalez, and Wolfram Schulz

Abstract Researchers using quantitative methods for describing data from observational or experimental studies often rely on mathematical models referred to as latent variable models. The goal is to provide quantities that allow generalization to future observations in the same subject domain. A review of selected current and historical examples illustrates the breadth and utility of the approach, ranging from a worldwide used system for ranking chess players, to finding hidden structure in genetic data, to identifying common factors that can explain patterns of volatility of assets examined in financial modeling. This chapter describes how latent variable models are used in educational measurement and psychometrics, and in the studies of the International Association for the Evaluation of Educational Achievement (IEA) in particular. Within this domain, these models are used to construct a validity argument by modeling individual and system level differences as these relate to performance on large-scale international comparative surveys of skills, such as those commissioned by IEA.

Keywords Ability estimation · Educational measurement · Latent regression · Latent variable models · Psychometrics · Validity argument

M. Von Davier (✉)
Lynch School of Education and Human Development, Campion Hall, Boston College, 140 Commonwealth Avenue, Chestnut Hll, MA, USA
e-mail: vondavim@bc.edu

E. Gonzalez
Educational Testing Service (ETS), 44 Leamington Rd, Princeton, Boston, NJ MA 02135, USA
e-mail: egonzalez@ets.org

W. Schulz
Australian Council for Educational Research (ACER), Camberwell, Australia
e-mail: wolfram.schulz@acer.org

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_11

11.1 Introduction

International surveys of student learning outcomes, such as those conducted under the auspices of IEA, rely on advanced statistical methodologies developed in fields commonly called psychometrics or educational measurement. The goal of these fields is to provide mathematically rigorous methods for quantifying skills and knowledge based on observable data, mainly responses to survey questions.

The responses of students on tests of reading literacy or mathematical knowledge, or students' responses to questionnaire items designed to measure students' attitudes toward learning at school provide these observables. In international comparisons, there is a strong emphasis on selecting only those types of observables that tap into the comparable types of skills and attitudes across countries. In addition to the mathematical rigor of these methods, which enables researchers to check whether the responses are indeed reflecting a common construct or latent trait, expert knowledge regarding content, learning, child development, and skill acquisition and skill decline are important areas that have to be taken into account when using statistical methods to compare the performance of educational systems and their contexts across the world.

The analytic techniques used in IEA assessments follow best practices and use latent variable models developed over several decades. Their presentation in this chapter does not delve into the depths of the mathematical formalism, but uses equations and mathematical expressions whenever, in our judgement, a purely verbal description of important concepts would not be sufficient to provide an accurate representation.

In this chapter, we do not cover the full breadth of models used in educational measurement, but focus only on those psychometric models that were further developed and adapted for the application to data from international assessments of student learning outcomes.

The foundation of the approaches presented here are methods that aim at deriving quantitative information with regards to latent traits of interest (cognitive or non-cognitive) based on how respondents answer test or questionnaire items that are designed to assess the desired constructs. These constructs can be so-called cognitive constructs, such as reading literacy, or scientific literacy or mathematical skills, or non-cognitive constructs, such as mathematics self-efficacy, perceptions of classroom climate, or attitudes toward learning in the case of questionnaire-type surveys of attitudes or perceptions.

Typically, some foundational assumptions are made that enable quantitative measures to be derived from responses of students given to test or questionnaire items. These central assumptions can be summarized as follows:

(1) Each response can be scored to reflect a degree of correctness (in the dichotomous case as either correct or incorrect) or the degree of appraisal on a rating scale, which in a way reflects the amount of the constructs that are to be measured;

(2) Responses to all items are associated with a defined attitude or skill, i.e., the responses to the items are a non-random reflection or manifestation of the variable of interest, and not explainable by other (nuisance) variables; and
(3) The same variable of interest underlies the responses to the items, and affects the responses in the same way, across participating countries.

This chapter explains why researchers using quantitative methods for describing data from observational or experimental studies tend to rely on mathematical models that are usually referred to as latent variable models. A brief review provides a few current and historical examples in order to illustrate the breadth and utility of the approach, ranging from a worldwide used system for ranking chess players, to finding hidden structure in biological data, to identifying common factors that can explain patterns of volatility of assets examined in financial modeling. The examples are organized by complexity rather than chronologically, enabling latent variable models to be considered as a solution to a problem science faces on a regular basis: scientists aim to describe what is observable by relating observations to broader and more general principles that can be considered as underpinning the phenomena. We end this chapter with a more detailed description of how latent variable models are used in educational measurement and psychometrics. Within this domain, we focus on how these models are used for modeling individual and system level differences in performance on large-scale international comparative surveys of learning outcomes such as IEA's Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), International Civic and Citizenship Education Study (ICCS), and International Computer and Information Literacy Study (ICILS), but also, as these are using rather similar approaches, surveys such as the Organisation for Economic Co-operation and Development's (OECD's) Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC).

A variable is called "latent" if it cannot be directly observed, although it may be possible to observe many of the effects, or at least correlates, of the hypothesized quantity on other related variables. While some may argue in favor of summarizing observed data without a reference to latent variables, there are striking reasons to use these types of models with variables. The assumption of "latent" variables (common in social and educational sciences) often raises concerns about assuming the existence of variables that cannot be measured directly. However, here we provide examples from natural sciences and economics to illustrate that many other disciplines facing similar problems use latent variable models. These are not only convenient but often turn out to be central to the solution of a scientific problem.

A simple example may clarify this: any person who has tried to calculate an average has in fact made an attempt to obtain a proxy to a latent variable. Whether the average is taken using test scores, ratings, or school grades, or whether batting averages in cricket, baseball, or softball are considered, the aim is the same: to obtain a more meaningful and stable summary of the underlying mechanism or cause (the average is also called a measure of central tendency) of outcomes in a certain domain.

At the same time most would agree that an average, while more stable than an individual score, is not to be confused with "true ability" (whatever that may be), as summary representations may be affected by particular circumstances at the time of measurement. A marathon runner will know a few stories to tell about which location, which time of year, and which temperature are more beneficial for faster finishing times. Usually this is not taken as evidence that the fastest runner in the world became consistently slower only because they happened to undertake the past few runs in extreme heat, in pouring rain, or at high elevations, conditions under which (almost) every runner is slower. Rather, people take task difficulty into account and assume that observed performance may vary based on these characteristics, while the underlying ability to run a fast marathon is virtually unchanged under constant and controlled conditions. This is the central tenet of latent variable models: different underlying mechanisms are assumed to affect observed performance. Some represent the agent's ability, some task difficulty, and potentially, some others represent additional situational characteristics such as time on task, or test taking motivation or engagement (see e.g., Rose et al. 2017; Ulitzsch et al. 2019).

Lewin (1939) described the dependency of observed behavior on personal characteristics and contexts in the following simple, semi-formal equation:

$$B = F(P, E)$$

Here, behaviors (B) are a function of person characteristics (P) as well as environmental factors (E). While it is uncertain whether this expression was already based on an understanding of the concept of broad behavioral tendencies in terms of latent variables, it appears that both the person P and the environment E are considered variables, and the behavior is considered a dependent variable, i.e., a different person P' may or may not show the same behavior B or B' when doing this influenced by the same environmental variable(s) E. Also, the same person P may show the same or different behavior if confronted with different environmental factors E'.

Interestingly, many latent variable models can be expressed in essentially the way Lewin (1939) described the dependency of observed behaviors on person variables and variables defining the situation in which a behavior is shown. The basic aim is to describe person attributes in a way that generalizes across items (environments) and, at the same time, to describe items in meaningful ways that allow comparisons and predictions about whether a certain person confronted with a certain environment is likely to behave in a certain way.

In Sects. 11.2 and 11.3, we describe how this basic aim was implemented in areas such as rating players in competitive games (chess, Call of Duty®, etc.), ordering or classifying biological observations according to their most likely genetic makeup, and finally describing factors that are responsible for group differences in skills, abilities, or affective variables measured in large-scale educational assessments. While on the surface chess player rating and quantitative genetics do not seem to have much in common with describing what students know and can do, we show that the methods used in these applications share a lot with what is done in educational

measurement. To illustrate this we use examples that share underlying assumptions about mechanisms operationalized as latent variables. Another common element is using controls for situational factors operationalized as variables, such as opponents, problem types, environments, time dependencies, or political events.

In Sect. 11.4, we put all these components together and show how gradual differences as well as systematic differences between groups, and the identification of common factors underlying a large number of observed variables are used in educational large-scale assessments. While these methods are broadly applied in educational measurement and other areas of data- and evidence-driven research, we focus on the quantitative methodologies as used in educational measurement in the context of international large-scale assessments (ILSAs) such as TIMSS, PIRLS, ICCS, ICILS, PISA, or PIAAC.

11.2 Modern Educational Measurement: Item Response Theory

Documented evidence of measuring human performance goes back at least to the second century BC in ancient China (e.g., Franke 1960). Humans have competed in sports and games for millennia (Murray 1913, 1952). Competitive, rule-based games such as chess, go, backgammon, and the ancient game of *hnefatafl* (Murray 1913, 1952) are most interesting when players are well matched, and opponents provide a challenge to the other player. In order to match players, a good match needs to be defined. One common approach is to assume that player strength can be quantified objectively, independent of the tasks or opponents this strength is measured against. Just as in the measurement of physical characteristics such as height and weight, the result of measurement should be (largely) independent of the scale or standard used to obtain the measure.

In the case of chess, rankings of players are well established. Elo (1978) came up with a method to adjust chess rankings of a player quasi on the fly, after each match, based on whether it had resulted in a win, loss, or a tie. The method Elo (1978) devised is an elegant way to provide players with an estimate of their strength (their ELO score) based only on the games they played. The mathematical foundations of this approach are based on what researchers describe as pairwise comparisons (Bradley and Terry 1952; Luce 1959). Under this paradigm, pairs of objects are compared, once or more than once, and either an observer declares a preference for one of the objects, or a rule-based system determines which of the objects has a higher ranking.

11.2.1 From Chess Ranking to the Rasch Model

The mathematical foundations of Elo's (1978) approach go back to Zermelo (1929), who developed a method for estimating player strength in chess tournaments where not all players compete against all others. In this approach, the probability of player A winning over player B is modeled as:

$$P(X = 1|\theta_A, \theta_B) = \frac{\exp(\theta_A - \theta_B)}{1 + \exp(\theta_A - \theta_B)}$$

where a data point $X = 1$ indicates a win of the first player over the second, and θ_A and θ_B denote the strength of players A and B, respectively. Zermelo (1929)[1] wrote the equation originally as:

$$P = \frac{w_A}{w_A + w_B}$$

where W_A and W_B denote the numbers of wins for players A and B respectively. This equation is mathematically equivalent to the above by setting $w_A = \exp(\theta_A)$. This form of the equation may seem more intuitive to some as it relates directly to repeated trials and winning and losing of players. If only two players are considered, and they compete 30 times, and player A wins 20 out of 30 while player B wins 10, the probability of player A winning becomes:

$$P_{AB} = \frac{20}{20 + 10} = \frac{2}{3}$$

However, player B is also matched with player C, where player C wins 15 times and player B wins 15 times, resulting in:

$$P_{BC} = \frac{15}{15 + 15} = \frac{1}{2}.$$

Intuitively, it could be inferred that players B and C are equally strong and it could also be assumed that in a match between A and C, the chances of A winning would be again 2/3. However, this does not need to be the case. This additional assumption may not be altogether solid, as C could be better at certain chess-related things (openings, for example) than B, while B could be better at endgames. Therefore, on average, the players may appear of equal strength, even though A is weaker than B at openings

[1] Economists tend to be familiar with Zermelo's name in the context of early game theory (Zermelo 1913; Schwalbe and Walker 2001) while mathematicians often know about Zermelo as one of the first scholars to formulate fundamental set theoretic results. His chess ranking model was independently rediscovered by Bradley and Terry (1952) and others. However, the estimation methods originally presented by Zermelo (1929) can be considered a more efficient maximum likelihood approach than that proposed by later authors who rediscovered the method independently (S.J. Haberman, personal communication 2016).

but extremely strong at endgames. In this case we would have an "endgame skill" as well as an "opening skill", and would need to devise two separate competitions to disentangle these skills with fidelity. Very often, however, chess players will be stronger or weaker than others in different kinds of chess-related situations. For example, a chess grandmaster tends to be excellent at openings, mid- and endgames, and will likely beat a novice in almost any match.

Related to the question whether we indeed measure just a single chess skill (as in assumption 2 in Sect. 11.1): how can we make sure we assess skills in a way that we can indeed generalize from seeing A and B playing as well as B and C to a hypothetical situation where A is confronted with a new task, say playing against C?

Much of the remainder of this chapter is concerned with how educational measurement is applied in ILSAs and builds models that allow these types of generalizations. Educational assessments assess individual students solving problems, for example when they answer questions about a text, or solve a mathematical problem. Models such as the ones introduced in Sects. 11.2–11.4 aim at deriving variables that allow these types of generalizations. In other words, these models aim at constructing proficiency measures that generalize beyond the specific set of test questions students see in educational assessments and permit more general statements about student's learning outcomes in different subject domains (such as mathematical or reading literacy) or other learning-related perceptions.

There is a striking similarity between the equations used in Zermelo's and Elo's chess ranking system and the model equation of the Rasch model (Rasch 1960; von Davier 2016). Georg Rasch was a Danish mathematician whose influence on educational measurement cannot be overstated (Olsen 2003). Many scholars have used the Rasch model as the basis for a variety of developments with broad impact on psychometric and educational measurement (see Fischer and Molenaar 1995; von Davier and Carstensen 2007). The Rasch model for dichotomous item responses is given by

$$P(X = 1|\theta_A, \beta_i) = \frac{\exp(\theta_A - \beta_i)}{1 + \exp(\theta_A - \beta_i)}$$

and θ_A, as before, denotes a strength, a skill, an ability or more broadly an attribute of person A, while β_i denotes the characteristic or difficulty of a task, indexed by $i = 1, \ldots I$. These tasks may be are a series of chess problems (as in: "checkmate in three moves") or a mathematics item ("solve for x: $3x + 8 = 20$" etc.) on a test, or some other exercise to check motor functions, or candidates selected by voters (Poole 2005). Fischer (1981) used the results provided by Zermelo (1929) to prove uniqueness and existence of maximum likelihood estimators of the Rasch model, and pointed out that the Rasch model is indeed a special case of the Zermelo (1929) model where two distinct sets Ω_θ and Ω_β of objects are always combined in a pairwise manner, while two objects from the same sets are never compared directly. More specifically, in the Zermelo approach all players can in principle be matched against all other players, the Rasch model assumes that human agents (test takers, respondents) are always paired with problems (tasks, test items) and, so to speak,

compete against tasks but not against each other. It is interesting to note that, apart from this particular feature, the two approaches are mathematically identical.

11.2.2 Characteristics of the Rasch Model

The Rasch model is one of the most successful item response theory models (IRT) (Lord and Novick 1968) and has been used for both large-scale international survey assessments as well as school-based and state assessments around the world. While there are more general models such as the two-parameter logistic (2PL) and three-parameter logistic (3PL) IRT models (Lord and Novick 1968), the Rasch model is considered one of the most elegant approaches as it has several mathematical properties that other models do not provide to the same extent (von Davier 2016). Among the applications of the Rasch model are the operational analyses and the reporting of proficiency scores from the initial reports from IEA's TIMSS 1995 grade 8 results, and those from PISA from 2000 until 2012, as well as those from IEA's ICCS and ICILS. While the Rasch model can be characterized as one of the most elegant approaches for measuring, it can be shown that it does not predict the observables as well as some more general models (see Sect. 11.2.3). TIMSS 1999 started using a more general approach, as did PIRLS, and PISA finally started using a more general model in 2015.

The reasons for this are best understood when comparing a test that assesses a multitude of topics in a variety of ways with the introductory chess example. A test, even one that a teacher may produce as a quick assessment of the students, contains different types of questions, requiring different types of answers. TIMSS and PIRLS items all assess a common domain (science, mathematics, reading literacy), but do so in a variety of ways. Chess matches are (to some extent) essentially always driven by the same objective: to achieve a checkmate (i.e., to capture the opposing side's king). Therefore, a model such as the Rasch model, which was originally developed to provide measures based on tests with extremely similar item materials, may need to be revised and extended in order to be suitable for broad, encompassing assessments such as TIMSS and PIRLS.

The Rasch model is not only one of the central tools of psychometrics and educational measurement but also an approach which is either reinvented frequently or highlighted as being particularly useful for complex situations which aim at deriving quantitative information from binary trials. The complexity often arises from the need to provide measures that enable comparisons even in situations where not all students are assessed on all tasks in a subject domain (which resembles the case of Zermelo's chess ranking model where chess players cannot play against all other players).

In many situations, measurement in education contexts cannot exhaustively assess all respondents on all possible tasks. Nevertheless, the aim is to make generalizable statements about the extent to which the task domain was mastered overall, and at what level this was the case. Ideally, the comparison between respondents should

not depend on what instrument they have been assessed with, just as the comparison of two chess players based on their ELO score should be independent of which opponents they faced in the past and, indeed, also independent of whether they ever played against each other.

In the Rasch model, this can be seen by calculating the logarithm of the odds (often referred to as log-odds) using the previous model equation. This provides

$$LO(\theta_A, \beta_i) = \log\left[\frac{P(X = 1|\theta_A, \beta_i)}{P(X = 0|\theta_A, \beta_i)}\right] = \theta_A - \beta_i$$

in the Rasch model. Further, when we calculate the difference

$$LO(\theta_A, \beta_i) - LO(\theta_B, \beta_i) = \theta_A - \theta_B$$

to compare any two respondents A and B, this difference turns out to be independent of which item β_i was used in the comparison. In order to compare any two tasks i, j we can use

$$LO\big(\theta_A, \beta_j\big) - LO(\theta_A, \beta_i) = \beta_i - \beta_j$$

which results in the same difference independently of which respondent was assessed.

In terms of practical comparisons, all respondents with the same total score receive the same estimate of the underlying latent trait θ under the Rasch model. In the situation of an educational test, for example, the probability of getting item i correct for respondent A can be estimated based on the relative frequency of getting the item correct based on the total group of respondents that has the same total score as respondent A. While the estimation methods used to generate optimal estimators are somewhat more sophisticated (Rasch 1960; Andersen 1970; von Davier 2016), comparisons can also be carried out across raw score groups and it is possible to calculate approximate differences (von Davier 2016) based on these conditional success probabilities in homogeneous score groups.

11.2.3 More General IRT Models

While the Rasch model can be considered the gold standard in terms of the ability to directly compare test takers independent of items, as well as items independently of the samples used to estimate item characteristics, this model puts rather strong constraints on how ability and the probability of successful task completion are related (von Davier 2016). There are more general IRT models that relax these assumptions and allow more flexibility when modeling the relationships between ability and task success. Among these, the models proposed by Birnbaum (1968) are the most commonly used ones. Since the 1999 cycle of the study, TIMSS has used the 3PL model for the multiple choice items, as does PIRLS.

As alternatives to the more constrained Rasch model, the 2PL and 3PL models are defined as

$$P_i(X = 1|\theta) = P(X = 1|\theta, a_i, b_i) = \frac{1}{1 + \exp(-a_i[\theta - b_i])}$$

and

$$P_i(X = 1|\theta) = P(X = 1|\theta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{1}{1 + \exp(-a_i[\theta - b_i])}$$

where c_i is considered a guessing parameter (and assumed to be $c_i = 0$ in the Rasch model and the 2PL model) and $a_i > 0$ is a slope parameter and, as before, θ, b_i are ability and difficulty parameters, respectively. The 3PL model can be applied for multiple-choice test items in order to take the guessing probability.

It is important to note that all three models, the Rasch model, and the 2PL and 3PL models, assume a single ability variable θ underlying a measured learning outcome, and that with increasing ability, the probability of successful task completion (in the case of educational testing) increases. Similarly, both for the Rasch and 2PL model, with increasing endorsement of a particular affective dimension there is also a higher likelihood of providing higher ratings (in the case of measuring non-cognitive outcomes, for example when using items with response categories reflecting levels of agreement or disagreement). The respective item parameters (such as a_i, b_i, c_i) do not change this fundamental relationship, and it can be shown that all three models lead to rather similar results (Molenaar 1997). This is not a surprise as research has shown that all three unidimensional IRT models are mathematically equivalent to certain types of unidimensional item factor models (Takane and DeLeeuw 1987).

Perhaps the most important property of IRT, however, is the ability to provide estimates in circumstances in which not all respondents are administered all items, but are assessed with different subsets of tasks, which all measure the same subject domain. This is particularly important in the context of large-scale assessments, since these not only attempt to assess relatively broad domains with large numbers or items but also renew the sets of tasks used for a variety of reasons, including the need to cover new content while maintaining a strong connection to previous rounds of the assessments. In Sect. 11.2.4, we review two additional assumptions that are made when deriving customary IRT models that are very useful in this context.

As an example, TIMSS, like all other current IEA studies assessing student achievement, uses a design in which each student only receives a subset of the tasks. This allows the administration of many items to each of the participating countries in order to broadly cover the subject domains measured by TIMSS, without overburdening the students with endless testing sessions. In TIMSS 2015, there were a total of 28 blocks of items: 14 blocks of mathematics items and 14 blocks of science items (Fig. 11.1). Each booklet contained only two blocks of test items for each of the two domains, giving a total of four blocks per booklet. Each block appears in two different booklets, each time in a different location. This balances the exposure of

	Assessment Blocks			
Student Achievement Booklet	Part 1		Part 2	
Booklet 1	M01	M02	S01	S02
Booklet 2	S02	S03	M02	M03
Booklet 3	M03	M04	S03	S04
Booklet 4	S04	S05	M04	M05
Booklet 5	M05	M06	S05	S06
Booklet 6	S06	S07	M06	M07
Booklet 7	M07	M08	S07	S08
Booklet 8	S08	S09	M08	M09
Booklet 9	M09	M10	S09	S10
Booklet 10	S10	S11	M10	M11
Booklet 11	M11	M12	S11	S12
Booklet 12	S12	S13	M12	M13
Booklet 13	M13	M14	S13	S14
Booklet 14	S14	S01	M14	M01

Fig. 11.1 TIMSS 2015 student achievement booklet design at grades 4 and 8. *Notes* M indicates a block of mathematics items, S indicates a block of science items. *Source* Martin et al. (2013, p. 91). Copyright © 2013 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College

the items to the test takers and, at the same time, by having the blocks overlap with the others, allows to make inferences on items that were not taken by the students.

Just as in the chess example, using the Rasch model or more general IRT models allows these types of inferences if the same skill is measured by all the items within the domain. TIMSS and PIRLS employ a sophisticated set of statistical tools to check whether the items are indeed measuring the same skills across blocks, across positions in the assessment booklets, as well as over time and across participating populations.

11.2.4 Central Assumptions of IRT and Their Importance

IRT models are often referred to as item-level models that describe the probability of a correct response, given examinees proficiency θ and some item-specific parameters (such as a_i, b_i). However, this is not how IRT models are actually applied. Initially the item parameters and the proficiency θ are unknown, and all that analysts can rely on

is a series of answers to not one, but often 10, 20, or more items. What is needed is a model for many responses, and one that makes assumptions that enable inferences to be made about the unknown parameters given the observed responses. Three central assumptions underlying IRT modeling are unidimensionality, local independence, and measurement invariance.

11.2.5 Unidimensionality

We now assume that there are several items, and we denote the number of these items with I and the response variables with $X = (X_1, \ldots, X_I)$. Assuming unidimensionality means that a single quantity, the latent variable of interest, is sufficient to describe the probabilities of responses on each of the items, and that this is the same quantity regardless of the items, for a given person v.

So, for example, let P_{iv} and P_{jv} denote the probability of person v scoring 1 for items i and j, then, if unidimensionality holds, this can be re-expressed as

$$P_{iv} = P_i(X = 1|\theta_v)$$

and

$$P_{iv} = P_j(X = 1|\theta_v)$$

with θ_v being a real number.

Roughly, unidimensionality ensures that the same variable is measured with all the test items. This of course holds only if the assessment development aims at producing a set of items that indeed are designed to assess the same dimension. Unidimensionality would (very likely) not hold, for example, if half of the items in a skills test consisted of multiplication problems, and the other were assessing gross motor skills such as shooting a penalty kick, two seemingly unrelated skills.

In TIMSS and PIRLS, it is assumed that all items measure the same skill, or latent variable; in TIMSS this holds for mathematics and science separately, of course. At the same time, these assessments also allow the quantification of subscale or subdomain skills. This is a concept that relates to the fact that complex tests have multiple facets or topics that can be collected into distinct groups of items that tap into similar contexts or contents (Rijmen et al. 2014). While statistically there is good reason to report across these subscales using an overall mathematics, science, or reading score, there are situations where researchers may wish to analyze how groups of examinees perform in certain subdomains. One reason may be that these subdomains may "act" like they appear to be a single domain within each country while they also show distinct differences in performance across countries. This may be due to curricular differences across countries, and these differences can be studied in terms of their effect on subdomains. While here space limitations prevent a detailed explanation of this approach, the ideas have been developed and outlined in the TIMSS and PIRLS

technical reports (available for download on the IEA website; www.iea.nl). Further reading on the topic can be found in Verhelst (2012) and Feinberg and von Davier (2020).

11.2.6 Local Independence

The assumption of local independence states that the joint probability of observing a series of responses, given an examinees' proficiency level θ, can be written as the product of the item level probabilities, that is:

$$P(\mathrm{X} = x_1, \ldots, x_I|\theta) = \prod_{i=1}^{I} P_i(X = 1|\theta)^{x_i}[1 - P_i(X = 1|\theta)]^{1-x_i}$$

In particular, for any two items i and j, it can be assumed that

$$P\big(X_i = x_i \wedge X_J = x_j|\theta\big) = P(X_i = x_i|\theta)P\big(X_J = x_j|\theta\big)$$

While this assumption appears to be a rather technical one, it can be made more plausible by the following consideration. As already mentioned, the proficiency variable we intend to measure is not directly observable; it is only possible to observe behaviors that we assume relate to this proficiency, for example by means of the assumptions made in the IRT models. The assumption of local independence facilitates these inferences, in that it is assumed that once a respondent's proficiency is accounted for, all responses are independent from each other. That is, for example, knowing whether a respondent taking a test has correctly answered the previous question does not help predict their next response, assuming the respondent's true proficiency is already known.

This assumption can, of course, be inadequately supported by the data and the items in the assessment. While, in TIMSS, most of the items are independent units, there are cases where multiple questions are associated with a single item stem. Assessments of reading often contain multiple questions that relate to a single reading passage. This means that the assumption of local independence is potentially threatened by such a set of nested questions, but experienced item developers work to reduce these effects by making sure each question relates to an independent unit of knowledge or part of the text. In addition, statistical tools such as residual analysis and scoring approaches can alleviate the effect (e.g., Verhelst and Verstralen 1997).

The local independence assumption can not only be applied to the dependency of one item's responses on other items on the same test but also to other types of variables. Based on this more general understanding of the local independence assumption, if the model is correct, no other variables are helpful when predicting the next answer of a respondent, either given next on the test, or in three weeks' time. Only the underlying proficiency is mainly "responsible" for the probability of giving

correct item responses. In this sense, local independence is the assumption that it is only necessary to make inferences about the underlying cause of the behavior, not other observables and how they relate to test responses. If local independence holds, the latent variable provides all available information about the proficiency domain. It turns out that this expanded understanding of local independence is related to another assumption that is discussed next.

11.2.7 Population Homogeneity/ Measurement Invariance

One last central, but often only implicit assumption of IRT models is that the same relationships between item location (and other parameters), the respondents' latent trait, and item responses hold for all respondents. This seems to be a trivial assumption, but it turns out to be a centerpiece, one that cannot easily be ignored, particularly for the comparison of learning outcomes across countries or other types of population specifications. If the association between response probabilities and the underlying latent trait changes across groups, there would be no way to ensure that differences in their responses reflect actual differences in the measured learning outcomes, and not any other factors.

Formally, this assumption can be expressed as follows: if there are two respondents to a test, v and w, with $\theta_v = \theta_w$, denoting that both have the same proficiency, and $g(v) \neq g(w)$ indicating that the two respondents belong to different groups as defined by a grouping variable (such as gender, ethnicity, or country of residence), then population homogeneity holds if for these, and any two respondents v, w with $\theta_v = \theta_w$, we have

$$P_i(X = 1|\theta_v, g(v)) = P_i(X = 1|\theta_v) = P_i(X = 1|\theta_w) = P_i(X = 1|\theta_v, g(w))$$

In other words, the response probabilities for these two respondents depend only on their proficiency levels $\theta_v = \theta_w$ and item parameters, but not on the grouping variable $g(\cdot)$ or their group membership.

The assumption of population homogeneity means that response probabilities are fully determined by item characteristics and the measured latent trait. This assumption is central to the possibility of comparing learning outcomes across members of different groups. Note that this assumption does not say anything about the distribution of the measured latent trait in the different groups. It may well be that the average performance of respondents differs across groups. This is a seemingly subtle difference: the probability of getting the item right might be (on average, across all members) low in one group and high in another group. However, if we pick two test takers from each group, and it turns out that both test takers have the same ability level, then their chances of getting the same item right are the same, independently of which group they belong to.

While we do not know the "true" latent traits of respondents, we can assume that for each population, proficiencies are distributed according to some statistical

probability distribution, and with the assumption of population homogeneity we can estimate the total probability of responses based on this distribution alone, without considering any other variables. While it is not the only option, it is customary to assume a normal (or Gaussian) distribution. For a population of respondents, it is possible to assume that

$$\theta \sim \phi(\theta; \mu.\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{[S-\mu]^2}{\sigma^2}\right)$$

where μ is the mean of the distribution of the measured latent trait while σ denotes its standard deviation. Further, we may also assume more complex distributions, namely mixtures of distributions that consist of multiple Gaussians. In order to produce a statistical model for all the responses on an assessment, it is now straightforward to calculate the marginal probability as:

$$P(x_1, \ldots, x_I) = \int_{\theta} \prod_{i=1}^{I} P_i(X = 1|\theta)^{x_i}[1 - P_i(X = 1|\theta)]^{1-x_i}\phi(\theta; \mu, \sigma)d\theta$$

which is the standard marginalized form for models that contain unobserved variables, such as the latent trait measured in an IRT model. Among other things, this marginal probability takes center stage in the actual estimation of item parameters (such as a_i, b_i, c_i as well as the parameters of the proficiency distributions, μ and σ in this case). This expression is also important in the evaluation of how well this model (which could be considered an arbitrary choice) actually succeeds in predicting the observed responses of test takers within and across countries.

11.3 Simultaneous Modeling of Individual and Group Differences

Sums of many randomly varying quantities tend to approximately follow a normal distribution (e.g., Feller 1968); this is referred to as the central limit theorem. The normal distribution is probably the first and most important distribution taught in introductory statistics classes, and is fully described by only two parameters, a mean and a variance.

If, for example, a respondent to a test attempts many tasks that carry the same success probability, say $p = 0.5$, the summed number of successes can be well approximated by the normal distribution once there are more than 50 attempts, and with more certainty once more than 100 tasks have been attempted (Hays 1981). As an example, let $x = 1$ denote a success, and $x = 0$ an unsuccessful attempt, then for 50 independent attempts there is an expected value of $S = \sum x_i \sim 25$ successful attempts, and a variance of $p(1-p) \times 50 = 12.5$. While the probability distribution of the total number of success and failure can be described exactly by the

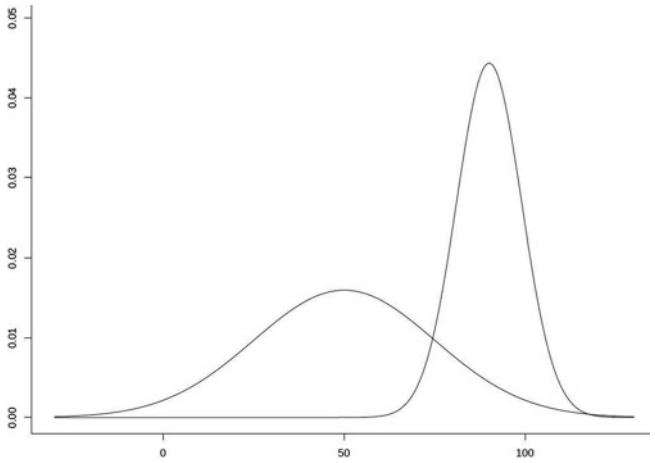

Fig. 11.2 Two normal distributions with means $\mu_1 = 50 = 0.5 \times 100$ and $\mu_2 = 90 = 0.9 \times 100$ and variances $\sigma_1^2 = 25 = 0.5 \times 0.5 \times 100$ and $\sigma_2^2 = 9 = 0.1 \times 0.9 \times 100$, respectively. It can be seen that for p = 0.9 a normal distribution does not provide a good approximation for the binomial with 100 trials, as substantial mass of the density is located at values larger than 100, while for p = 0.5 most of the mass of the approximation is located between 0 and 100

binomial distribution (e.g., Feller 1968; Hays 1981), the normal distribution provides a reasonable approximation by using

$$P(S) = \frac{f(S)}{Z} \text{and} f(S) \approx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{S-\mu}{\sigma}\right]^2\right)$$

with $\mu = 25$ and $\sigma = 12.5$ and $Z = \sum_{Q=0}^{50} f(Q)$ as a normalizing constant.[2] Consider two normal distributions (more accurately, normal densities) with expected values 30 and 90, and variances 21 and nine, respectively (see Fig. 11.2). These may correspond to two players who attempt gaming tasks (say playing against a chess program) with a constant success probability of 0.3 (weak player) and 0.9 (strong player), respectively.

However, there are many important cases where a normal distribution is not appropriate, even though many individual measures are averaged. One example was examined by Karl Pearson, who published, among many other things, several important pieces on a mathematical theory of evolution. Hence, Pearson (1894) is often also credited with having invented the mixture of normal distributions, a discovery that McLachlan and Peel (2000) traced back to Newcomb (1886). The relevance of this approach can be best understood by looking at an example. While Pearson (1894) was rather concerned with distributions related to the size of organisms (crabs, to be

[2] Strictly, the constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ is not needed in the equation above if the normal is used to approximate a discrete distribution in the way described here.

specific) and how these depend on not directly observed genetic factors, education provides also examples of unobserved causes of differences resulting in distributions that are not necessarily symmetric.

Pearson (1894) studied a case in which an observed distribution is a composition (mixture) or two (or more) components, while each of these components can be thought of as a normal distribution with a component-specific mean and variance. Formally, if $F(\theta)$ denotes the distribution of a random variable θ and $f(\theta)$ the associated density, then the simplest case of a discrete mixture distribution can be written as

$$f(\theta) = p_1 \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left[\frac{\theta - \mu_1}{\sigma_1}\right]^2\right) + p_2 \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left[\frac{\theta - \mu_2}{\sigma_2}\right]^2\right)$$

The above equation writes the marginal distribution of a random variable as consisting of two normally distributed components with different means and standard deviations, here μ_1, σ_1 and μ_2, σ_2 that are specific to each component. Schork et al. (1996) reviewed the use of its general approach in human genetics research and pointed out that discrete mixture distributions are:

> … intuitive modelling devices for the effects of underlying genes on quantitative phenotypic (i.e. trait) expression. In addition, mixture distributions are now used routinely to model or accommodate the genetic heterogeneity thought to underlie many human diseases. Specific applications of mixture distribution models in contemporary human genetics research are, in fact, too numerous to count.

One of many examples from educational testing is the analysis of spoken English proficiency among medical students taking the United States Medical Licensing Examination Step 2 Clinical Skills exam described by Raymond et al. (2009), where the authors stated that the distribution of scores of more than 29,000 test takers did not appear to be well fitted by a normal distribution (see Fig. 11.3, which shows 3000 randomly drawn data points from such a distribution).

Even without a formal test it appears obvious that the histogram (Fig. 11.3) is not well approximated by a normal distribution. However, when splitting the sample into test takers with English as their first language, versus test takers who report English as their second language, each subsample can be well approximated by a normal distribution (similar to that seen in Fig. 11.2).

The differences observed using only the performance data may lead to the conclusion that there are disproportionally many respondents with very low skill levels relative to the average performance in the population. However, knowing that there are respondents with different language backgrounds may lead to a different conclusion, namely that there are group differences that should be considered. The examples (Figs. 11.2 and 11.3) imply very different realities. One starts from the notion of different populations with different skill distributions (Fig. 11.2), but this is not immediately evident in the second example (Fig. 11.3), which is based on the assumption that all test takers are random samples from the same distribution without considering the difference between test takers attending international versus

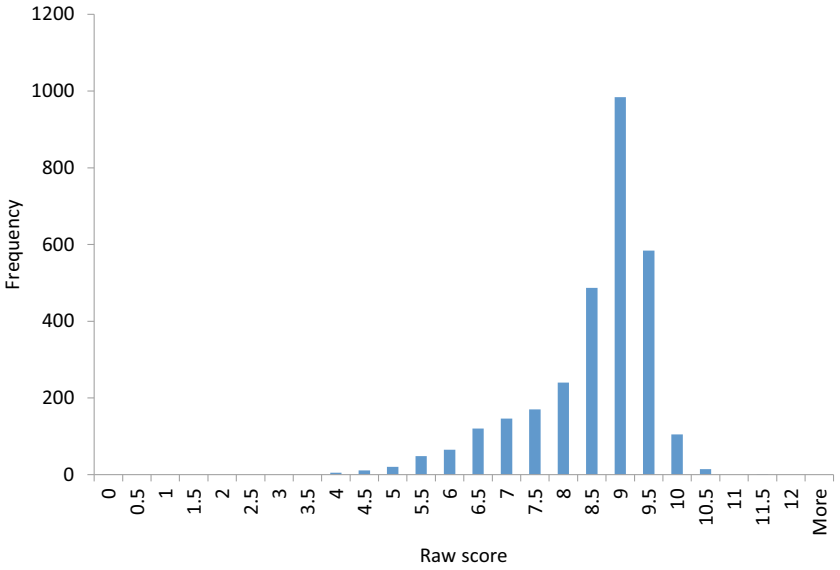

Fig. 11.3 Distribution of spoken language test scores obtained from 29,000 medical students taking the United States Medical Licensing Examination Step 2 Clinical Skills exam. The graph shows 3000 randomly drawn data points based on the distribution described by Raymond et al. (2009)

national schools when producing the graph that represents these as two separate but overlapping distributions.

Obviously, these types of group differences can be frequently encountered in populations that are composites of several subpopulations, for example those defined by regional, ethnic, or language differences. In TIMSS and PIRLS, these types of subpopulations can lead to important differences with respect to average performance as well as variability of skills in participating countries. Addressing this issue, and finding ways to incorporate the fact that group differences need to be assessed accurately has led to the types of multiple and mixture distribution models discussed in this section.

More generally, it is important to consider the existence of multiple populations, and that each population takes a translated and/or adapted version of the assessment in its source version, and that the outcome of the assessment depends on the underlying latent trait. This leads to several questions that we address in Sect. 11.4: how can we properly check whether the same latent trait is measured in all populations, how can we account for differences between populations with respect to trait distributions, and how can we examine whether the trait we are measuring is a valid representation of the construct we are interested in?

11.4 Statistical Modeling of Individual and Group Differences in IEA Survey Data

The famous statistician George Box is credited with the saying that "all [statistical] models are wrong, but some are useful," an insight which Box and coauthors have referred to in a variety of places (e.g., Box et al. 1978). This statement reflects the observation that statistical models can never include all possible evidence, and that even the quantities that are included may not entirely describe the way they truly interact. Statistical models can only be considered simplified formalisms that aim at providing useful summaries of observed data.

Statistical models used in ILSAs are certainly no exception. While similar models are used in national survey assessments (e.g., von Davier et al. 2007), the specific features that distinguish ILSAs are related to the added complexity due to the analyses of data from multiple countries surveying respondents with the same instruments in multiple languages (e.g., von Davier and Sinharay 2014; von Davier et al. 2009). This increased level of complexity is reflected in the following list of desiderata for statistical modeling:

(1) To account for and measure the extent to which groups differ on variables educational researchers are interested in;
(2) To specify how performance on a range of test items relate to underlying proficiency variables;
(3) To examine whether the assumed relationship between proficiency and observed responses is the same across countries;
(4) To integrate different sources of information such as responses and background information to facilitate reporting; and
(5) To provide variables for a database aimed at secondary analyses of proficiency data.

We will discuss how these desiderata are addressed in the approach used in large-scale assessments in the following subsections. International studies of educational outcomes translate these into a modeling approach in providing outcome variables that are comparable across countries and that facilitate secondary analyses as well as primary reporting. These reports aim at comparing the association of outcomes with policy-relevant contextual variables collected using student, teacher, parent, school and/or national context questionnaires.

11.4.1 Comparability as Generalized Measurement Invariance

When evaluating statistical modeling approaches for multiple populations, it is important to recognize that estimates of central quantities in the model, such as item parameters, as well as the estimates of distributional parameters, such as means and

standard deviations, may differ across populations. While distributional differences across populations are expected, and a focus of interest in terms of cross-country variance requires explanation, item parameter differences are undesirable with regard to the comparability of outcomes, and therefore should ideally be eliminated or at least reduced to negligible levels.

To illustrate this, we present a simple example with only three test items and two populations. We use a very much simplified IRT model with only two ability levels $\theta \in \{-1, +1\}$ with probabilities $P(\theta = -1) = 0.7 = 1 - P(\theta = +1)$ and one single binary item that differs with respect to how well it measures the ability variable θ in two groups A and B. This can be regarded as an item response model with a discrete latent variable (e.g., Follman 1988; Formann 1992; Haberman et al. 2008; Heinen 1996; Macready and Dayton 1977; von Davier 2005). Models of this type are studied in the context of mastery modeling, and more recent variants of these models are commonly referred to as diagnostic classification models (e.g., von Davier and Lee 2019).

Let us assume that the probabilities of a correct response in group A are given by $P_A(X = 1|\theta = -1) = 0.15$, and $P_A(X = 1|\theta = +1) = 0.85$, and in group B by $P_B(X = 1|\theta = -1) = 0.3$, and $P_B(X = 1|\theta = +1) = 0.70$.

The question is, what inferences can be drawn from observing a test taker who solves the single item on this test? We know that the test taker has an ability level of $\theta = +1$ with probability 0.3 and $\theta = -1$ with a probability of 0.7. Let us assume that the test taker succeeds in solving the item. Then we can apply Bayes theorem (Bayes 1763), one of the most fundamental equations in probability theory:

$$P_A(\theta = +1|X = 1) = \frac{P_A(X = 1|\theta = +1)P(\theta = +1)}{P_A(X = 1|\theta = +1)P(\theta = +1) + P_A(X = 1|\theta = -1)P(\theta = -1)}$$

Applying this theorem to our example then yields

$$P_A(\theta = +1|X = 1) = \frac{0.3 \times 0.85}{0.3 \times 0.85 + 0.7 * 0.15} \approx 0.71.$$

This can be translated to an estimated ability by using the expected-a posteriori (EAP) value, and in this simple mastery model, we obtain $\theta_{EAP|A} = -1 * 0.29 + 1 * 0.71 = 0.42$.

This means that our prior knowledge that mastery, $\theta = +1$, is only observed in 30% of population A is updated through the observation of one correct response by a test taker from population A and leads to a posterior probability of this test taker being in the group of masters of 71%. It appears that observing a correct response changes the likely mastery state of test takers in population A considerably. When looking at the same calculation in population B we obtain

$$P_B(\theta = +1|X = 1) = \frac{0.7 \times 0.3}{0.3 \times 0.7 + 0.7 \times 0.3} = 0.5.$$

As an EAP estimate we obtain $\theta_{EAP|B} = -1 * 0.5 + 1 * 0.5 = 0$ for population B. The change from identical prior knowledge (prevalence of masters with $\theta = +1$ is 30% also in population B) to posterior likelihood of mastery state is not quite as large among test takers in population B and we only would expect with a probability of 50% that they belong to the group that masters the material.

By drawing inferences about how likely it is that a test taker responds in a certain way, this constructed example illustrates that differences between groups in terms of item functioning can have profound consequences. In this example, our estimated probability that a test taker masters the material by solving an item decreases from 71 to 50% when comparing group B with group A. The reason is of course that the probabilities of correct responses on this item are very different for groups A and B.

What we implicitly assumed to be "the same item", on closer inspection, turns out not to function in the same way across the two comparison groups. Obviously, this was caused by the different probabilities of correct responses in the ability levels within groups, 0.7 versus 0.85 for respondents mastering the tasks and 0.3 versus 0.15 for those who fail to do so. The difference between those who fail and those who master the material is much larger in group A than in group B and, hence, if Bayes' theorem is applied, the information gained from observing a correct response in group A leads to a different adjustment of the posterior probability than in group B.

Therefore, one central commonly stated requirement is that items should work the same across all groups that are to be compared on the basis of a particular set of items. This is equivalent to the assumption of population homogeneity introduced earlier. Similar assumptions have been introduced in applied statistical modeling; this is referred to as measurement invariance (e.g., Meredith 1993; Millsap and Meredith 2007).

In IEA assessments, a careful review of items is conducted to ensure observed variables have the same relationship to the variable of interest, the skill and attitude variables that are intended goal of the assessment. The reports available for TIMSS and PIRLS, for example, allow insight into how translation accuracy is ensured, how items are reviewed, and how statistical tools are used to ensure invariance of the items across participating countries.

Even after items have been reviewed for appropriateness by all countries, and a translation into all languages of the assessment has been conducted, the quality control is not finished. As a hypothetical example, if an item was found to be impossible to solve in one specific country at the time of the first field testing of the item, this would trigger a series of investigations into, for example, the quality of translation, the review by countries, or the scoring guide. As a result, an item may be revised or eliminated, or the scoring guide adjusted for the affected country in order to ensure that the item that was found to violate invariance is either eliminated or changed in order to ensure integrity of the measurement.

11.4.2 Multiple-Group IRT Models

The idea behind IRT models is sometimes misunderstood as implying that the way items function, namely the way in which the measured trait determines item response probabilities, is independent of the groups that are being assessed. This is a prevalent misconception that appears to be based on a confusion of the concepts of a "population" versus a "sample (from a population)". It is sometimes said that IRT provides "sample-free" (e.g., Wright 1968) estimates (which is of course not true, a sample of observations is needed to estimate parameters). IRT (and in particular the Rasch model) are known to provide parameters that are (in expectation) invariant under sampling from the same population. Rasch (1966) spoke of specific objective comparisons: item difficulties can be compared independent of the respondents that were sampled to estimate the parameters.

In Sect. 11.3, we discussed how these types of misconceptions find their way into practice by implicit assumptions as a case of population homogeneity or measurement invariance. There is absolutely nothing in the models discussed here that will prevent parameters from varying from population to population. Population invariance is a feature of a test or an item that content experts have to work towards. That is why IEA, among many other quality control measures for curriculum-referenced studies such as TIMSS or PIRLS, performs curriculum coverage analyses to ensure that all items that become part of TIMSS or PIRLS have been vetted as appropriate for the student populations that are being tested. Statistics and psychometrics cannot enact population invariance, but rather they provide tools to test for invariance or approximate measurement invariance (e.g., Glas and Jehangir 2014; Glas and Verhelst 1995; Muthén and Asparouhov 2014; Oliveri and von Davier 2011; Yamamoto and Mazzeo 1992).

A customary approach to checking whether item parameters can be assumed to be invariant is estimation of multiple population versions of the statistical model under consideration. In IRT, these types of models have come to be known as multigroup IRT models (e.g., Bock and Zimowski 1997; von Davier 2016; von Davier and Yamamoto 2004). The basic assumption is that there is a finite number of populations denoted by $g \in \{g_1, \ldots, g_G\}$ and that the probabilities of correct response $P_{ig}(X = 1|\theta) = P_i(X = 1|\theta, g) = P\big(X = 1|a_{ig}, b_{ig}, c_{ig}, \theta\big)$ *may* depend on the group g as well as the ability variable. The same applies to the group specific ability distributions that can be mathematically described as:

$$\phi_g(\theta) = \phi(\theta|g) = \phi\big(\theta; \mu_g, \sigma_g\big).$$

While the model allows for deviations across multiple groups, ideally item parameters should be equal (invariant) across groups, or at least very similar (approximate invariance) in order to compare the latent trait estimates, so that situations like the one illustrated in the example in Sect. 11.4.1 do not occur. Note that, as already pointed out, $P_i(X = 1|\theta)$ should only depend on θ (which reflects the same latent trait across groups) and not on any other variables in order to meet assumptions of

population homogeneity or (strict) measurement invariance. This means the item response should only depend on the skill measured, not on the language version of the item, or the country where the item is administered. While this may be easier in chess and mathematics, as solving a system of linear equations is the same task no matter which language was used to ask the student to do this, it should also be possible to ensure in science and reading. Asking for the central agent in a reading passage should be possible, and should lead to the same response depending only on reading skills in the language of administration, and not on other variables such as cultural context.

The assumption of invariance would entail $a_{ig} = a_i, b_{ig} = b_i, c_{ig} = c_i$ for all population groups g. In terms of international studies, this would mean that the goal is to aim for the same shape of item functions across countries. If the items have the same item functions in all countries, the targeted skill has the same relationship to how the item is likely answered across participants from different countries.

Note that in cases of data collections where respondents come from multiple countries, the fact that each respondent was sampled in their country of (current) residence can be used to augment the data. Instead of only obtaining responses to items, we now have at our disposal the combined data, item responses x_i and group (country) membership g,

$$d_n^{IG} = \left[x_{n1}, \ldots, x_{nI}, g(n)\right]$$

where $g(n)$ represents the group (country) in which test taker n was tested. To estimate the marginal probability of the responses in the test takers group $g = g(n)$, we obtain

$$P(x_{n1}, \ldots, x_{nI}|g(n)) = \int_{\theta} \prod_{i=1}^{I} P_{ig(n)}(X = 1|\theta)^{x_{ni}} \left[1 - P_{ig(n)}(X = 1|\theta)\right]^{1-x_{ni}} \phi_g(\theta) d\theta$$

which, when assuming that all respondents complete their assessment independently can be used to define the likelihood of the data of all respondents as

$$P\left(d_1^{IG} \ldots d_N^{IG}\right) = \prod_{n=1}^{N} P(x_{n1}, \ldots, x_{nI}|g(n))$$

This is the basis for estimating the parameters μ_g, σ_g but also a_{ig}, b_{ig}, c_{ig}, typically starting by assuming that all item parameters are the same across groups, $a_{ig} = a_i, b_{ig} = b_i, c_{ig} = c_i$ for all g. While this is only a starting point, there exist elaborate procedures that allow items (and sub-groups or countries) to be identified for which this assumption is not met (more details about these procedures can be found in Glas and Jehangir 2014; Glas and Verhelst 1995; Oliveri and von Davier 2011; von Davier and von Davier 2007; Xu and von Davier 2006; Yamamoto and Mazzeo 1992).

Multiple-group IRT models are also used to link across cycles of international and national assessments where link items are included for the measurement of change

over time (Yamamoto and Mazzeo 1992) or where assessments are linked across different delivery modes of paper-based versus computer-based delivery (e.g., von Davier et al. 2019). The important feature of multiple-group IRT models in this context is the capacity to identify where there are item-by-country interactions that indicate that it is not possible to assume the same parameter across all countries or sub-groups. Technical reports on scaling assessments, such as those available online for TIMSS and PIRLS, show the linkage design both in terms of graphical displays and the number of link items involved. Link items may be used over two or more assessment cycles so that several data collections can indicate whether the items change over time or their retain measurement properties over multiple assessments (von Davier et al. 2019).

As a result of such analyses, some programs discard those items that do not meet the assumption of measurement invariance. Such an approach has the disadvantage that items are no longer used even though they may provide valuable information to increase the reliability of estimates for subgroups within countries. The fact that across countries these non-invariant items do not contribute to the comparability of scale means does not make them useless, so discarding items seems a rather extreme measure. However, there are more sophisticated approaches. It is also possible to maintain one set of item parameters that is used with common item parameters for all countries or sub-groups, while allowing parameters for some items to deviate in some countries or sub-groups. Such a procedure leads to a partial invariance model that maximizes the fit of the statistical model to the observed data, while maintaining the largest possible number of common parameters that meet criteria of measurement invariance (more details about a practical application of the approach can be found in von Davier et al. 2019).

11.4.3 Population Models Integrating Test and Background Data

Data collections in ILSAs tend to be fairly comprehensive. Aside from information about which items respondents completed in the assessment and how they scored, data from ILSAs also contain many additional variables about contextual variables collected in background questionnaires (from students, teachers, parents, and/or schools). These background data provide a rich context that allows secondary analysts to explore how students from different background perform on the assessment. Background variables further augment what is known about students participating in an assessment. We can write the complete data as

$$d_n^{IGB} = (x_{n1}, \ldots, x_{nI}, g_n, z_{n1}, \ldots, z_{nB})$$

where $z_{n1}, \ldots, z_{nB}$ represent the responses given by test taker n to the questions on the background questionnaire, g_n is the group membership (country of testing),

and $x_{n1}, \ldots, x_{nI}$ are the responses to the cognitive test items. The background data may contain additional variables from other sources (e.g., from school principal and teacher questionnaires) but, for simplicity, here we assume that we only make use of respondents' self-reports.

The background data can be assumed to be statistically associated with how respondents complete an assessment. What is measured by the assessment is quantified through the latent trait variable, θ, and readers are reminded that one of the central assumptions made in the previous sections was that the probability of observed successful performance is only related to θ, and no other variable. However, it was assumed that the distribution of this ability may depend on group membership and/or other (background) variables.

The population model consequently follows this line of reasoning by building a second level of modeling that predicts the expected value μ_n of the proficiency θ_n as a function of background variables $z_{n1}, \ldots, z_{nB}$:

$$\mu_n = \sum_{b=1}^{B} \beta_{g(n)b} z_{nb} + \beta_{g(n)0}$$

Furthermore, the proficiency variable is assumed as a normally distributed latent trait around this expected value, namely

$$\theta_n \sim N(\mu_n, \sigma)$$

Together, this provides a model for the expected proficiency given the background data $z_{n1}, \ldots, z_{nB}$. In other words, the expectation is that the distribution of proficiency depends on the background data used in the model. Such an assumption was mentioned in Sect. 11.3, when illustrating possible differences in learning outcomes between native speakers and second language speakers, and we also already mentioned the assumption of group-specific (e.g., across countries) latent trait means μ_g and standard deviations σ_g. However, the sheer amount of background data is much larger than the number of countries typically participating in an assessment. Therefore, if background variables are selected in such a way that suggests correlations with ability, it can be expected that the distribution around this expected value of $\sum_{b=1}^{B} \beta_{g(n)b} z_{nb} + \beta_{g(n)0}$ is somewhat narrower than a country-level distribution of abilities.

Formally, this is a multiple (latent) regression model that regresses the measured latent trait on background data collected in context questionnaires. The estimation problem is addressed separately within countries, since it cannot be assumed that background data have the same regression effects across different national contexts. Mothers' highest level of education, for example, is well known as a strong predictor of student performance, but this association can be moderated by other factors at the level of educational systems, so that in some countries it may be stronger than in others.

There are several ways to address the estimation of the regression parameters. In IEA assessments and other ILSAs, the latent trait is determined by an IRT model estimated across countries. Then the (latent) regression model is estimated using the item parameters obtained in the IRT estimation as fixed quantities. This ensures that the invariance properties that were determined through IRT estimation across countries will be applied equally to each national dataset (see, e.g., Mislevy et al. 1992; Thomas 1993; von Davier and Sinharay 2014; von Davier et al. 2007).

11.4.4 Group Ability Distributions and Plausible Values

The goal of the psychometric approaches described above is to produce a useful database that contains useful and reliable information for secondary users of the assessment data. This information comes in the form of multiple imputations or plausible values (PVs; see Mislevy 1991; Mislevy et al. 1992) of latent trait values for all respondents given all the responses to the assessment, as well as the knowledge about how respondents completed questions in the background questionnaire. Integrating the IRT model described in the first part of this chapter with the regression model introduced in the previous section, we can estimate the probability of the responses, conditional on background data, as

$$P_g(\mathbf{x}_n|\mathbf{z}_n) = \int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni}|\theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta$$

This equation provides the basis for the imputation of plausible proficiency values for each respondent. To allow a more compact notation, we use $P_{ig}(x_{ni}|\theta) = P_{ig}(X = 1|\theta)^{x_{ni}}\left[1 - P_{ig}(X = 1|\theta)\right]^{1-x_{ni}}$

This model allows making inferences about the posterior distribution of the latent trait θ, given both the assessment items $x_i \ldots x_I$ and the background data $z_1 \ldots z_B$.

This Posterior Distribution Can Be Written as

$$P_g(\theta|\mathbf{x}_n, \mathbf{z}_n) = \frac{\prod_{i=1}^{I} P_{ig}(x_{ni}|\theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right)}{\int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni}|\theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta}$$

The posterior distribution provides an estimate of where the a respondent n is most likely located on the measured latent trait, for example by means of

$$E_g(\theta|\mathbf{x}_n, \mathbf{z}_n) = \int_\theta \theta P_g(\theta|\mathbf{x}_n, \mathbf{z}_n) d\theta$$

as well as the posterior variance, which provides a measure of uncertainty around this mean:

$$V_g(\theta|\mathbf{x}_n, \mathbf{z}_n) = E_g\left(\theta^2|\mathbf{x}_n, \mathbf{z}_n\right) - \left[E_g(\theta|\mathbf{x}_n, \mathbf{z}_n)\right]^2.$$

These two quantities allow PVs to be established (Mislevy 1991), quantities that characterize each respondent by means of a sample of imputed values representing their location on the latent trait measured by the assessment. PVs are the basis for group level comparisons and they contain information not only about the respondents' latent trait but also their group membership, such as a country or education system, as well as responses given to background questions (which may include attitude and interest scales, and self-reports about socioeconomic home background, such as books at home, parents' education, or parents' occupation).

PVs are random draws

$$\tilde{\theta}_{ng} \sim N\left(E_g(\theta|\mathbf{x}_n, \mathbf{z}_n), \sqrt{V_g(\theta|\mathbf{x}_n, \mathbf{z}_n)}\right)$$

that depend on response data x_n as well as background data z_n and group membership g, which in international assessments often relates to the country or education system where the respondent was assessed. That means two respondents with the same item scores but different background data will receive a different predicted distribution of their corresponding latent trait. This, on the surface, may appear incoherent when not considering the underlying statistical properties. The reason for the need to include all available (context or background) information into the model used for generating the PVs can be best understood when looking at the research on imputation methods (e.g., Little and Rubin 1987). The latent ability variable is not observed for any respondent, and must be inferred by imputation. When leaving out important predictors of ability, this imputation will lead to biased estimates of abilities as the relationship between abilities and context or background factors is ignored: Von Davier et al. (2009) illustrated this phenomenon in a simulation study that is modeled after large-scale assessments used by IEA and other organizations.

All available data needs to be included to derive quantities that allow unbiased comparisons of population distributions (e.g., Little and Rubin 1987; Mislevy 1991; Mislevy et al. 1992; von Davier et al. 2009). Importantly, PVs should never be referred to, used, or treated as individual assessment scores, because the term score commonly implies that the quantity depends only on the test performance of the individual respondent, and not on contextual data.

11.5 Conclusions

By design, although we only provide a cursory treatment of the psychometric methods underlying the scaling of large-scale assessment data as used when reporting statistics and building research databases for secondary analyses, we have illustrated the general principles and foundations of measurement used in ILSAs. Note that these methods have been developed in order to tackle the problem of estimating latent traits based on observed, but incomplete data, and that the goal is to provide quantities that allow generalization to future observations in the same subject domain.

Most of the major international student and adult skill assessments use methods that are closely related to the approaches presented here (von Davier and Sinharay 2014). The general principles used in international assessments also apply to many national assessments, however, national evaluations usually lack the complexity introduced by assessing multiple populations and in multiple languages. IRT models, the measurement modeling approach most commonly used, are also widely applied in high stakes testing, school-based testing, certification, licensure testing, and psychological and intelligence testing.

The latent regression model used to generate PVs is best understood as a general-purpose operational imputation approach that enables the integration of multiple sources of information while using computationally efficient preprocessing methods. At a conceptual level, the population model is best understood as a complex imputation model that allows complete data to be generated under certain conditional independence assumptions using incomplete data collection designs.

It is worth noting that while IRT was developed and has traditionally been applied when testing cognitive aspects of learning, it is also increasingly used to scale data derived from questionnaires, in particular in the context of international studies such as TIMSS, PIRLS, PISA, ICCS, and ICILS (see Martin et al. 2016, 2017; OECD 2017; Schulz and Friedman 2015, 2018). When applied to questionnaire data, IRT tends also to be used for analyzing measurement invariance (see Schulz and Fraillon 2011), often in combination with other analytical approaches, such as multiple-group analyses (see Schulz 2009, 2017).

Further information on the general modeling approach used to integrate background data, item responses, and invariance information to generate proficiency distributions for large-scale educational surveys can be found in von Davier and Sinharay (2014) and von Davier et al. (2007), and Rutkowski et al. (2014) described many aspects of the methodological approaches used in these studies. An accessible introduction to why grouping and background data is needed for unbiased estimates of proficiency distribution can be found in von Davier et al. (2009).

References

Andersen, E. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society Series B, 32,* 283–301.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society, 53*, 370–418. http://doi.org/10.1098/rstl.1763.0053

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis, and model building*. New York, NY: John Wiley & Sons Inc.

Bradley, R. A., & Terry, M. E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika, 39,* 324–45.

Elo, A. E. (1978). *The rating of chess players, past and present*. New York: Arco Publishing.

Feinberg, R., & von Davier, M. (2020). Conditional subscore reporting using iterated discrete convolutions. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998620911933.

Feller, W. (1968). *An introduction to probability theory and its applications, Volume 1* (3rd ed.) New York, NY: John Wiley & Sons, Inc.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46*(1), 59–77. https://doi.org/10.1007/BF02293919.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53,* 553–562. https://doi.org/10.1007/BF02294407.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87,* 476–486.

Franke, W. (1960). *The reform and abolition of the traditional Chinese examination system*. Harvard East Asian Monographs, Volume 10. Boston, MA: Harvard University Asian Center.

Glas, C. A. W., & Jehangir, K. (2014). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 97–115). New York, NY: Springer.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69–95). New York, NY: Springer-Verlag.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions*. ETS Research Report RR-08-45. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x.

Hays, W. L. (1981). *Statistics for the social sciences* (3rd ed.). New York, NY: Holt, Rinehart and Winston.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Advanced Quantitative Techniques in the Social Sciences, Volume 6. Thousand Oaks, CA: Sage Publications.

Lewin, K. (1939). Field theory and experiment in social psychology: Concept and methods. *American Journal of Sociology, 44*(6), 868–896. https://doi.org/10.1086/218177.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons Ltd.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: John Wiley & Sons Ltd.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*(2), 99–120.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2013). TIMSS 2015 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2015 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. http://timssandpirls.bc.edu/timss2015/frameworks.html.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1–15.312). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., Fishbein, B., & Liu, J. (2017). Creating and interpreting the PIRLS 2016 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 14.1–14.106). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-14.html.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics. New York, NY: John Wiley & Sons, Inc. http://dx.doi.org/10.1002/0471721182

Meredith, W. M. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and new directions* (pp. 131–152). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161.

Molenaar, W. (1997). Lenient or strict application of IRT with an eye on practical consequences. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 38–49). Münster, Germany/New York, NY: Waxmann Verlag.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5,* 978. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00978/full.

Murray, H. J. R. (1913). *A history of chess*. Oxford, UK: Oxford University Press.

Murray, H. J. R. (1952). *A history of board games other than chess*. Oxford, UK: Clarendon Press.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics, 8*(4), 343–366.

OECD. (2017). *PISA 2015 technical report*. Paris, France: OECD. https://www.oecd.org/pisa/data/2015-technical-report/.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling, 53,* 315–333.

Olsen, L. W. (2003). *Essays on Georg Rasch and his contributions to statistics*. Ph.D. thesis. Institute Of Economics, University of Copenhagen, Denmark. https://www.rasch.org/olsen.pdf.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 185,* 71–110. https://doi.org/10.1098/rsta.1894.0003.

Poole, K. T. (2005). *Spatial models of parliamentary voting*. Cambridge, UK: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Volume 1 of Studies in mathematical psychology. Copenhagen, Denmark: Danmarks Paedagogiske Institut (Danish Institute for Educational Research).

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19,* 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x.

Raymond, M. R., Clauser, B. E., Swygert, K. A., & van Zanten, M. (2009). Measurement precision of spoken English proficiency scores on the USMLE Step 2 Clinical Skills examination. *Academic Medicine, 84*(10 Suppl.), S83–S85.

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 38,* 32–60. https://doi.org/10.3102/1076998614531045.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82,* 795–819. https://doi.org/10.1007/s11336-016-9544-7.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2014). *Handbook international large-scale assessment: Background, technical issues, and methods of data analysis.* London, UK: CRC Press (Chapman & Hall).

Schork, N. J., Allison, D. B., & Thiel, B. (1996). Mixture distributions in human genetics research. *Statistical Methods in Medical Research, 5*(2), 155–178. https://doi.org/10.1177/096228029600500204.

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series Volume 2 (pp. 113–135). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_05.pdf.

Schulz, W. (2017). Scaling of questionnaire data in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 384–410). Chichester, UK: John Wiley & Sons Ltd.

Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation, 17*(6), 447–464.

Schulz, W., & Friedman, T. (2015). Scaling procedures for ICILS questionnaire items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley, & E. Gebhardt (Eds.), *International Computer and Literacy Information Study 2013 technical report* (pp. 177–220). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/icils-2013-technical-report.

Schulz, W., & Friedman, T. (2018). Scaling procedures for ICCS 2016 questionnaire items. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (139–243). Amsterdam, the Netherlands: IEA. https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report.

Schwalbe, U., & Walker, P. (2001). Zermelo and the early history of game theory. *Games and Economic Behavior, 34*(1), 123–137.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408. https://doi.org/10.1007/BF02294363.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2,* 309–322.

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12188.

Verhelst, N. D. (2012). Profile analysis: A closer look at the PISA 2000 reading data. *Scandinavian Journal of Educational Research, 56*(3), 315–332. https://doi.org/10.1080/00313831.2011.583937.

Verhelst, N. D., & Verstralen, H. H.F. M. (1997). Modeling sums of binary items by the partial credit model. Measurement and Research Department Research Report 97-7. Arnhem, Netherlands: Cito.

von Davier, M. (2005). *A general diagnosis model applied to language testing data.* ETS Research Report RR-05-16. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2005.tb01993.x.
von Davier, M. (2016). The Rasch model. In W. van der Linden (Ed.), *Handbook of item response theory, Volume 1* (2nd ed.) (pp. 31–48). Boca Raton, FL: CRC Press. http://www.crcnetbase.com/doi/abs/10.1201/9781315374512-4.
von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution Rasch models.* New York, NY: Springer.
von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages.* New York, NY: Springer.
von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Boca Raton, FL: CRC Press.
von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*, 115–124. https://doi.org/10.1027/1614-2241.3.3.115.
von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT Models: An extension of the generalized partial-credit model. *Applied Psychological Measurement, 28*(6), 389–406. https://doi.org/10.1177/0146621604268734.
von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series Volume 2 (pp. 9–36). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf.
von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 1039–1055)., Psychometrics North Holland: Elsevier.
von Davier, M., Yamamoto, K., Shin, H.-J., Chen, H., Khorramdel, L., Weeks, J., et al. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy and Practice, 26*(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642.
Wright, B. D. (1968). Sample-free test calibration and person measurement. In B.S. Bloom (Ed.), *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85–101). Princeton, NJ: Educational Testing Service.
Xu, X., & Von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data.* ETS Research Report RR-06-08. https://doi.org/10.1002/j.2333-8504.2006.tb02014.x.
Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linking in NAEP. *Journal of Educational Statistics, 17*(2), 155–173.
Zermelo, E. (1913). On an application of set theory to the theory of the game of chess. Reprinted in E. Rasmusen (Ed.). (2001). *Readings in games and information.* Oxford, UK: Wiley-Blackwell.
Zermelo, E. (1929). The calculation of tournament results as a maximum-likelihood problem [in German]. *Mathematische Zeitschrift, 29,* 436–460.

Matthias Von Davier research focuses on developing psychometric models for analysing data from complex item and respondent samples and on integrating diagnostic procedures into these methods. His areas of expertise includes topics such as item response theory, latent class analysis, classification and mixture distribution models, diagnostic models, computational statistics, person-fit, item-fit, and model checking, as well as hierarchical extension of models for categorical data analysis, and the analytical methodologies used in large scale educational surveys.

Eugenio Gonzalez is a Principal Research Project Manager at Educational Testing Service (ETS), and director of the IEA-ETS Research Institute (IERI), a collaborative effort between the International Association for the Evaluation of Educational Achievement (IEA) and ETS that focuses on improving the science of large-scale assessments. IERI undertakes activities around three broad areas of work that include research studies related to the development and implementation of large-scale assessments; professional development and training; and dissemination of research findings and information gathered through large-scale assessments. Dr Gonzalez is also responsible for the technical documentation and international database training activities for PIAAC and PISA.

Dr. Gonzalez was formerly head of the Research and Analysis Unit at the IEA Hamburg (2007–2012), the director of quality control and field operations for the National Assessment of Educational Progress (NAEP) (2004–2006), and director of international operations and data analysis in the TIMSS & PIRLS international study center (ISC) at Boston College (1994–2004). In this last role, he oversaw the development and implementation of international operations, data analysis, and reporting procedures for the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). Since 1993, he has conducted database training activities for several research and governmental organizations, and has been a developer and technical lead of customized software for analyzing international large-scale assessment databases such as the IEA's IDB Analyzer and the Data Explorer. Dr. Gonzalez has also served as external consultant to several national and international large scale assessment programs, and has held teaching appointments at Boston College and the University of Massachusetts, Amherst. He has a PhD in Educational Research, Measurement, and Evaluation from Boston College, and an undergraduate degree in Psychology from the Universidad Católica Andres Bello in Caracas, Venezuela.

Wolfram Schulz is a Principal Research Fellow (formerly Research Director International Surveys) at the Australian Council for Educational Research (ACER) where he has worked on a large number of national and international large-scale assessment studies. He is International Study Director of the IEA International Civic and Citizenship Education Study (ICCS) and Assessment Coordinator for the IEA International Computer and Information Literacy Study (ICILS). He is also a member of the IEA Technical Executive Group (TEG).