Chapter 14
CTT and No-DIF and ? = (Almost) Rasch Model

Matthias von Davier

Abstract Assuring the absence of differential item functioning (DIF) is one of the central goals when constructing a test using either classical test theory (CTT) or item response theory (IRT). One of the most prominent methods of DIF detection is the Mantel Haenzel (1959) procedure that was suggested for this purpose by Holland and Thayer (1986). This test is not only used for DIF detection, a fact sometimes forgotten by educational measurement practitioners, and is often also called the Cochran-Mantel-Haenzel test. The basis of this test is a comparison of odds-ratios of several 2 by 2 tables, which is utilized in educational testing in the context of conditional 2 by 2 tables given the different ordered categories of a variable that represents proficiency or skill levels. In this note, I am expanding existing work that relates the Cochran-Mantel-Haenzel test used in conjunction with a simple sum score variable to the Rasch model. As I have pointed out in previous publications, the simple raw score, being the sum of binary scored responses, has certain desirable features, but is also limited in the sense of how information is used (e.g. von Davier 2010, 2016; von Davier and Rost 2016). In the context of CTT, as well as the use of the Cochran-Mantel-Haenzel procedure in CTT, and its relationship to the assumptions made in the Rasch model, however, the use of the sum score in conditional odds ratios is what brings these important approaches in applied test theory together on a formal mathematical basis.

14.1 Introduction

There have been prior attempts to relate CTT and IRT (Holland and Hoskens 2003; Bechger et al. 2003) as well as attempts to compare extensions of the Rasch model and MH-DIF approaches (Linacre and Wright 1989; Paek and Wilson 2011). However, the current approach, while rooted in the findings of these prior studies tries to approach the issue from a slightly different angle: In this note I focus on

M. von Davier (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: mvondavier@ets.org

M. Rosén et al. (eds.), *Cognitive Abilities and Educational Outcomes*,
Methodology of Educational Measurement and Assessment,
DOI 10.1007/978-3-319-43473-5_14

what is missing, or at least not stated explicitly in the use of (a) the MH-DIF procedure and (b) item-total regressions for test assembly in conjunction with CTT, and how slightly stronger versions of these requirements of ‘good items’ relate to a model that is virtually identical to IRT or the Rasch model. The basis for this chapter is an examination of tests consisting of binary (correct/incorrect) response variables. However, most results generalize in straightforward ways to tests with polytomous ordinal responses or mixed binary/polytomous tests (e.g. von Davier and Rost 1995; von Davier 2010).

Comparing item analyses to approaches in Biometrics and other fields (Cochran 1954; Armitage 1955), one finds similarities to the assessment of associations between a binary variable and an ordered categorical variable (in test theory this will be very often the sum score). For example, Armitage (1955) states:

> One frequently encounters data consisting of a series of proportions, occurring in groups which fall into some natural order. The question usually asked is then not so much whether the proportions differ significantly, but whether they show a significant trend, upwards or downwards, with the ordering of the groups.

Note that this whole paper is obsolete once it is understood that IRT and some types of nonlinear factor analysis are equivalent (e.g. Takane and DeLeeuw 1987 and more recently Raykov and Marcoulides 2016), and that all of test theory can be covered using a common, unified framework (McDonald 1999). In this note, we hope to add to this discussion by means of a one-by-one comparison of the assumptions made, in particular those of

(a) Absence of DIF versus local independence,
(b) True score + error versus sufficiency of (weighted) total score,
(c) Positive item correlation versus strict monotonicity.

This note also offers a perspective of how these weaker assumptions made in CTT need to be only slightly strengthened to arrive at a model that is virtually identical to the Rasch model or the subgroup of IRT models with sufficient statistics for the person variable (e.g. OPLM & 2PL in the binary case).

Also, several proponents of classical test theory (CTT) on the one hand and the Rasch model (or IRT) on the other hand may need some additional gentle nudging toward the insight that using either approach leads to good test instruments that are compatible with the (seemingly) competing other approach. This insight has its roots in the fact that both approaches, CTT and (unidimensional) Rasch and IRT models are special cases of generalized latent variable models (e.g. Moustaki and Knott 2000; Rabe-Hesketh et al. 2004).

The need to provide a generalized class of models such as the general diagnostic model (von Davier 2005, 2008) grows out of the understanding that several competing hypothesis about the structure of the variables we are aiming to assess can be directly compared and tested against each other if they are specified in a coherent statistical framework. However, there may be cases where several competing hypothesis are indeed providing very similar descriptions of the data while profoundly differing in the underlying assumptions made (von Davier et al. 2012). In other case, several models that appear to be different, or even extensions of another approach may turn out to be mere equivalent versions that can be covered in the generalized modeling framework by means of a reparameterization (von Davier 2013, 2014).

Generalizations of CTT to linear factor models are the predecessors of these generalized (linear and nonlinear) latent variable models. Also, several tests around the world, some of them high-stakes instruments used for highly consequential decisions are still being designed and assembled using the principles of ’vanilla’ CTT, together with customary tools to ensure psychometric quality. Among these, procedures for assessing (ensuring the absence of) differential item functioning (DIF) are one of the central foci when constructing a test using either classical test theory (CTT) or item response theory (IRT). Not only Lord and Novick (1968) and others (e.g. Wainer 1988) emphasize the importance of items and their resulting scores as fundamental to the test score. Moreover, ensuring the absence of DIF is considered one of the fundamental goals of test construction in order to provide fair assessments (Dorans 2013). One of the most prominent methods of DIF detection is the Mantel-Haenzel procedure that was suggested for this purpose by Holland and Thayer (1986). This test is not only used for DIF detection, a fact sometimes forgotten by educational measurement practitioners, and often also called the Cochran-Mantel-Haenzel test as methods related to this test have been discussed by Cochran (1954) and Mantel and Haenzel (1959). The basis of this test is a comparison of odds-ratios of several 2 by 2 tables, which is utilized in educational testing in the context of conditional 2 by 2 tables given the different ordered categories of a variable that represents proficiency or skill levels, very often taking the form of observed score groups.

In this note, I am expanding existing work that relates the Cochran-Mantel-Haenzel test used in conjunction with a simple sum score variable to the Rasch model. As I have pointed out in previous work, the simple raw score, being the unweighted sum of scored item responses, has certain desirable features, but is also limited in the sense of how information is used (e.g. von Davier 2010, 2016; von Davier and Rost 2016). In the context of CTT, as well as the use of the Cochran-Mantel-Haenzel procedure in CTT, and its relationship to the assumptions made in the Rasch model, however, the use of the sum score in conditional odds ratios is what brings these important approaches in applied test theory together on a formal mathematical basis. This chapter reviews the assumptions made in the Rasch model and how tests that fulfill these assumptions turn out to be ‘good’ tests in light of CTT measures of quality, and vice versa, when taking the definition of absence of DIF broadly. While related work pointed out other types of similarities, a direct argument of equivalency under these separately developed sets of preconditions for test quality has not been attempted to my knowledge.

14.2 Notation

Let $X_1, \ldots, X_K$ denote $K > 2$ binary random variables, and let Ω denote the population of respondents on which these random variables can be observed. We will assume for simplicity that the population is finite, and that we can sum over variables observed on samples from Ω. The random variables are representing the scored item responses on a test, and the respondents are represented as members $u \in \Omega$ of a population of potential test takers.

For each respondent, we may define the probability of responding correctly to each of these items X_i. More specifically, let

$$p_{ui} = P(X_i = 1|u)$$

denote that respondent u produces a correct response on item i. Note that this is not the marginal probability of a correct response, but the probability of a correct response for a given $u \in \Omega$.

Then, for considering a realization of the random variables, let

$$x_{ui} = X_i(u) \in \{0, 1\}$$

denote the binary item response to item i by a respondent u from the population Ω. The code $X_i = 1$ represents correct responses, while $X_i = 0$ represents incorrect responses.

The observed score, the total number of correct responses for a respondent u, aggregated across the K response variables will be denoted as

$$s_X(u) = \sum_{i=1}^{K} x_{ui}.$$

In addition to the item responses, there may be additional random variables that are defined for the population Ω. As an example, background information about each of the potential respondents $u \in \Omega$ can be represented as random variables z_j with $j = 1, \ldots, J$. For a variable that represents gender, for example, z_j, for some j may be defined as

$$z_{ui} = Z_i(u) \in \{male, female\}$$

which may also be coded as $\{0, 1\}$ as it is possible for any binary variable. Note that the additional variables could also represent other types of data, such as answers to items on other tests or questionnaires.

One additional random variable should be considered, one that represents the target of inference. Tests are typically given to make inferences about a skill, or an attribute, often quantitative in nature, which underlies test performance. A 10-item mathematics test is supposed to test more than the performance of students on the

10 items, but rather represent something that speaks more generally about the skill or ability of these students to solve these and similar mathematics problems.

Formally, each respondent u is assumed to possess a level of ‘skill’, mathematically a continuous random variable Υ, with

$$\tau_u = \Upsilon(u)$$

representing the skill level of respondent u. In different approaches to test theory, there will be different instances of this ‘skill level’ variable τ. In classical test theory (CTT), the true score, T, can be viewed as a version (a function of) τ specific to a test form, and in item response theory and Rasch models, the ‘skill level’, τ, will appear in the form of the person parameter, θ, which can also be assumed to be a function of τ.

14.3 Classical Test Theory in a Nutshell

CTT assumes that the observed score $S_X(u)$ can be written as the sum of two components. The foundational equation of the CTT is

$$S_X(u) = T_X(u) + e_X(u)$$

and much has been written about how to interpret these components. The most common setting is that $T_X(u)$ is the expected score on test X for respondent u, assuming either that the test can be repeated indefinitely, or that, based on additional model assumptions, an expected score can be calculated (see the corresponding section below).

Note that this definition of

$$T_X(u) = E(S_X|u)$$

as conditional expectation leads to a number of implications. First, $T_X(u)$ is often referred to as the ‘true score’, even though it is more accurately described as the conditional expectation of the sum score given respondent $u \in \Omega$.

This conditional expectation is vanishing for all respondents, so for any subset of respondents $U \subseteq \Omega$ we also have $E(e_X|U) = 0$. In particular for subsets of the type

$$U_T = \{u \in \Omega | T_X(u) = T\}$$

As a corollary we obtain that

$$E(e_X|T) = E(e_x|U_T) = \int\limits_{\{u \in \Omega | T_X(u) = T\}} E(e_X|u)p(u)du = 0$$

for any true score T.

Hence, the error variable, e_X, and T_X are, by definition of T_X, uncorrelated in Ω. Therefore, the total variance of the scores $V(S_X)$ can be written as

$$V(S_X) = V(T_X) + E[V(e_X)].$$

Note that this equation decomposes the total variance into the variance of T_X in Ω and the expected variance of the error term e_X. This result follows directly from the definition of T_X and e_X. Measures of reliability and the extent to which a score has validity are, at least in the traditional understanding of these concepts in CTT (Thurstone 1931), based on the correlation of the true score to true scores on other tests that are measures of the same underlying concept, or by means of correlations of the true score and other types of measures that are potentially difficult or expensive to collect, but can be considered the underlying target of inference.

In addition to the foundational assumption of CTT, measures of quality assurance include the selection and assembly of the items as components of the total score. Among these, the most prominent assumptions, or better selection criteria for items, are the absence of differential item functioning (no-DIF) and the presence of (moderate to high) correlations between the item score, X_i, and the total score, S_X.

More specifically, for the absence of DIF, it is assumed that for a number of grouping variables that separates the population into two groups, f and r, the conditional response probabilities by group membership and by total score are the same, that is

$$P(X_i = 1|S_X, f) = P(X_i = 1|S_X, r) = P(X_i = 1|S_X).$$

Expressed as odds ratio for the binary grouping $G: \Omega \rightarrow \{f, r\}$, this equality becomes

$$O_{S_X}(X_i, G) = \frac{P(X_i = 1|S_X, f)}{P(X_i = 0|S_X, f)} \times \frac{P(X_i = 0|S_X, r)}{P(X_i = 1|S_X, r)} = 1.$$

Basically, traditional uses of DIF restrict the study to grouping variables that are of policy relevance such as gender and ethnic minority status. However, there is nothing in the definition that would prevent us from applying the MH-DIF concept broadly, to any binary grouping variables, including those of other items. This use of another item response for splitting the sample is common practice in testing assumptions of the Rasch model (e.g. van den Wollenberg 1982; Verhelst 2001; von Davier 2016).

The Mantel Haenzel (MH) statistic uses a quantity that can be understood as the average odds ratio to test for DIF, more specifically if

$$MH(i,G) = \frac{\sum_{s=1}^{K-1} \frac{N\left(X_i=1\wedge f\middle|s\right)N\left(X_i=0\wedge r\middle|s\right)}{N(s)}}{\sum_{s=1}^{K-1} \frac{N\left(X_i=0\wedge f\middle|s\right)N\left(X_i=1\wedge r\middle|s\right)}{N(s)}} \approx 1$$

we may assume that there is no DIF for item i with respect to grouping variable, G. The expression $N(s)$ represents the frequency of score s. The notation $A \wedge B$ represents the conjunction of events A and B, that is, "A and B" was observed. As an example $N(X_i = 1 \wedge f|s) = N(X_i = 1 \wedge G = f|s)$ denotes the frequency of item i being solved in the focus group given score s. Note that the sum does not include terms for the total scores 0 or K since $P(X_i = 1|S_X = 0) = 0$ and $P(X_i = 0|S_X = K) = 0$ if the item score, X_i, is part of the sum score S_X.

In essence, the MH test statistic is used to check whether the conditional probabilities of success are the same across a variety of subpopulations. Traditionally, DIF analyses includes gender and ethnicity based groupings, but other types of groupings can obviously be used as well.

The positive item-total correlation criterion is based on the rationale that the covariance of the item score, X_i, and the total score, S_X, of which X_i is an additive component, should be in the same direction for all items. The underlying assumption is that the probability of a correct response should increase with increasing true score, which is the expectation of the observed score as defined above.

This covariance can be written as

$$\begin{aligned} cov(X_i, S_X) &= \sum_{x=0}^{1}\sum_{s=0}^{K} P(X_i = x, S_X = s)x \cdot s - E(X_i)E(S_X) \\ &= [E(S_X|X_i = 1) - E(S_X)]P(X_i) \end{aligned}$$

which is nonnegative whenever

$$E(S_X|X_i = 1) \geq E(S_X).$$

Alternatively, the cross product part of the covariance can also be written as

$$\sum_{s=0}^{K}[P(X_i = 1|S_X = s) \cdot s]P(S_X = s) = E[S_X \cdot P(X_i = 1|S_X)]$$

and

$$cov(X_i, S_X) = E[S_X \cdot P(X_i = 1|S_X)] - E(X_i)E(S_X)$$

One straightforward way to ensure positivity is postulating that the conditional probabilities of solving the item given a specific score increase with increasing total score. That is, one may assume

$$P(X_i = 1|S_X = s) \geq P(X_i = 1|S_X = t)$$

for any two scores with $s > t$. This basically ensures that there are more test takers expected to solve the item in groups with higher total scores.

14.4 Rasch Model

With the notations above, the Rasch model assumes the following association between person skill level τ_u and expected performance on a response variable. For all $u \in \Omega$ it is assumed that

$$P(X_i = 1|u) = \frac{\tau_u}{d_i + \tau_u}, \tag{14.1}$$

and, customarily, this definition is used with the transformations $\exp(\theta_u) = \tau_u$ and $\exp(b_i) = d_i$. Hence, the above definition is equivalent to

$$P(X_i = 1|u) = \frac{\exp(\theta_u)}{\exp(b_i) + \exp(\theta_u)} = \frac{\exp(\theta_u - b_i)}{1 + \exp(\theta_u - b_i)} \tag{14.2}$$

which is the form commonly recognized as the dichotomous Rasch model (e.g. Rasch 1966; von Davier 2016). The θ_u is commonly referred to as the person parameter and the b_i is referred to as the item parameter.

Then, for the set of response variables, $X_1, \ldots, X_K$, it is assumed that conditional independence holds. This translates to the assumption that the joint probability of observing responses $x_1, \ldots, x_K$ is given by

$$P(X_1 = x_1, \ldots, X_K = x_K|u) = \prod_{i=1}^{K} \frac{\exp(x_i[\theta_u - b_i])}{1 + \exp(\theta_u - b_i)} \tag{14.3}$$

the product of the item specific responses. the above equation it is easily verified by noting that

$$P(X_i = 0|u) = 1 - P(X_i = 1|u) = \frac{1}{1 + \exp(\theta_u - b_i)}.$$

The expression for the joint probability in Eq. (14.3) can be rearranged so that

$$P(x_1,\ldots,x_K|\theta) = A(x_1,\ldots,x_K)\cdot B[S_X(u),\theta]\cdot C(\theta) \tag{14.4}$$

with

$$A(x_1,\ldots,x_K) = \prod_{i=1}^{K}[\exp(-x_{ui}b_i)]$$

and

$$B[(S_X(u),\theta)] = \exp[S_X(u)\theta]$$

and

$$C(\theta) = \prod_{i=1}^{K}\left[\frac{1}{1+\exp(\theta-b_i)}\right]$$

for any skill level $\theta\in\mathbb{R}$. This result can be utilized to calculate the probability of a response pattern given the raw score, S_x. This is done by calculating

$$P(S_X|\theta) = B[S_X(u),\theta]\cdot C(\theta)\left[\sum_{\{(x_1,\ldots,x_k)|\sum x_i=S_x\}} A(x_1,\ldots,x_K)\right]$$

the sum of the probabilities of all response patterns according to Eq. (14.4). For any given response pattern $(x_1^*,\ldots,x_k^*)$ with sum score $\sum x_i^* = S_X$ the conditional probability of observing this particular response vector among those with the same score becomes

$$P(X_1 = x_1^*,\ldots,X_k = x_k^*|S_X) = \frac{\prod_{i=1}^{K}[\exp(-x_i^*b_i)]}{\sum_{\{(x_1,\ldots,x_k)|\sum x_i=S_x\}}\prod_{i=1}^{K}[\exp(-x_ib_i)]}. \tag{14.5}$$

The above expression is obtained by integrating out the latent skill variable θ, exploiting that the identity holds for every level of θ. The expressions

$$\gamma_K[\mathbf{b}=(b_1,\ldots,b_K),S_X] = \sum_{\{(x_1,\ldots,x_k)|\sum x_i=S_x\}}\prod_{i=1}^{K}[\exp(-x_ib_i)]$$

are commonly referred to as the symmetric functions (e.g. Gustafson 1980; von Davier and Rost 1995; von Davier 2016) for $S_X = 0,\ldots,K$ and S_X is called the 'order' of the function. The result of importance here is that this expression can be utilized to find

$$P\left(X_j = 1|S_X\right) = \sum_{\{(x_1,\ldots,x_k)|\sum x_i = S_x, x_j = 1\}} P(X_1 = x_1, \ldots, X_j = 1, \ldots, X_K = x_K|S_X) \tag{14.6}$$

for any item j and any raw score $S_{X.}$ Equations (14.5) and (14.6) show that S_X is the minimally sufficient statistic (Fisher 1922) for parameter θ in the Rasch model. It can be further shown that

$$P\left(X_j = 1|S_X\right) = \frac{-\left(\frac{\partial \gamma_K[\mathbf{b}, S_X]}{\partial b_j}\right)}{\gamma_K[\mathbf{b}, S_X]},$$

that is, that the derivative of the symmetric function with respect to item difficulty b_i can be used in an expression to calculate the conditional score probabilities. The sum in the above Eq. (14.6) runs over all response vectors with the same raw score S_X and with the additional condition that for the item of interest, $x_j = 1$. Most importantly, in the Rasch model the probability of a correct response on item j for raw score group S_X can be calculated without any assumptions about the skill level θ, or its distribution in the population, or about the true score $T_X = E(S_X)$.

14.5 From Rasch Model to CTT

If it can be shown that if the Rasch model holds for a test $\mathbf{X} = (X_1, \ldots, X_K)$, then the classical test theory summary score S_X has 'good' properties, in the sense of that the sum score of this test will provide a satisfactory summary of the data at hand. Hambleton and Jones (1993) pointed out that item response theory (IRT) [and the Rasch model] are strong models, in the sense of that model assumptions made allow derivation of stronger results. As an example, sample independence of parameters and specific objectivity (Rasch 1966) can be derived from these model assumptions, while these cannot be obtained from CTT without making additional assumptions (von Davier 2010, 2016).

14.6 Sufficiency and Total Score

The Rasch model as defined above has some outstanding mathematical features. One of the most salient features is that it turns out that if the Rasch model holds, the total score, $S_X(u)$, is a sufficient statistic for the person parameter, θ_u. In mathematical statistics, a statistic $S = f(X_1, \ldots, X_K)$ is sufficient for a parameter θ if

$$P(X_1, \ldots, X_K|\theta) = P(X_1, \ldots, X_K|S)P(S|\theta)$$

or, equivalently, if

$$\frac{P(X_1, \ldots, X_K|\theta)}{P(S|\theta)} = P(X_1, \ldots, X_K|S).$$

The property of sufficiency can be described as the ability to separate (or eliminate) parameters by conditioning on the sufficient statistics when calculating the unconditional probability of the observed data.

For the Rasch model, the sufficiency of the total score, S_X, allows us to predict the distribution of the response variables, X_i, for all i based on the item parameters, $b_1, \ldots, b_K$. This result means that, if the Rasch model holds, the sum score $S_X(u) = \sum_i X_i$ is all that is needed to summarize the data.

The statistic S_X is the score typically utilized in CTT as the basis for inferences. The fact that this is the sufficient statistics in the Rasch model—a probability model for predicting item responses at the individual level—gives substantial credence to this common choice in CTT. Note that the choice of the unweighted sum score $S_X = \sum x_i$ is, while arguably the simplest form of aggregation, nevertheless a completely arbitrary one (Gigerenzer and Brighton 2009; von Davier 2010). In addition, other IRT models exist that use different assumptions leading to other types of sufficient statistics, not the simple total number correct. As such, there is a clear connection between many, if not the vast majority, of applications of CTT and the Rasch model in that the simple sum score, that is, the total number of correct responses, plays a central role in both approaches.

14.7 Local Independence, True Score, and Error Variance

The assumption of local independence as given in Eq. (14.3) provides a basis for looking at what the expected score for a person u might be. Note that the expected score on a test is what forms the basis of the additive decomposition of observed score, $S_X(u)$, into true score, $T_X(u)$, and error component, $e_X(u)$.

The reasoning is as follows: If the Rasch model holds, we can assume local independence, so that the expected true score can be calculated based on the model equation, summing up the conditional response probabilities across items. That is, we can write

$$E[S_X(u)] = T_X(u) = \sum_{i=1}^{K} P(X_i = 1|\theta_u) = \sum_{i=1}^{K} p_{ui}$$

for all u. In addition, the error variance of $e_X(u) = S_X(u) - T_X(u)$ can be written as

$$V[e_X(u)] = \sum_{i=1}^{K} p_{ui}(1 - p_{ui})$$

since independence given u holds.

This means that the Rasch model (and more general IRT) will provide direct estimates of the true score and the error variance, if the person parameter, θ_u, is known. This can be used, and is being used, for example in the prediction of expected scores on test forms that have not been taken by a respondent, by means of what is known as 'true score equating'.

14.8 No-DIF

The Rasch model is based on assumptions that apply to all respondents in the population, that is, for all $u \in \Omega$ it provides an expression that relates the probability of success to an item difficulty and a person skill level through

$$p(X_i = x|u) = \frac{\exp(x[\theta_u - b_i])}{1 + \exp(\theta_u - b_i)}.$$

Note that there is no person dependent variable other than θ_u included in the definition of this probability. More specifically, this implies that if the Rasch model holds for all $u \in \Omega$, as given in the expression above, we can conclude that the same probability hold for all levels of θ.

However, there is an even more direct way to show that if the Rasch model holds with items parameters, b_i, for all $i = 1, \ldots k$, we can expect that the MH-test for DIF will turn out such that there is no indication of DIF. More specifically, recall the result that shows how to calculate the conditional probability of a response for a score group. We have obtained

$$P\big(X_j = 1|S_X\big) = \sum_{\{(x_1,\ldots,x_k)|\sum x_i = S_x, x_j = 1\}} P(X_1 = x_1, \ldots, X_j = 1, \ldots, X_K = x_K|S_X) \tag{14.7}$$

for any item j and any raw score S_X if the Rasch model holds. For each grouping variable $G : \Omega \to \{r,f\}$ that separates the population in into members of a focal versus a reference group, we obtain estimates of the relative frequencies

$$\hat{P}\big(X_j = 1|S_X,f\big) = \frac{N(X_i = 1 \wedge S_X \wedge f)}{N(S_X \wedge f)}$$

the relative frequency of a success on item j of persons with score S_X in the focus group and

$$\hat{P}\left(X_j = 1|S_X, r\right) = \frac{N(X_i = 1 \wedge S_X \wedge r)}{N(S_X \wedge r)}$$

the relative frequency of a success on item j of persons with score S_X in the reference group. It directly follows from the weak law of large numbers that these relative frequencies converge to $P\left(X_j = 1|S_X\right)$ if the Rasch model with given parameters holds in Ω. This trivially implies that the odds also converge to the same expected odds

$$\frac{\hat{P}\left(X_j = 1|S_X,f\right)}{\hat{P}\left(X_j = 0|S_X,f\right)} \rightarrow \frac{P\left(X_j = 1|S_X\right)}{P\left(X_j = 0|S_X\right)} \leftarrow \frac{\hat{P}\left(X_j = 1|S_X, r\right)}{\hat{P}\left(X_j = 0|S_X, r\right)}.$$

Finally, this result implies that with growing sample size, all odds ratios in all score groups will converge to the values calculated based on the true parameters and the symmetric functions as given in Eq. (14.7) if the Rasch model holds with item parameters $b_1, \ldots, b_K$ in the population Ω.

Note that there are straightforward extensions that allow for added features to the Rasch model to account for DIF. As an example, for given groups $\{f, r\}$ one could assume that the Rasch model holds, but with different sets of parameters such that

$$P(X_i = 1|\theta, g) = \frac{\exp\left(\theta - b_{ig}\right)}{1 + \exp\left(\theta - b_{ig}\right)}$$

in group $g \in \{f, r\}$. This modification allows for group specific item difficulties so that b_{ir} and b_{if} are not necessarily the same (e.g. von Davier and Rost, 1995, 2006, 2016).

However, if the Rasch model holds with the same set of item parameters in all of the whole population, Ω, it follows that there is no DIF for any grouping variable.

14.9 Positive Item Regressions

In CTT, items are typically selected for multiple criteria. Aside from No-DIF and appropriate difficulty level, the main selection criterion is that of assuring positive correlation of the item score variable X_i with the total score S_X. Note that Armitage (1955) and others already aim for a stronger criterion of strict monotonic increasing proportions with increasing score variable (or some other ‘natural’ ordering of respondents). In the case that the Rasch model can be assumed to hold for a test in some population Ω it is straightforward to show that all item-total correlations are positive.

Recall that the expected item score is given by

$$E(X_i|\theta) = P(X_i = 1|\theta) = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}$$

which is strict monotonic increasing in θ. Also, the expected value of the observed score is the true score, which can be calculated as

$$E(S_X|\theta) = T_X(\theta) = \sum_{i=1}^{K} P(X_i = 1|\theta) = \sum_{i=1}^{K} E(X_i|\theta) \tag{14.8}$$

and is also strict monotonic increasing in θ. Finally, the covariance of the item score variable and the total score S_X can be expressed as

$$cov(X_i, S_X) = \int_{\theta} [E(X_i|\theta) - E(X_i)][E(S_X|\theta) - E(S_X)]f(\theta)d\theta$$

which is positive due to the strict monotonicity of $E(X_i|\theta)$ and $E(S_X|\theta)$ and that there exists a θ^* for which $E(S_X|\theta^*) = E(S_X)$ and by means of equation (14.8) and commutativity of finite sums and integration it follows that $E(X_i|\theta^*) = E(X_i)$. Hence, when the Rasch model holds, item-total correlations are positive.

14.10 CTT + Generalized No-DIF + Strict Monotone Item Regression = (Almost) IRT

The previous section has shown that a test designed to follow the Rasch model produces an outcome that has very satisfactory properties when looking at the test from the perspective of CTT. A test constructed by using the Rasch model as a guideline will produce a test in which the simple total score carries all information needed to estimate person skill level, the true score and the error variance can be calculated based on simple item level expected scores, and the test will not have DIF and all item-total correlations are positive.

In this section, the reverse direction is explored. When assembling a test using the basic assumptions of CTT and the customary measures of quality assurance, do we produce an instrument that can be fitted with an IRT model, in particular, the Rasch model?

14.11 CTT Total Score and the Rasch Model

The simple total number of correct responses, also often referred to as the total score

$$S_{\mathbf{X}}(u) = \sum_{i=1}^{K} x_{ui}$$

with binary responses $x_{ui} \in \{0, 1\}$ is compatible with the assumptions made in the Rasch model. It was shown in section that the total score S_X is a sufficient statistic, minimally suffcient statistic, in the Rasch model for the person parameter θ. A more general choice would be

$$W_{\mathbf{X},\mathbf{w}}(u) = \sum_{i=1}^{K} a_i x_{ui}$$

with (typically positive) real-valued weights a_i for $i = 1, \ldots, K$. There is no reason to prefer one over the other just by means of the defnition, indeed, the simple total score is a special case of the weighted score, i.e., $S_X(u) = W_{X,1}(u)$ (von Davier 2010). However, there are legitimate practical reasons to use the unweighted score, in particular if there is little or no information about how to calculate or determine the weights (e.g. Gigerenzer and Brighton 2008; Davis-Stober 2011).

However, there may be good reasons for choosing weights, either based on maximizing the predictive power of a score with respect to some external criterion, or with respect to some unobserved latent variable, or simply in terms of improving the prediction of item scores given the estimate of a person's skill level. It turns out that a number of cases can be identified for which different weighting schemes exhibit a direct correspondence to the sufficient statistic for person ability in an IRT model. Table XYZ gives three prominent examples, the Rasch model (Rasch 1960), the OPLM (Glas and Verhelst 1995) and the 2PL model (Birnbaum 1968).

	Score	Model	$P(X_i = 1\|\Theta)$
Simple total score (all weights equal to 1)	$\sum_{i=1}^{K} X_{ui}$	Rasch	$\frac{\exp\cdot(\Theta - b_i)}{1 + \exp\cdot(\Theta - b_i)}$
Pre-specified integer weights ($l_i \in \{0, 1, 2, \ldots\}$)	$\sum_{i=1}^{K} l_i X_{ui}$	OPLM	$\frac{\exp\cdot(l_i\cdot[\Theta - b_i])}{1 + \exp\cdot(l_i\cdot[\Theta - b_i])}$
Single factor model with positive weights ($a_i \in R^+$)	$\sum_{i=1}^{K} a_i X_{ui}$	2PL	$\frac{\exp\cdot(a_i\cdot[\Theta - b_i])}{1 + \exp\cdot(a_i\cdot[\Theta - b_i])}$

The above table provides another indication of how Rasch model and CTT are conceptually and mathematically connected. In both approaches, the simple total score is the central summary of observed response behavior. In the Rasch model this is a consequence of the assumptions made, while in CTT, the simple

(=unweighted) total score is often the central statistic chosen to represent a fallible measure of the true score on a test.

14.12 Absence of DIF—No-DIF 2.0

The no-DIF case when tested will be indicated by a value of the MH-statistic close to 1, see the Sect. 14.3 above. This value represents the odds ratio for the item probabilities in focus and reference group, averaged over total scores. Typically, this average odds-ratio is tested only for a handful of grouping variables such as gender and/or race/ethnicity. However, as pointed out above, the MH-DIF statistic can be calculated for any binary grouping variable.

At this point we need to deviate from the customary checks and propose additional conditions to make the CTT assumptions indeed commensurate with IRT assumptions. Hence, it is being acknowledged that CTT with the usual set of procedures is not based on strong enough assumptions to make the approach equivalent to IRT. However, it should be noted that the assumptions made in addition do not violate customary assumptions or directives for item selection. The absence of MH-DIF is tested by calculating the average over odds ratios, for example, while all that is needed is a slightly stronger assumption that requires the odds ratios in each of the score groups to be 1, that is, instead of the average odds ratio being 1, it is assumed that

$$\frac{P(X_i = 1|S_X,f)}{P(X_i = 0|S_X,f)} \cdot \frac{P(X_i = 0|S_X,r)}{P(X_i = 1|S_X,r)} = 1$$

for all $S_X = 1, \ldots, K - 1$. One may argue that this only provides what was intended when Mantel and Haenzel defined the MH-statistic, namely that across various groupings, the odds ratio is always 1, i.e., that the conditional probabilities in focal and reference group are the same given the conditioning on the total score. This extension, together with the absence of this type of DIF for any other binary grouping variables yields an assumption equivalent to local independence that is common in IRT models. Note that Linacre and Wright (1989) do indeed conjecture that if the same average odds ratio is to be expected in all types of groupings (intervals of total scores or similar) then each of the odds ratios should be in expectation the same. Here we take a slightly different approach and state this as an explicit assumption leading to a stricter criterion for item selection.

More specifically, the response to another item on the test, or an additional item that is not part of the test could also be used to group respondents. Let us assume for items $i \neq j \in \{1, \ldots, K\}$

$$\frac{P(X_i = 1|S_X, X_j = 1)}{P(X_i = 0|S_X, X_j = 1)} = \frac{P(X_i = 1|S_X, X_j = 0)}{P(X_i = 0|S_X, X_j = 0)}$$

so that respondents who solve item j, i.e., $X_j = 1$, are being treated as the focus group and $X_j = 0$ is equivalent to the reference group. Using the definition of conditional probabilities we have

$$P(X_i = 1|S_X, X_j = 1) = \frac{P(X_i = 1 \wedge S_X \wedge X_j = 1)}{P(S_X \wedge X_j = 1)} = \frac{P(X_i = 1 \wedge X_j = 1|S_X)}{P(X_j = 1|S_X,)}$$

so that

$$\frac{P(X_i = 1 \wedge X_j = 1|S_X)}{P(X_i = 0 \wedge X_j = 1|S_X)} \frac{P(X_i = 0 \wedge X_j = 0|S_X)}{P(X_i = 1 \wedge X_j = 0|S_X)} = 1$$

which equivalent to X_i, X_j being independent given S_X. This means that the stronger MH condition applied to one item response variable X_i and another item variable X_j viewed as the grouping variable yields local independence, conditional on the total score. Hence we can write

$$P(X_1 = x_1, \ldots, X_K = x_K|S_x) = \prod_{i=1}^{K} P(X_i = x_i|S_X)$$

as the pairwise local independence extends to the full response pattern probability by the same argument.

14.13 Item-Total Regression 2.0

The previous sections showed how a slightly stronger MH criterion applied to focal and reference groups defined by responses to another item yields local independence given total score. A similar approach will be taken in this section with the goal to extend and strengthen the positive item-total regression criterion. More specifically, recall that the positivity of the covariance of item score and total score can be studied by looking at the cross product of conditional response probability and total score, namely

$$cov(X_i, S_X) = E(P(X_i|S_X) \cdot S_X) - E(X_i)E(S_X)$$

with

$$E(P(X_i|S_X) \cdot S_X) = E(S_X|X_i = 1) \cdot E(X_i).$$

These equivalencies illustrate that higher conditional item response probabilities associated with higher total scores yield a more positive item-total covariance. The criterion of positive item-total covariance can hence be strengthened by assuming conditional item response probabilities to increase strictly with total scores. That is, the strong(er) version of a positive item-total regression requires

$$P(X_i = 1|s) > P(X_i = 1|t)$$

for all total scores $s > t \in \{0, \ldots, K\}$. This condition implies that

$$P(X_i = 1|S_x = 0) < P(X_i = 1|S_x = 1) < P(X_i = 1|S_x = 2) \\ < \cdots < P(X_i = 1|S_x = K).$$

Note that the 'spirit' of the positive item-total correlation was not abandoned but strengthened: All items that meet the slightly stronger assumption will also meet the weaker assumption that the item-total correlation is positive.

14.14 An Approximate IRT Model Based on Strengthened CTT Assumptions

The above sections introduced the total score S_X as the basic unit of analyses in CTT and showed that the same quantity is the minimal sufficient statistic for the person ability parameter in the Rasch model. In addition, two slightly strengthened CTT requirements were introduced. One that extends the MH approach of no-DIF requirement to additionally requiring all total score based odds ratios to be equal to 1. Finally, the positive item-total regression requirement was strengthened to the criterion that conditional item success probabilities are required to be strictly increasing with the total score.

These assumptions, and often even the weaker original assumptions with regard to item selection in CTT constructed tests commonly lead to a set of items that, when using the sum score or some other proxy to the true score or underlying ability, align in very systematic ways along the construct we want to measure. An early example can be found in Thurstone (1925) who plotted the relative frequency of success on a number of tasks used in developmental research against the age of respondents in calendar years. Figure 14.1 presents this association. Other examples can be found in Lord (1980) illustrating item-sumscore regressions.

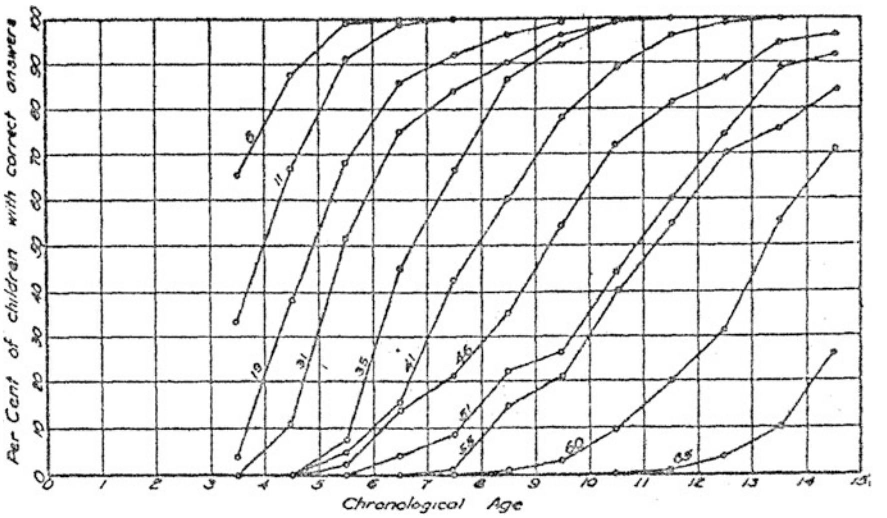

Fig. 14.1 Thurstone's (1925) illustration of item regressions, the relative frequencies of success are depicted as a function of age in calendar years

Given the obvious resemblance of the strictly monotonic item regressions in Fig. 14.1 and the item characteristic curves defined by the Rasch model or more general IRT models, the following approach is proposed: With the assumption that the strengthened versions of the customary CTT item selection requirements are met for $i = 1, \ldots K$ items X_i, define

$$\delta_{i,s} = \log\left[\frac{P(X_i = 1|s)}{P(X_i = 0|s)}\right]$$

for all $s \in \{1, \ldots, K - 1\}$ and note that

$$P(X_i = x|s) = \frac{\exp\left(x \cdot \delta_{i,s}\right)}{1 + \exp\left(\delta_{i,s}\right)}.$$

Note that Hessen (2005) defined constant log odds ratio (CLOR) models, and also studies the obvious relation of these to the MH procedure. In the in the context of the quantities defined above, CLOR models would be based on $\omega_{ij}(s) = \frac{\delta_{i,s}}{\delta_{j,s}}$ and an assumption is made that these log odds ratios are constant for all ability levels (here: total scores), which specifies that

$$\omega_{ij} = \frac{\delta_{i,s}}{\delta_{j,s}}$$

is a constant for all score groups. It turns out that this is a rather strong assumption, and CLOR models can be shown to be special cases of Rasch models with a (potentially) constrained ability range (Maris 2008). In our context, we will not use the above assumption but build up the argument from the assumed positive item total correlation, or its somewhat strengthened version, the monotonicity of conditional P+ in score groups. While the strengthened assumption is not logically implied by its weaker form (if it was, it would be redundant), it appears that it is often implicitly assumed when studying proportions in levels of a ‘natural ordering’ of respondents (Armitage 1955).

To continue the line of argumentation, there is the obvious requirement that all probabilities are non-vanishing, so that the δ are well defined. If the strengthened CTT assumption of strict monotonicity of proportions in score groups holds for the data at hand, we have

$$\delta_{i,1} < \delta_{i,2} < \delta_{i,3} < \cdots < \delta_{i,K-1}$$

for all items $i = 1, \ldots, K$. Next we define item-wise and score-wise effects as well as the grand mean of the δ. Let

$$\mu = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{s=1}^{K-1} \delta_{i,s}$$

and let

$$\beta_i = \mu - \frac{1}{K-1} \sum_{s=1}^{K-1} \delta_{i,s}$$

and finally

$$\tau_s = \frac{1}{K} \sum_{i=1}^{K} \delta_{i,s}.$$

by definition we have $\sum_i \beta_i = 0$. Then we can define

$$\hat{\delta}_{i,s} = \tau_s - \beta_i.$$

These $\hat{\delta}$ parameters can be used as approximation to the δ parameters. We can define a probability model by means of

$$\hat{P}(X_i = 1|s) = \frac{\exp(\tau_s - \beta_i)}{1 + \exp(\tau_s - \beta_i)}.$$

The similarity of this model to the Rasch model is evident, and relationships to log-linear Rasch models (e.g. Kelderman 1984, 2006) are obvious. However, there is need to assess how well this approximation works, since strict monotonicity in S and main effects in i are not necessarily assurance enough that the $\hat{\delta} = \tau - \beta$ are close to the δ. Alternatively, one could look at this as an optimization problem and miminize the difference

$$\sum_{i=1}^{K}\sum_{s=1}^{K-1}\left(\delta_{i,s} - [\alpha_i\tau_s - \beta_i]\right)^2.$$

In this case, the derived IRT like model turns out to be

$$\hat{P}(X_i = 1|s) = \frac{\exp(\alpha_i\tau_s - \beta_i)}{1 + \exp(\alpha_i\tau_s - \beta_i)}$$

and similarities to the 2PL IRT model can be observed.

With the implied conditional independence in score groups these yield a model for the full item response vectors. The strict monotonicity of the $\delta_{i,s}$ in s provides support for the use of a simple linear approximation rather than one that utilizes higher order moments of s or τ_s. However, more complex models such as

$$\delta_{i,s} = \sum_{m=0}^{M}\gamma_{i,m}s^m + e$$

can be considered. Given the strict monotonicity and restricted range of item total regressions, however, a linear approximation can be expected to perform well. Note that these models make use of the consequences of assumptions that are slightly stronger than those commonly made in CTT and arrive at models that look a lot like IRT.

14.15 Conclusions

This paper presents an (or yet another) attempt to relate the practices and customary procedures of classical test theory to the assumptions made in the Rasch model and IRT. While Wright and Linacre (1989) and Holland and Hoskens (2003), Bechger et al. (2003), as well as most recently Paek and Wilson (2011) all tackle slightly different angles of this issue, it appears that all parties attempting these types of endeavors agree on some basic similarities. CTT assumes the observed (typically unweighted) sum-score of (often binary) test items as the foundation of all analyses. Note however, that this choice of the aggregate is not ‘natural’ or ‘best’ by any means, but that different choices are possible and common in factor analysis as well as in IRT (McDonald 1999; Moustaki and Knott 2000; von Davier 2010). The basis of the sum score S_X as the person measure is extended by showing that the likelihood of solving an item, given this score, is unchanged in different groups under the stricter MH-no-DIF criterion. This yields local independence, a fundamental assumption made in many IRT models. Finally a slightly more rigorous requirement of strict monotone item-total regression yields strictly monotone log-odds, which are finally used to approximate the conditional response probabilities used in MH-DIF and item regressions by IRT type models.

The other direction, deriving 'good' CTT properties based on the Rasch model is much more straightforward. The Rasch model (and other unidimensional IRT models) make sufficiently rigorous assumptions that allow to derive satisfactory adherence to summary statistics used in CTT (unweighted total, or integer weighted, or real valued weighed sum score) as well as the requirement of no-DIF, and finally positive item-total correlations, if the items selected for a test follow these models. DIF can be incorporated in IRT models in a variety of ways, from multiple group IRT models (Bock and Zimowski 1997) with partial invariance (Glas and Verhelst 1995; Yamamoto 1998; Oliveri and von Davier 2014) to models that explicitly examine what split of the sample exhibits direct evidence of item by group interactions (e.g. von Davier and Rost 1995, 2006, 2016).

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics (International Biometric Society), 11*(3), 375–386. doi:10.2307/3001775. JSTOR 3001775.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*(5), 319–334.

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

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics, 10*, 417–451.

Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika, 76*, 650–669.

Dorans, N. (2013). *Test fairness*. Princeton, NJ (ETS RR-xx-13).

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A, 222*, 309368. doi:10.1098/rsta.1922.0009 (JFM 48.1280.02. JSTOR 91208).

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107–143. doi:10.1111/j.1756-8765.2008.01006.x

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.

Gustafsson, J.-E. (1980). A solution of the conditional estimation problem for long test in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40*(2), 377–385 (T270201 R).

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 3847.

Hessen, D. J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika, 70*(3), 497–516.

Holland, P. W., & Hoskens, M. (2003, March). Classical test theory as a first-order Item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika, 68* (1), 123–149.

Holland, P. W., & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure*. Technical Report No. 86 69. Princeton, NJ: Educational Testing Service.

Kelderman, H. (1984) Loglinear Rasch model tests. *Psychometrika, 49*(2), 223–245.

Kelderman, H. (2006). Loglinear multivariate and mixture Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models*. Springer: New York.

Linacre J. M., & Wright B. D. (1989). Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions, 3*(2), 52–53.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22(4), 719748. doi:10.1093/jnci/22.4.719

Maris, G. (2008). A note on “constant latent odds-ratios models and the Mantel-Haenszel null hypothesis”. *Psychometrika, 73*(1), 153–157.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika, 65*(3), 391–411.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*(1), 1–21. doi:10.1080/15305058.2013.825265

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison With MantelHaenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement, 71*(6), 1023–1046.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika ,69*, 167–190.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research: Copenhagen.

Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89–107).

Raykov & Marcoulides. (2016). One the relationship between classical test theory and item response theory: From one to the other and back. *Educational and Psychological Measurement, 76*, 325–338.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393-408.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451.

Thurstone, L. L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems* (113 p). Ann Arbor, MI, US: Edwards Brothers. doi:10.1037/11418-000.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123–139.

Verhelst, N. (2001). Testing the unidimensionality assumption of the Rasch model. *Methods of Psychological Research Online, 6*(3), 231–271. Retrieved from http://www.dgps.de/fachgruppen/methoden/mpr-online/issue15/art2/verhelst.pdf

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. Research Report RR-05-16. ETS: Princeton, NJ.

von Davier, M. (2008, November). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

von Davier, M. (2010). Why sum scores may not tell us all about test takers. In L. Wang (Ed.), Special issue on Quantitative Research Methodology. *Newborn and Infant Nursing Reviews, 10*(1), 27–36.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model—Two variants of a model equivalency. *BJMSP, 67*, 4971. doi:10.1111/bmsp.12003/abstract

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series.* doi:10.1002/ets2.12043/abstract

von Davier, M. (2016). The Rasch model (Chapter 3). In W. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, 2nd ed.). Berlin: Springer.

von Davier, M., Naemi, B., & Roberts, R. D. (2012). Factorial versus typological models: A comparison of methods for personality data. *Measurement: Interdisciplinary Research and Perspectives, 10*(4), 185–208.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models—Foundations, recent developments and applications* (pp. 371–379). New York: Springer.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26). Psychometrics. Amsterdam: Elsevier.

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models (Chapter 23). In W. van der Linden (Ed.), *Handbook of Item response theory* (Vol. 1, 2nd ed.). Berlin: Springer.

Wainer, H. (1988). *The future of item analysis*. Princeton, NJ: ETS (ETS Research Report No. RR-88-50).

Yamamoto, K. (1998). Scaling and scale linking. In T. S. Murray, I. S. Kirsch, & L. B. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first international adult literacy survey* (pp. 161–178). Washington, DC: National Center for Education Statistics.