Psychological Test and Assessment Modeling, Volume 60, 2018 (1), 3-27

Testing psychometric properties of the CFT 1-R for students with special educational needs

Jörg-Henrik Heine[1], Markus Gebhard[2], Susanne Schwab[3], Phillip Neumann[4], Julia Gorges[4] & Elke Wild[4]

Abstract

The Culture Fair Intelligence Test CFT 1-R (Weiß & Osterland, 2013) is one of the most used tests in Germany when diagnosing learning disabilities (LD). The test is constructed according to the classical test theory and provides age specific norms for students with LD in special schools. In our study, we analyzed the test results of 138 students in special schools and 166 students with LD in inclusive settings in order to test the measurement invariance between students with LD, who are educated in these two different educational settings. Data were analyzed within an IRT framework using a non-iterative approach for (item) parameter recovery. This approach parallels with the principle of limited information estimation, which allows for IRT analyses based on small datasets. Analyses for Differential Item Functioning (DIF) as well as a test for global and local model violations with regard to both subgroups were conducted. The results confirmed the assumption of measurement invariance across inclusive and exclusive educational settings for students with LD.

Keywords: Measurement invariance, Rasch model, item parameter recovery, limited information estimation, learning disabilities

[1] *Correspondence concerning this article should be addressed to:* Jörg-Henrik Heine | Technical University of Munich, TUM School of Education, Centre for International Student Assessment (ZIB), Arcisstr. 21 | D-80333 München, Germany; email: joerg.heine@tum.de

[2] Technische Universität Dortmund

[3] University of Wuppertal Germany & North-West University, Vanderbijlpark, South Africa

[4] Universität Bielefeld

Introduction

In Germany and many other countries, the construct of Learning Disabilities (LD) refers to children, who have significant academic difficulties in school and need additional special educational support for which neither other disabilities (e.g., sensory impairment, mental retardation, or emotional and behavioral disorders) nor lack of schooling can be found as a cause (Lloyd, Keller, & Hung, 2007). In almost all school systems, these children are labeled with LD to give them a legal right for additional assistance and support in school. However, the concepts of LD, the assessment procedures and the diagnostic criteria, as well as their interpretation, vary widely from country to country; but, they generally agree that general cognitive abilities, as measured by standardized IQ tests, are an important aspect. In the identification process of special educational needs (SEN), an intelligence test is often used combined with academic performance tests (Bundschuh & Winkler, 2014). In German speaking countries, a below average IQ outcome was considered the most effective diagnostic criterion of LD during the 1960s and 1970s because this was a general "objective" assessment of the cognitive performance of a child without a school reference (Grünke, 2004). One of the most used tests for this purpose is the Culture Fair Intelligence Test CFT 1-R (Weiß & Osterland, 2013). The CFT 1-R is a language-free intelligence test, constructed according to the classical test theory to measure basic aspects of intelligence for children aged from five to eleven years. The German adaption of the CFT 1-R provides standardized tests-scores also for students in special schools. The test especially measures fluid intelligence, the ability to understand and process complex information (Cattell, 1963). The concept of fluid intelligence should not be influenced or rather confounded by the language and the cultural background of a specific test taker. Thus, children with limited language skills in German should not be disadvantaged by the CFT 1-R. The test is a group test and has a satisfactory reliability ($r = .95$), particularly differentiating the lower levels of intelligence. Therefore, it is recommended as a diagnostic intelligence inventory for students with SEN (Büttner, 1984). When using intelligence tests such as the CFT 1-R for the purpose of diagnostic differentiation between subgroups that are solely defined by their test outcome, the issue of (strong) measurement invariance immediately arises. Especially in the case of diagnosing LD, an assumption of measurement invariance regarding a lower proficiency subsample is a crucial assumption to be verified (Schwab & Helm, 2015). Local distortions from the general assumption of between group measurement invariance are discussed in the literature via the term differential item functioning (DIF); see e.g., Holland (1993) for a general overview and Zwick, Donoghue, and Grima (1993); Zwick (2012) for a summary of principles of DIF detection in the framework of student assessment. Furthermore, some classical reviews of different DIF detection methods are given for example by Rudner, Getson, and Knight (1980), Mellenbergh (1982) and Osterlind (1983), as well as newer developments given by Khalid and Glas (2014) and Lee and Geisinger (2015). The detection of DIF itself is usually related to the application of models from Item Response Theory (IRT – G. Fischer & Molenaar, 1995; Millsap, Gunn, Everson, & Zautra, 2015). Unfortunately, such IRT-based DIF-analyses in general must be based on sufficient sample sizes for both subgroups to be tested against invariant outcome measurement. This is true to greater extent when parametric, specifically iterative and likelihood based, IRT methodology is to be applied (Zwick, 2012). The general challenge is to achieve stable model parameter estimates against the backdrop of lacking data or rather small sample sizes (Heine & Tarnai, 2015). Such small dataset usually arises when examining marginal groups such as highly gifted students or students with SEN.

Assessing general intelligence of students with LD

The use of intelligence tests in general, and specifically the use of the CFT, has a long tradition of diagnosing students with SEN. Based on its outcome, decisions are made regarding the future academic career of the student, special learning support, and recommendations to attend special schools (Heimlich, Lotter, & März, 2005; Schuck, 2011). Furthermore, the CFT is often used in research focusing on students with LD (e. g. Hövel, Hennemann, Casale, & Hillenbrand, 2015; Gebhardt, Schwab, Krammer, & Gasteiger, 2012; Sonntag, 2010; Voß et al., 2014). The CFT was used, for example, in the first large studies on the effectiveness of special schools and inclusive schools in studies in Switzerland (Haeberlin, Bless, Moser, & Klaghofer, 1998) and in Germany (Tent, Witt, Bürger, & Zschoche-Lieberum, 1991). These studies showed positive results towards inclusion of students with LD that were similar to recent studies (Kocaj, Kuhl, Kroth, Pant, & Stanat, 2014; G. Lindsay, 2007). Since the research tradition of Alfred Binet, intelligence has been seen as an important indicator of future school development, and thus it serves as a criterion for deciding the future school career of students with LD (Bundschuh & Winkler, 2014). Specifically, the CFT 1-R is one of the most used tests in practice to identify LD. German students with LD are in general older in comparison to students without LD. This is due to delayed school enrolment and decelerated schooling career—the first three years of special schools covers standard schools' first two years (Biewer, 2001). In secondary school, students with LD learn basic mathematical skills that are normally taught to regular students in primary school (Gebhardt, Zehner, & Hessels, 2014). In Germany, students with severe disabilities are more likely to attend special schools (Gebhardt, 2015), and students with LD in special school settings generally have a lower IQ and lower academic performance compared to students with LD in inclusive settings (Kocaj et al., 2014; Myklebust, 2002). Therefore, it is unclear whether students with LD in both educational settings can be considered part of the same population based on measurement invariance and other psychometric properties of the CFT 1-R. However, the test is constructed based on classical test theory as well as existing verifications of its psychometric properties in the field of LD. In the framework of classical test theory, the CFT 1-R shows good reliability and validity, and it considers students with LD in special schools in its latest revision. Admittedly, a proof of the reliability and measurement invariance in the framework of IRT is still missing for students with special needs who are educated in inclusive settings. Moreover, when measuring latent variables such as intelligence, the application and assumptions of classical test theory and the concept of true scores may only represent an operationalist view of the measurement process, but not an underlying formal structure that relates test scores to the hypothesized latent trait (Borsboom, 2005, p. 49). The later assumptions are better fulfilled in latent-variable measurement models, primarily used in educational testing, which came to be known as Item Response Theory (IRT) models. In general, these models provide a useful theoretical and verifiable model for the emergence of observed manifest student responses based on an assumed latent trait-intelligence in case of the CFT 1-R. Specifically, the Rasch model (RM) is not only useful for modeling student's responses in performance tests, such as the CFT 1-R, but is also a necessary prerequisite for summative scaling when the number of correct items is used for individual diagnostic purposes (Kubinger, 2005). However, studies for the CFT 1-R with regard to specific populations like students with special needs are still missing. Therefore, the present study aims at examining the psychometric properties of the CFT 1-R for students with LD in inclusive settings and special schools in the framework of IRT.

A psychometric Item Response Theory for practical applications

Scaling

As pointed out in the above section, there is a lack of research concerning the psychometric properties of the CFT 1-R. This applies to two key problems: first, whether the implicit assumption of measurement invariance holds true across students with LD in both inclusive schools and special schools and second, the need to analyze the CFT 1-R in the framework of Item Response Theory (IRT). In this sense Kuhn, Holling, and Freund (2008) analyzed and judged the quite similar CFT 20 R (Weiß, 2008) to show good psychometric properties and measurement invariance for highly gifted students in comparison to a student population with normally distributed general intelligence. This investigation also showed strong measurement equivalence with regard to the two subgroups of highly skilled students and students with average skill levels. Kuhn et al. (2008) had to fall back on introducing a second model parameter by applying the 2-PL model to fit their data. Although interesting from the perceptive of the mere data analyst, who is mainly interested in a sophisticated and precise explanation of the data generating process, such a procedure does not necessarily fulfill the needs of practical applications, where (unweighted) sum scores are used for diagnostic purposes on an individual level.

The core idea of any psychometric item response model is to make the nature of the empirically discovered data matrix explainable via a formal, mathematical link of different assumed model parameters. More precisely, the binary logistic test model, originally introduced by Georg Rasch (1960), formalizes the response probabilities of a person for each of two predetermined response categories (e.g., correct = 1 and false = 0) based on two (model) parameters, σ for the item difficulty and θ for the person ability. The Rasch model (RM) holds a special unique advantage over other IRT models, which, however, share some general properties of the RM. By parsimoniously introducing only two types of model parameters, it gives the basic conditions for a fair and objective comparison of both items and persons relating to the modeled latent variable. In short, the term ‘specific objectivity’ of the estimation as introduced by Rasch (1964, p. 17) means that at any point on the latent continuum – that is at any degree of trait level – all items share the same kind of measurement quality as represented by their difficulty estimates on a common scale. In other words, specific objectivity demands that the item difficulty hierarchy is relative invariant across person abilities (Fisher, 2010). As discussed in Heine and Tarnai (2015), specific objectivity can be seen as a prerequisite of scientific inferences in general (see also Rasch, 1977). However, specific objectivity is especially given when applying the RM, when scaling response data (e. g. G. H. Fischer, 1988; Scheiblechner, 2009). With regard to specific objectivity Irtel (1987) mentioned that next to the Rasch model also the ordinal independence model allows for specifically objective comparisons for psychodiagnostic measurement, but only on ordinal scale level. However, the ordinal independence model plays an important role for the principle foundation of the Rasch model (Irtel, 1987). If successfully applied to a dataset, the Rasch model implies an equally unweighted consideration of every test item contributing to the scale. This in turn might be seen as a prerequisite for the justification of the usage of item sum scores as a measure of trait. In contrast to the theoretical assumptions of the 2-PL model that implies a weighted summation of item scores, the manual of the CFT 1-R advises using an unweighted summation of item scores – as most test manuals do. Thus, unweighted unidimensionality of any psychometric scale should be a prerequisite for using the sum score in individual diagnostics (Wright, 1977). This is especially true when raw values are regarded as interval-scaled (or rather ratio scaled) and used in the evaluation with the CFT 1-R for the purpose of diagnosis on an individual level. Additionally, with the introduction of an item specific varying slope parameter such as in the 2-PL model, a particularly unfavorable consequence is that the items are no longer uniformly related to the ability parameter θ. In other words, the property of specific objectivity is abandoned in favor of a more flexible model adjustment. As a result, persons may be differentially rated on the latent trait continuum θ, depending on the parameters of the specific item, i.e. the slope of the Item Characteristic Curve (ICC). In turn, when using the 2-PL model for scaling, a certain weighted sum score of the individual responses to the items should rather serve as an estimator for the person’s characteristic expression (Sijtsma & Hemker, 2000). The item slopes then represent not the difficulties, but the different weighting coefficients of the items (Rost, 2004). In favor of models with more than only one item parameter (e.g. 3- and 4-PL models), it must be noted that in general the more parameters such models imply the better they fit the empirical data (e. g. Aitkin & Aitkin, 2011, p. 42). While all scientific models in general imply some kind of pragmatic simplification of empirical data (e. g, Stachowiak, 1973), the question of usefulness of a psychometric model should be a more important criterion for the selection of a model of proficiency scaling (see also Box, 1979, p. 202). Moreover regarding misfitting items due to hidden multidimensionality, Crişan, Tendeiro, and Meijer (2017) recently showed that applying a unidimensional scaling model nevertheless leads to unbiased θ parameter estimates.

Because we are aiming to verify the approach of using sum scores in practical settings for diagnostic purposes as in the CFT 1-R, the usefulness of a unidimensional specific objective scaling model such as the RM is essential for the present study.

Method of parameter estimation

In the history of psychometric research, several parameter estimation techniques for applying IRT-models have been proposed. In the context of student assessment and social sciences overall, the three main, most prevalent types are Joint Maximum Likelihood (JML), Conditional Maximum Likelihood (CML) and Marginal Maximum Likelihood (MML) estimation (see Heine, Sälzer, Borchert, Siberns, & Mang, 2013). Linacre (1999) classifies the parameter estimation methods within IRT more broadly into iterative and non-iterative approaches. The pairwise approach used in the present paper falls into the second (non-interative) class of techniques for parameter recovery see Heine and Tarnai (2015), for a more detailed introduction and discussion of the principle of pairwise item parameter recovery in the framework of IRT).

A common principle in all of the other ML-based methods is that they find the model parameters as the margins of the empirical data by maximizing their Likelihood in an iterative process – usually a Newton-Raphson type (Linacre, 2004). Another, perhaps more practical, commonality of those three iterative estimation methods is that they all require usually quite large sample sizes, or rather should only be seriously applied on larger datasets. Such datasets with sufficient sample sizes are prevalent in international educational surveys like PISA, TIMMS and others. With a sufficient sample size, such ML-based methods usually result in consistent parameter estimates. With regard to CML estimates, Linacre (2004) argued that consistency and unbiasedness holds only when extreme person scores (zero and perfect response vectors) are excluded from data contributing to the likelihood, which is to be maximized. Moreover, the consistency of MML-estimates relies heavily on the distributional assumption of normality of the trait to be estimated based on the underlying sample (e. g. Rost, 2004). In turn when scaling marginal groups such as SEN student samples with ML-based methods, (1) optimal sample size requirements, which may fulfill the assumption of normality, are often not met and (2) extreme response vectors are more likely to occur. In line with this argumentation, Andrich and Luo (2003) showed that because of low category frequencies – likely due to small sample sizes – the estimates of the corresponding item parameters turn out to be unstable.

A standard ML-based estimation method is full information maximum likelihood (FIML) via the expectation-maximization (EM) algorithm (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988). As stated by Forero, and Maydeu-Olivares (2009), the term full information is derived from the principle of using the full response pattern based information when estimating the model parameters. Specifically, the main problem with the assumption of a full information approach, in connection with small sample sizes, lies in the rather unrealistic assumption of a data driven model definition based on a full set of response pattern. For example, estimating the 1-PL model for a scale of 15 binary items (the number of items in only one of the CFT 1-R subscales) would imply the theoretical assumption of $c = m^k = 215 = 32,768$ cells [1] , or different response patterns, to fulfill the asymptotic requirements for sufficient estimation and statistical inference on the model to be fitted. The asymptotic efficiency of estimates based on the FIML approach is that in theoretical samples approaching infinite size, no other estimator yields parameter estimates with smaller variances (Forero & Maydeu-Olivares, 2009). Conversely as expressed by the relationship between sample size n and model size c (i.e., the fraction of number of observations and number of cells n/c), the empirical type I error rates of inferential model fit-statistics (e.g., Pearson's χ^2) tend to become inaccurate with increasing sparseness of the data (Maydeu-Olivares & Joe, 2006). To overcome such problems with inferential model testing against the backdrop of sparse contingency tables, Maydeu-Olivares and Joe (2005) proposed the use of limited information methods (LI) for estimation and model testing which use only univariate and bivariate information (see also Maydeu-Olivares, 2001; Bolt, 2005; Maydeu-Olivares & Joe, 2006; Maydeu-Olivares, 2006; Joe & Maydeu-Olivares, 2010). Maydeu-Olivares and Joe (2005) have proved that Pearson's full information χ^2-based test statistics can be seen as special cases of a family of LI test statistics. Furthermore, they investigated the asymptotic distribution of full-information test statistics (as Pearson's χ^2) based on parameter estimates preserved by LI procedures and showed that these methods result in superior and stable estimates when sample sizes are limited (Maydeu-Olivares & Joe, 2005). These LI methods also parallel the least-square estimation methodology often used in Structural Equation Modeling (Bollen, 1996).

The *pairwise* procedure and the resulting least-square (item) parameter estimates used in this paper can be seen as LI- estimators because they use only bivariate information of the pairwise item response frequencies (see e. g. Millsap & Maydeu-Olivares, 2009, p. 194). For the purpose of model testing in the present paper the parameter estimates based on the pairwise (limited information) principle were used to calculate different (established) model fit-statistics – see method section below. This general principle of the LI procedure is described in (e. g. Maydeu-Olivares, 2015, p. 113), including a description of the derivation of full information χ^2-test-statistic as a special case of a limited information χ^2-test-statistic. The pairwise limited information approach can be seen as being part of a more general theory of composite (quasi or pseudo) likelihood approaches (B. G. Lindsay, 1988; Varin, 2008; Varin, Reid, & Firth, 2011). One may argue that the application of such approach should be restricted to situations when the full likelihood is computationally unmanageable or very complicated, due to complex models; which might not be an issue when applying a rather sparse model like the Rasch model. However, there can be several other reasons for the use of such partial likelihood approaches. As already pointed out by Cox (1975) such reasons include aspects like the reduction of dimensionality in presence of nuisance factors – e.g. lack of

[1] with c being the number of cells in the multidimensional contingency table, with m equals the number of response categories (equally for all items) and k equals the number of items.

distributional assumptions with regard to normality due to censored data, as in marginal selective samples – and the striving for robustness (e. g. Cox & Reid, 2004) in parameter estimation.

Research question

The present research addresses two main research issues: substantive and methodological issues. First, regarding substantive issues there is little research investigating the psychometric properties of the CFT 1-R for students with LD in inclusive settings and special schools. Yet, Heydrich, Weinert, Nusser, Artelt, and Carstensen (2013) argued for the necessity of inclusion of students with SEN during largescale assessments.

Therefore, this article's purpose is to examine whether the CFT 1-R has sufficient, and furthermore, desirable psychometric properties within the framework of Item Response Theory (IRT). Specifically, the usability of the test for students with LD in different educational settings was investigated (special schools vs. inclusive settings). These students are on average older than the target population that was taken as a basis for the latest norming revision of the test. When applying the CFT 1-R, sound psychometric properties are a fundamental prerequisite for reliable measurement and fair comparisons between any kinds of subgroups, especially so in heterogeneous surroundings that students with LD find themselves in. Based on the theoretical considerations in the previous sections, it is worth questioning whether the achieved test results of any groups of students with LD are comparable from a psychometric IRT -based perspective. The one- dimensional scalability of the CFT 1-R was tested by applying the 1-parameter logistic test model (Rasch model). To do so we scaled the whole 45-item pool as well as the three subscales separately, comprising 15 items each. Global model tests for each scaling approach were performed in order to support the hypothesized good psychometric properties of the CFT 1-R, which were based on earlier findings using methodology from classical test theory. Alongside the testing of a one dimensional IRT model, the assumption of measurement invariance across different subgroups was investigated. To do so, the present sample was divided into subsamples of students with LD in different educational settings (special setting vs. inclusive setting). Additionally, two other commonly applied principles of subsample splitting were examined: median-split and split by gender. With regard to these subgroups any local model violations were examined via tests for differential item functioning (DIF) to detect specific model violations (e. g. Glas & Verhelst, 1995).

Second, regarding research methodology, the question of applicability of the method used for the estimation of the model parameters and the different global and local fit indices for model fit is also important. The above-mentioned, non-iterative pairwise LI method approach was used in the present study. In addition to being computational simple and speedy (see e. g. G. H. Fischer, 1970), this principle handles sparse contingency tables in a theoretically straightforward manner (Heine & Tarnai, 2015; Wright & Masters, 1982;

G. H. Fischer, 1970; Choppin, 1968).

Because such sparse data often arises in research related to students with SEN, it is important to investigate whether the proposed pairwise methodology is a viable alternative to ML-based IRT estimation techniques, which typically require larger sample sizes. Instead of parameter estimation relying upon the full set of possible response patterns, the pairwise approach uses only bivariate item association information. Therefore, in the resulting model, additional test statistics therefore represent limited information fit-statistics. In the present paper we examined how such indices contribute to inferences with regard to model fit based on small sample sizes when applying a rather tight scaling model (1-PL model) in comparison to more relaxed models (e.g. 2-PL model) estimated via ML- based technique.

Method

Measure

The German version of the CFT-1-R (Weiß & Osterland, 2013) is a partial adaptation and revision of the 'Culture Faire Intelligence Tests – Scale 1' introduced by Cattell (1950). It is based on Cattel's (1941; 1963) theory of fluid and crystalized Intelligence. The full test comprises 150 items according to six subscales that are named as substitution (UT1 – 75 Items), labyrinths (UT2 – 15 Items), similarities (UT3 – 15 Items), series (UT4 – 15 Items), classification (UT5 – 15 Items) and matrices (UT6 – 15 Items). Each of the six subscales is related to a specific cognitive task, each of which contributes a varying amount to fluid intelligence. In the present study, only three of the six subscales (45 items in total) were administered in order to keep cognitive load at a minimum level. However, the three subscales give a sufficient coverage of the basic aspects of general intelligence (Weiß & Osterland, 2013). They included series (UT4; completing a series of numbers), classification (UT5; distinguishing one dissimilar figure among four other similar ones) and matrices (UT6; choosing a figure to complete the pattern). A more comprehensive, formal description and theoretical foundation of these three scales is presented in Weiß and Osterland (2013).

Sample and data

Students with LD in the fifth grade were administered the short version of the CFT 1-R (45 Items) comprising of three different cognitive tasks – series (UT4), classification (UT5) and matrix (UT6) – as part of a more general research project related to students in an inclusive educational setting (BilieF – Wild, Lütje-Klose, Schwinger, Gorges, & Neumann, 2017). The initial sample in this project comprised $n = 316$ students. However, for altogether 11 students no response data related to the three CFT 1-R subscales were available. These respective cases had to be excluded from further analysis, either due to (CFT 1-R) unit none-response (5 students) or data entry errors (6 students). Thus the remaining total sample for this study comprised $n = 304$ students. Out of these 138 students attended special schools while 166 students attended inclusive school settings with non- SEN classmates. The total sample included 58.8 % male students, which is in line with the gender ratio of SEN students in Germany (Hasselhorn & Autorengruppe Bildungsberichterstattung, 2014).

Approximately half (51.0 %) of students from the total sample were aged 12 years, followed by 32.9 % aged 11, 14.8 % at age of 13 years and only 1.3 % were at age of 14 years at the point of testing. As the CFT 1-R offers only norm tables up to age of 11 years and 11 months (11;00 - 11;11), those were taken to compute T-values for the whole sample to give some first descriptive impressions of the distribution of general intelligence within the sample. The total sample reached an average T-value of $M = 53.07$; $(SD = 9.08)$ with a range of $M = 31.00$ for the lower bound and $M = 72.00$ for the upper bound (see figure 1).

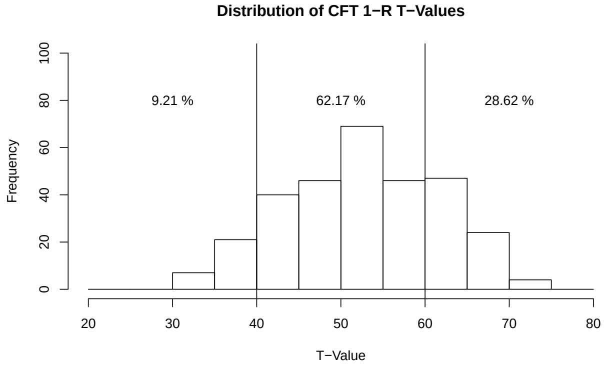

Figure 1:
Distribution of T-values from the CFT 1-R for $n = 304$ students with LD using Norms for students aged from 11;00 up to 11;11 years attending special schools (see Weiß & Osterland 2013, p. 92, table D1).

Even though the older students should theoretically achieve higher scores when applying norms related to students with age 11 years, they did not show better results than the younger part of the sample. The Spearman's correlation between the T-value and student's age revealed at $\rho = -0.04$. Students with SEN in inclusive educational settings showed significantly better T-values ($M = 55.49; SD = 8.66$) than students attending special schools ($M = 50.16; SD = 8.72$); ($t = -5.31, df = 302, p = .000$). The proportion of missing responses over all items for the present sample ranged from 0 %, items 4 (UT4) and 5 (UT5) to 11.8 %, item 15 (UT4). During parameter estimation, any missing values were treated as missing data points and thus not recoded as wrong answers.

IRT Analyses

IRT analyses were conducted within the *R* statistical environment (R Core Team, 2017), using the package `pairwise` (Heine, 2017). We choose this package because it implements a stable, non-iterative method for item parameter recovery, even under sparse data conditions, like in our study. In a first run, we applied a one-dimensional scaling approach to the total set of 45 items (from the three subscales). This aligns with the implicit assumption from the principle given in the CFT 1-R manual, which requires an invariant, one-dimensional proficiency continuum to additively combine the single items of the three subscales. To check for any sub dimensionality, possibly resulting from the theoretic foundations with regard to the three subscales, a Rasch residual factor analysis (RFA – Wright, 1996; Linacre, 1998) was performed, as well as three separate one-dimensional scaling procedures for each dimension to accomplish more differentiated analyses for each sub scale of the CFT 1-R. For these scaling approaches, both global and local model fit measures were calculated. For global model checks, the likelihood ratio based model-test (Andersen, 1973) was conducted using the three splitting criteria of gender, educational setting of schooling and median-split. Using the identified model parameters, weighted mean square item fit-statistics – *INFIT* and *OUTFIT* – (Wright & Masters, 1982) were evaluated to detect any local model violations. We again conducted those checks for the overall scaling approach and for scaling each of the three subscales separately.

In order to further test the respective model-fit on item level, analyses of differential item functioning (DIF) were carried out using test statistics based on the pairwise estimates, which can be also used based on CML or MML estimates (e. g. Glas & Verhelst, 1995). For the analysis of DIF effects across LD, gender and median-split subgroups, the item parameters were calculated based on the sub samples respectively and then compared to each other. Specifically, the test statistic S_i, as implemented in the *R*-package `pairwise` was evaluated on item level. This item fit statistic is also (perhaps misleadingly) named as 'Wald test' in other *R*-packages. According to (G. H. Fischer & Scheiblechner, 1970), the S_i statistic is defined in the following equation (1) given below (see also equation (3) in van den Wollenberg, 1982, p. 124).

$$S_i = \frac{\widehat{\sigma_i{}^{(1)}} - \widehat{\sigma_i{}^{(2)}}}{\sqrt{\left(SE^{(1)}_{\sigma_i}\right)^2 + \left(SE^{(2)}_{\sigma_i}\right)^2}} \tag{1}$$

Where $\widehat{\sigma_i{}^{(1)}}$ is the estimate of the item parameter of subsample one, $\widehat{\sigma_i{}^{(2)}}$ the estimate of the item parameter of subsample two and $(SE^{(1)}_{\sigma_i})$ and $(SE^{(2)}_{\sigma_i})$ are the respective standard errors. In (G. H. Fischer, 1974, p. 297) the resulting test statistic (as defined above) is labeled with Z_i as it is asymptotically normally distributed. Contrary to the 'Wald-type' test statistic W_i which was derived by Glas and Verhelst (1995) from the (general) χ^2 distributed test of statistical hypotheses concerning several parameters, as introduced by Wald (1943).

To further evaluate the relative model fit for the rather restrictive scaling model (Rasch 1- PL model) in comparison to the more relaxed 2-PL model, an alternative scaling approach using the *R*-package `TAM` (Robitzsch, Kiefer, & Wu, 2017) was performed. Contrary to the pairwise approach, the package `TAM` implements an ML-based approach for parameter estimation relying on Marginal Maximum Likelihood. We evaluated the relative global model fit, by inspecting the respective information theoretic indices' AIC (Akaike, 1974) and BIC (Schwarz, 1978). We calculated the person estimates for the CFT 1-R outcomes for both – based on the 1-PL modeling approach and based on the more differentiated 2-PL model. Lastly, we examined the practical consequences on individual person estimates when choosing between the tight scaling model (1-PL model) and the more differentiated model (2-PL model). Correlations were calculated for the CFT 1-R sum scores, pairwise (1-PL) WLE estimates and the TAM (2-PL) WLE estimates.

Results

Overall scaling

The results from applying the one-dimensional 1-PL model (Rasch model) to the total 45 item set for all 304 students revealed a far good scalability. WLE reliability reached an acceptable value of $r_{WLE} = .89$. Looking at the wright map, the test showed a quite good targeting for the sample of students with LD - apart from three items – 1 (UT6), 4 (UT5) and 2 (UT6) – which were too easy for the present sample (see figure 2). The global model test (Andersen, 1973), confirmed the model assumption of a one-dimensional scaling model. When testing our central hypothesis, no significant deviation from the model assumption was found for dividing by school setting ($\chi^2 = 31.42; df = 89; p = 0.99$). This was also true with a median-split ($\chi^2 = 108.58; df = 89; p = 0.08$) and when splitting by gender ($\chi^2 = 72.02; df = 89; p = 0.91$).

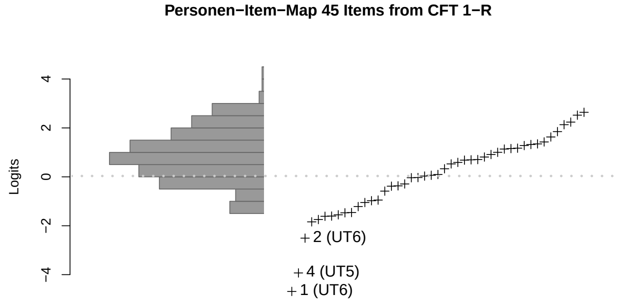

Figure 2:
Wright-Map for 45 Items (plus signs on the right panel) from the CFT 1-R and $n = 304$ students with LD (histogram of trait distribution on the left panel).

The three badly targeted items (see figure 2) were inspected with regard to their category frequencies. Based on the total sample ($n = 304$), item 4 (UT5) had a 98 % correct rate, item 1 (UT6) had 99 % correct, and item 2 (UT6) had 96 % correct. Similarly high percentage correct rates were found when comparing school settings, with even higher rates for students in inclusive settings. Moreover, for item 1 (UT6), a constant column vector of correct answers from all of the n = 166 students in inclusive school settings was found. Based on these practical grounds those three items were excluded from further analysis, but the WLE reliability remained at an acceptable value of $r_{WLE} = .89$ for the test with the reduced item set.

With regard to a graphical over all global model fit test using the split criteria school setting, which is related to our main research question, the data including both subgroups, can be adapted to the model assumptions sufficiently when eliminating the three items mentioned above (see figure 3).

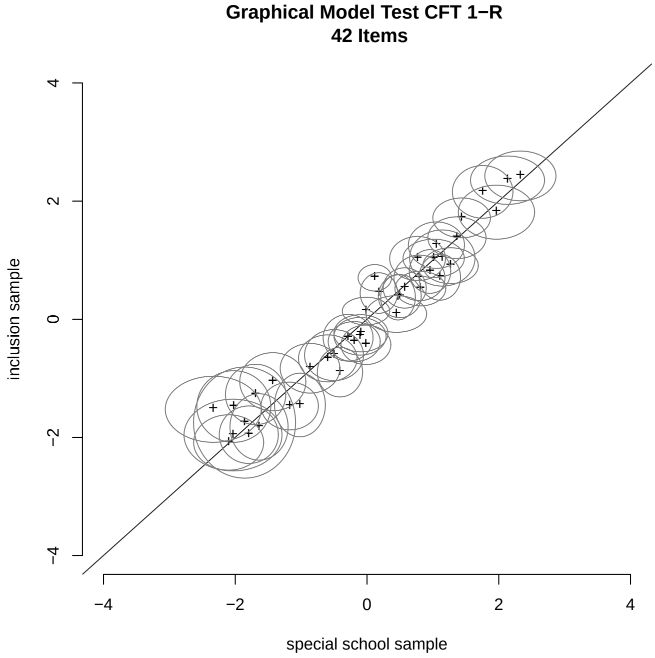

Figure 3:
Graphical model test with split criteria “school setting” for 42 remaining items from the CFT 1-R and $n = 304$ students with LD; ellipses represent confidence intervals for item parameter point estimates.

To confirm the results of the global model tests and better understand any possible causes of local model deviations, further analyses on item level were conducted. Also for the one-dimensional scaling approach, the results from the analysis on item level with the reduced item set are quite in line with the above findings from the global model tests. In summary only four items (Item 10 from UT 4 and Items 9, 11 and 14 from UT 5) show somewhat unambiguous deviations from the model assumption when simultaneously taking into account the results from the rout-mean-square statistics (*INFIT* and *OUTFIT*) and the results from the Fischer-Scheiblechner test in any of the three splitting conditions (see table 1).

Table 1:
Over all one dimensional scaling according to the Rasch model - Tests for local model deviations.

								school setting		gender		median	
Item	χ^2	*df*	p_{χ^2}	OUT_{MSQ}	OUT_{zSTD}	IN_{MSQ}	IN_{zSTD}	S_i	p	S_i	p	S_i	p
1 (UT4)	242.43	302	1.00	0.85	−0.65	1.03	0.32	0.95	0.34	1.54	0.12	−0.23	0.82
2 (UT4)	301.99	302	0.49	1.04	0.26	1.01	0.18	1.37	0.17	−1.00	0.32	−0.12	0.91
3 (UT4)	298.86	302	0.54	1.03	0.38	1.04	0.82	0.38	0.71	−0.72	0.47	0.65	0.52
4 (UT4)	218.71	303	1.00	0.76	−0.82	0.97	−0.24	0.46	0.64	−0.17	0.86	−0.18	0.86
5 (UT4)	304.48	301	0.43	1.05	0.59	0.96	−0.68	−0.05	0.96	0.85	0.40	−0.24	0.81
6 (UT4)	257.26	300	0.97	0.90	−0.77	0.98	−0.36	0.74	0.46	0.37	0.71	−0.42	0.68
7 (UT4)	254.13	300	0.97	0.89	−1.37	0.93	−1.29	1.44	0.15	−0.06	0.96	−1.25	0.21
8 (UT4)	286.68	291	0.56	1.03	0.34	1.02	0.38	1.03	0.30	0.10	0.92	0.56	0.57
9 (UT4)	263.86	290	0.86	0.95	−0.25	1.01	0.17	0.25	0.81	1.10	0.27	−0.23	0.82
10 (UT4)	203.85	285	1.00	0.76	**−2.39**	0.85	**−2.84**	−0.53	0.60	1.81	0.07	−2.72	0.01
11 (UT4)	291.21	286	0.40	1.06	0.67	1.03	0.60	0.38	0.71	−0.20	0.84	0.04	0.97
12 (UT4)	256.91	282	0.86	0.95	−0.53	0.95	−0.84	1.41	0.16	1.36	0.17	−0.42	0.67
13 (UT4)	310.45	278	0.09	1.16	1.64	1.05	0.84	−2.50	0.01	−0.72	0.47	0.24	0.81
14 (UT4)	268.94	271	0.52	1.03	0.31	0.99	−0.11	0.41	0.69	−1.19	0.24	0.60	0.55
15 (UT4)	327.10	267	0.01	1.27	1.62	0.97	−0.35	−0.54	0.59	2.06	0.04	0.29	0.77
1 (UT5)	298.56	300	0.51	1.04	0.22	1.00	0.08	−1.37	0.17	1.54	0.12	0.98	0.33
2 (UT5)	275.17	300	0.85	0.96	−0.03	0.96	−0.27	−1.57	0.12	−0.82	0.41	0.49	0.62
3 (UT5)	243.15	296	0.99	0.86	−0.32	0.99	−0.03	−0.14	0.89	−1.26	0.21	−0.10	0.92
5 (UT5)	382.61	303	0.00	1.30	1.20	0.98	−0.12	−1.29	0.20	−1.06	0.29	1.24	0.22
6 (UT5)	321.00	297	0.16	1.12	0.49	0.95	−0.37	−0.25	0.80	−0.86	0.39	1.76	0.08
7 (UT5)	295.76	299	0.54	1.03	0.21	1.01	0.10	−1.15	0.25	−1.86	0.06	0.88	0.38
8 (UT5)	359.11	294	0.01	1.26	1.43	1.05	0.68	−0.11	0.91	−0.14	0.89	2.23	0.03
9 (UT5)	369.20	294	0.00	1.30	**2.91**	1.25	**4.45**	−0.89	0.37	0.41	0.68	3.12	0.00
10 (UT5)	308.45	294	0.27	1.09	0.63	1.05	0.68	0.36	0.72	−2.05	0.04	0.84	0.40
11 (UT5)	391.93	293	0.00	1.38	**4.06**	1.29	**5.01**	−0.88	0.38	−0.43	0.67	3.34	0.00
12 (UT5)	297.39	293	0.42	1.06	0.60	1.07	1.32	1.31	0.19	−0.35	0.73	0.34	0.73
13 (UT5)	270.18	282	0.68	1.00	0.03	1.02	0.35	0.39	0.70	−1.52	0.13	0.05	0.96
14 (UT5)	351.22	284	0.00	1.28	**1.98**	1.10	1.28	−1.17	0.24	2.40	0.02	2.22	0.03
15 (UT5)	271.10	284	0.70	1.00	−0.01	1.01	0.20	−0.75	0.45	1.48	0.14	0.04	0.97
3 (UT6)	184.74	301	1.00	0.66	−1.01	0.99	−0.04	−0.03	0.98	−0.34	0.73	−0.23	0.82
4 (UT6)	192.54	302	1.00	0.68	−1.07	0.95	−0.34	0.36	0.72	−0.58	0.56	−0.28	0.78
5 (UT6)	282.71	302	0.78	0.98	−0.09	0.96	−0.61	1.73	0.08	−0.64	0.52	−0.55	0.59
6 (UT6)	253.28	302	0.98	0.88	−0.93	0.96	−0.71	0.11	0.92	−0.58	0.56	−0.50	0.62
7 (UT6)	246.43	302	0.99	0.86	−1.19	0.94	−1.11	0.63	0.53	0.20	0.84	−0.62	0.54
8 (UT6)	230.74	302	1.00	0.81	−1.71	0.90	−1.85	0.44	0.66	2.61	0.01	−1.39	0.17
9 (UT6)	257.74	302	0.97	0.90	−0.84	0.98	−0.37	1.39	0.17	2.35	0.02	−0.63	0.53
10 (UT6)	267.35	302	0.93	0.93	−0.85	0.94	−1.24	0.18	0.86	0.79	0.43	−0.63	0.53
11 (UT6)	262.19	302	0.95	0.91	−1.10	0.94	−1.21	0.43	0.67	1.19	0.24	−0.45	0.65
12 (UT6)	225.94	301	1.00	0.79	**−2.69**	0.84	**−3.25**	0.59	0.56	−0.97	0.33	−2.01	0.04
13 (UT6)	288.34	300	0.68	1.00	0.06	0.98	−0.43	−0.07	0.95	0.37	0.71	0.22	0.82
14 (UT6)	318.83	301	0.23	1.10	0.66	1.02	0.26	−0.21	0.83	−1.93	0.05	−0.10	0.92
15 (UT6)	320.05	301	0.22	1.11	1.01	1.12	1.86	−0.95	0.34	−0.44	0.66	0.77	0.44

Notes. Items UT6_1, UT5_4 and UT6_2 were omited from scaling; $p_{\chi^2} = p$-value for pearson χ^2-square test; S_i = test statistic for Fischer-Scheiblechner test, p = p-value for Fischer-Scheiblechner test, all $p < .05$ in bold face; OUT_{MSQ} = outfit-mean-square statistic (*OUTFIT*); IN_{MSQ} = infit-mean-square statistic (*INFIT*); $OUT_{zSTD} = z$-standardized outfit statistic (*OUTFIT*); $IN_{zSTD} = z$-standardized infit statistic (*INFIT*), values above 1.964 or below -1.964 in bold face.

In order to evaluate the relative model-fit of the 1-PL model compared to a more complex model, an alternative scaling approach applying the 2-PL model by using the R-software `TAM` (Robitzsch et al., 2017) was performed. Overall, the results of such comparison indicate no severe deviations of model fit between the two respective scaling models. On person level the WLE reliability for the 2-PL model reached a similarly good value of $r_{WLE} = .87$ ($r_{WLE} = .89$, for the pairwise approach). The inter correlations between the respective person estimates and the simple sum score (percent correct), as recommended in the CFT 1-R test manual to be used for individual diagnosis, reached both an almost perfect value of $r = .98$. Relative global model fit, as indicated by the information theoretic indices' AIC (Akaike, 1974) and BIC (Schwarz, 1978), both suggest a slightly better fit of the sparser Rasch 1-PL model (1-PL model: $AIC =$

$11782.71, BIC = 12091.23$; 2-PL model: $AIC = 12657.41, BIC = 12813.53$).

In the Rasch residual factor analysis (Linacre, 1998) to examine sub dimensionality, the theoretical assumptions of the three CFT 1-R subscales were upheld for the reduced overall 42 - item set – omitting items 1 (UT6), 4 (UT5) and 2 (UT6). The pattern of the item loadings upon the first main component roughly reflects the theoretically derived sub dimensionality of the CFT 1-R based on the sample of students with LD (see figure 4).

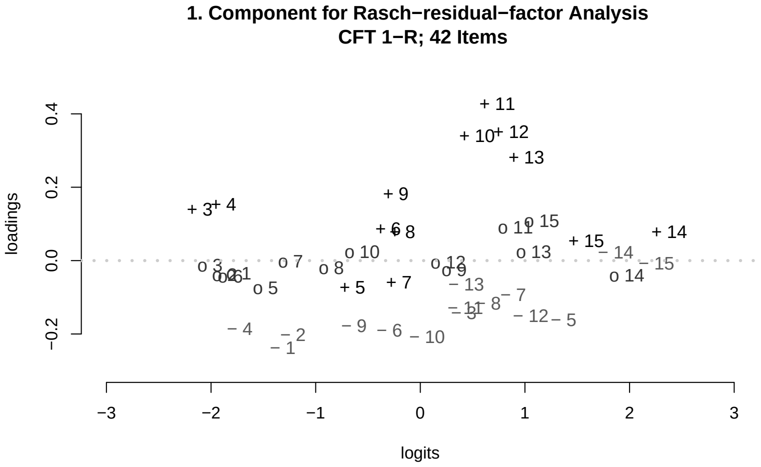

Figure 4:
First component from a Rasch residual factor analysis (Linacre, 1998), for 42 remaining items from the CFT 1-R and $n = 304$ students with LD; y-Axis: Loadings on the first main component; x-Axis: Item difficulty based on one dimensional Rasch scaling including 42 items; $+$ = UT6 - matrices, o = UT5 - classification, $-$ = UT4 - series.

The loadings of the residuals on the first main component of the Rasch residual factor analysis show a quite narrow range, $\lambda_{max} = .42$ to $\lambda_{min} = -.24$. Overall, the Rasch residuals from the items of sub scale UT6 (matrices) tend to show positive loadings on the first main component (except item 5 and 7), while those from the items of the sub-scale UT4 (series) show rather negative loadings (except item 14 and 15). The loadings of the Rasch residuals from the items of subscale UT5 (classification) cluster around zero (see figure 4). Based on the findings from the Rasch residual factor analysis, separate one-dimensional scaling approaches for each sub scale were performed – again omitting the three items mentioned above due to their insufficient distribution of category frequencies.

Analysis of subscales

In summary, the results for the more differentiated analyses for each of the three sub dimensions show that the model assumption holds for all three scales, based on the Andersen likelihood-ratio global model test. For the dimension series (UT4) including all items and for the dimension classification (UT5; excluding item 4) and the dimension matrices (UT6; excluding items 1 and 2), no significant model deviation was found when using the split criteria school setting, gender and median-split. However, for scale classification (UT5) the p-value for the likelihood ratio test is close to the level of significance (but still above $\alpha = .05$) when splitting the sample based on median. Table 2 gives an overview of the global model tests for each scale using the three different splitting criteria.

Table 2:
Andersen Likelihood Ratio tests for three CFT 1-R subscales.

CFT 1-R Subscale	Split criterion	χ^2	*df*	*p*
Series (UT4)	school setting	8.905	29	0.99
	median	35.505	29	0.19
	gender	18.581	29	0.93
Classification (UT5)	school setting	11.723	27	0.99
	median	37.987	27	0.08
	gender	30.879	27	0.28
Matrices (UT6)	school setting	7.453	25	0.99
	median	6.819	25	0.99
	gender	21.413	25	0.67

Notes: One dimensional Scaling according to the Rasch model for three subscales of the CFT 1 R respectively; Items 1 (UT6), 4 (UT5) and 2 (UT6) were omitted from scaling in the respective scale.

In line with the findings related to the global model tests, the majority of the items show no severe deviation from ideal model fit based on the respective graphical model test when splitting based on educational setting (see figure 5).

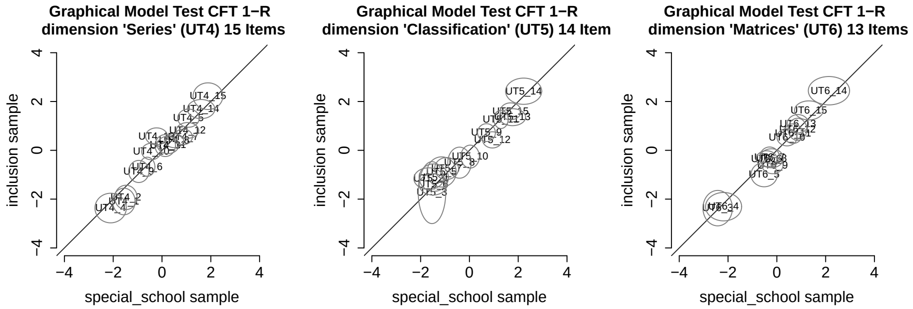

Figure 5:
Graphical Model-Test with split criterion educational setting for three subscales of the CFT 1-R based on one dimensional Rasch scaling; 15 Items for series (UT4 – left panel); 14 Items for classification (UT5 – middle panel); 13 Items for matrices (UT6 – right panel).

Considering all item fit-statistics (*INFIT*, *OUTFIT* and Fischer-Scheiblechner test) simultaneously, no consistent, distinct item misfit based on all fit-statistic was observed, but with regard to the scale UT6 (matrices), some items show a somewhat larger DIF based on the Fischer-Scheiblechner test under a median-split (see table 3).

Table 3:
Model fit on item level based on one dimensional scaling for three scales of the CFT 1-R respectively.

									school setting		gender		median	
Scale	Item	χ^2	*df*	p_{χ^2}	OUT_{MSQ}	OUT_{zSTD}	IN_{MSQ}	IN_{zSTD}	S_i	p	S_i	p	S_i	p
Series (UT4)	1	219.56	302.00	1.00	0.81	−0.57	0.97	−0.30	1.09	0.27	1.09	0.27	−0.01	0.99
	2	287.83	302.00	0.71	1.03	0.21	0.98	−0.14	1.22	0.22	1.22	0.22	0.40	0.69
	3	297.19	302.00	0.57	1.06	0.55	1.10	1.55	−0.25	0.80	−0.25	0.80	0.89	0.37
	4	178.05	303.00	1.00	0.67	−0.85	0.93	−0.53	0.65	0.51	0.65	0.51	−0.12	0.91
	5	296.14	301.00	0.57	1.06	0.49	1.00	0.01	−0.81	0.42	−0.81	0.42	0.17	0.86
	6	246.58	300.00	0.99	0.90	−0.54	0.99	−0.11	0.23	0.82	0.23	0.82	−0.46	0.64
	7	233.37	300.00	1.00	0.86	−1.21	0.96	−0.63	1.17	0.24	1.17	0.24	−0.65	0.52
	8	262.16	291.00	0.89	0.98	−0.13	1.06	1.03	0.42	0.68	0.42	0.68	1.35	0.18
	9	246.97	290.00	0.97	0.93	−0.28	1.03	0.44	−0.28	0.78	−0.28	0.78	−0.38	0.71
	10	181.45	285.00	1.00	0.72	**−2.12**	0.85	**−2.35**	−1.60	0.11	−1.60	0.11	**−2.53**	**0.01**
	11	269.19	286.00	0.76	1.02	0.21	1.08	1.28	0.10	0.92	0.10	0.92	0.70	0.49
	12	214.56	282.00	1.00	0.84	−1.33	0.91	−1.57	0.87	0.39	0.87	0.39	−1.19	0.23
	13	299.41	278.00	0.18	1.16	1.26	1.11	1.69	**−2.82**	**0.01**	**−2.82**	**0.01**	0.33	0.74
	14	372.30	271.00	**0.00**	1.45	**2.61**	1.04	0.58	−0.24	0.81	−0.24	0.81	1.48	0.14
	15	381.35	267.00	**0.00**	1.51	**2.35**	0.99	−0.10	−0.75	0.45	−0.75	0.45	1.36	0.17
Series (UT5)	1	238.28	300.00	1.00	0.92	−0.21	1.00	0.04	−0.81	0.42	−0.81	0.42	0.02	0.98
	2	220.70	300.00	1.00	0.87	−0.42	0.98	−0.15	−1.15	0.25	−1.15	0.25	−0.17	0.87
	3	263.48	296.00	0.91	1.02	0.17	0.97	−0.19	0.51	0.61	0.51	0.61	−0.07	0.95
	5	268.70	303.00	0.92	1.02	0.15	0.94	−0.56	−0.66	0.51	−0.66	0.51	1.28	0.20
	6	214.04	297.00	1.00	0.85	−0.45	0.91	−0.65	−0.16	0.87	−0.16	0.87	−0.01	1.00
	7	238.06	299.00	1.00	0.93	−0.30	0.96	−0.44	−0.47	0.64	−0.47	0.64	0.08	0.94
	8	257.36	294.00	0.94	1.01	0.08	0.99	−0.06	0.26	0.79	0.26	0.79	0.15	0.88
	9	309.20	294.00	0.26	1.18	1.93	1.17	**3.20**	−0.11	0.92	−0.11	0.92	0.80	0.42
	10	264.66	294.00	0.89	1.03	0.26	1.02	0.35	0.86	0.39	0.86	0.39	0.05	0.96
	11	328.47	293.00	0.08	1.25	**2.79**	1.16	**2.99**	0.08	0.94	0.08	0.94	0.65	0.52
	12	251.72	293.00	0.96	0.99	−0.10	1.02	0.42	1.65	0.10	1.65	0.10	−0.23	0.82
	13	220.23	282.00	1.00	0.91	−0.98	0.95	−0.85	1.60	0.11	1.60	0.11	−1.15	0.25
	14	291.88	284.00	0.36	1.16	1.15	1.03	0.36	−0.56	0.57	−0.56	0.57	0.41	0.68
	15	213.23	284.00	1.00	0.88	−1.29	0.92	−1.52	0.30	0.76	0.30	0.76	−1.34	0.18
Series (UT6)	3	122.26	301.00	1.00	0.59	−0.96	0.82	−1.26	−0.08	0.94	−0.08	0.94	**−22.14**	**0.00**
	4	133.85	302.00	1.00	0.63	−0.94	0.85	−1.09	0.22	0.83	0.22	0.83	−0.04	0.97
	5	308.62	302.00	0.38	1.20	1.08	1.12	1.61	1.50	0.13	1.50	0.13	**2.98**	**0.00**
	6	260.02	302.00	0.96	1.04	0.31	1.11	1.52	−0.26	0.79	−0.26	0.79	**2.03**	**0.04**
	7	285.51	302.00	0.74	1.13	0.84	1.15	**2.15**	−0.05	0.96	−0.05	0.96	**2.56**	**0.01**
	8	242.13	302.00	1.00	0.98	−0.06	1.01	0.13	0.30	0.77	0.30	0.77	1.94	0.05
	9	245.47	302.00	0.99	0.99	0.02	1.00	−0.04	1.58	0.11	1.58	0.11	**2.38**	**0.02**
	10	210.58	302.00	1.00	0.88	−0.92	0.94	−0.97	−0.43	0.67	−0.43	0.67	1.13	0.26
	11	200.54	302.00	1.00	0.85	−1.16	0.89	−1.75	−0.06	0.95	−0.06	0.95	0.64	0.52
	12	177.64	301.00	1.00	0.77	−1.71	0.84	**−2.56**	−0.06	0.95	−0.06	0.95	−0.26	0.79
	13	226.89	300.00	1.00	0.94	−0.38	0.97	−0.48	−0.75	0.46	−0.75	0.46	1.11	0.27
	14	343.20	301.00	**0.05**	1.32	1.16	1.06	0.73	−0.59	0.55	−0.59	0.55	**2.42**	**0.02**
	15	453.21	301.00	**0.00**	1.69	**3.20**	1.24	**3.36**	−0.81	0.42	−0.81	0.42	**5.55**	**0.00**

Notes. Items UT6_1, UT5_4 and UT6_2 were omitted from scaling; $p_{\chi^2} = p$-value for pearson χ^2-square test; S_i = test statistic for Fischer-Scheiblechner test, p = p-value for Fischer-Scheiblechner test, all $p < .05$ in bold face; OUT_{MSQ} = outfit-mean-square statistic (*OUTFIT*); IN_{MSQ} = infit-mean-square statistic (*INFIT*); $OUT_{zSTD} = z$-standardized outfit statistic (*OUTFIT*); $IN_{zSTD} = z$-standardized infit statistic (*INFIT*), values above 1.964 or below -1.964 in bold face.

However, item 3 of scale UT6 showed the biggest deviation in the median-split Fischer-Scheiblechner test ($S_i = -22.14, p = 0.00$). This finding can be traced back to a boundary problem in estimation. Because there was a 100 % correct rate for scores above the median (for UT6 $md = 8$), the result for the Fischer-Scheiblechner test for this item in the median-split cannot be interpreted in a sensible way.

In line with the above findings of the overall unidimensional scaling, the WLE reliabilities for the subscales reached acceptable values of $r_{WLE} = .80$ for UT4, $r_{WLE} = .69$ for UT5, and $r_{WLE} = .80$ (UT6), taking into account the shortened scales compared to the overall scaling approach.

Discussion

This study examined the psychometric properties of the CFT 1-R with a specific focus on measurement invariance between students with LD in inclusive settings and special schools. Since inclusive education is increasingly implemented across the globe, it is important to have valid measures of key variables like intelligence to analyze the initial conditions for different school settings as a control variable. The relevance of measurement invariance for group comparisons of latent variables was ignored in most test construction and research studies in recent decades. However, measurement equivalence is a requirement of group comparisons. Especially for a widely used instrument like the CFT 1-R, it is of vital interest and necessary to demonstrate that there is no severe bias when comparing different groups with this instrument.

Our first result indicated that in general the CFT 1-R is a fair test to students with LD. The reliability for the WLE estimates turned out to be relatively high. In comparison to Kuhn et al. (2008), we showed that the use of a 1-PL model is possible. Furthermore, the AIC and BIC information criteria for the comparison of the two scaling approaches (i.e., the 1-PL model and the 2-PL model) support the sparser 1-PL model. Moreover, because the practitioner using the test calculates and interprets unweighted individual sum scores in the field, the application of the 1-PL model is more valid with regard to applied settings. In general, the analyses reported above conform to the summative allocation rule of the individual item scores as a measure of intelligence for individual diagnosis.

Additional Rasch residual factor analyses were able to uncover some slight sub dimensionality of the total item set, which might be traced back both to the theory of constructing the CFT 1-R and the different instructions (i.e., cognitive tasks) necessary to solve the items of the respective subscale. The three item sets, series (UT4), classification (UT5) and matrices (UT6), could be clearly distinguished by their loading pattern on the first main component of the Rasch residual factor analysis. For subscale series (UT4), the items 14 and 15 turned out to be rather hard to solve and so might stand out from the typical loading pattern of sub dimensionality in the series (UT4) subscale. Both subscales matrices (UT6) and classification (UT5) tend to stick together with regard to their loading pattern as both might require similar cognitive processes for solving the items. In classification (UT5), students have to discover one dissimilar figure among four other similar ones, and in matrices (UT6), students had to find one figure to complete a homogeneous matrix pattern. Thus, both tasks essentially are similarity (or rather dissimilarity) judgments. In principle, such similarity judgments can be performed in a simple sequential manner – even for complex stimuli. Therefore, the difficulty-generating rules for individual items in both subscales may be similar, but nevertheless more complicated for matrices (UT6) than for classification (UT5). In contrast, when solving the items in series (UT4), one must take into account the whole series given in the stimulus in a holistic way in order to properly select the next object that completes the series. Thus, series (UT4) might differ from matrices (UT6) and classification (UT5) in the fundamental principle of two different solution strategies – i.e., a holistic approach vs. a simpler and more focused sequential approach. However, all differences in the loadings of the residuals on the first main component of the Rasch residual factor analysis turned out to be quite small ($\lambda_{max} = .42$ to $\lambda_{min} = -.24$) when compared to the findings regarding multidimensionality by Linacre (1998). Thus our findings from the Rasch residual factor analysis on one hand could simply reflect the slightly different instructions for students and on the other hand, could justify the theoretical underpinnings of the construction of the CFT 1-R. Nevertheless, for practical applications, the one-dimensional principle of summation of all item scores as a measure for intelligence is not negatively affected by these findings.

According to our research goals, we measured students with SEN in the fifth grade. In the present study, the CFT 1-R was subject of psychometric review in the framework of IRT with regard to the target population of students with LD. Thus, even older students who were potentially not properly covered by the respective norms in the test manual were included in the sample. As a result, the adequacy of the outcome T-values with regard to a meaningful comparability to normal classes, might be questioned. However, this limitation of the present research seems to be a minor point, as it was not the aim to achieve correct T-values in the sense of a norming study. With regard to its psychometric properties in the framework of Item Response Theory, the test comprising all items showed good targeting for the sample of students with LD – except for three items that turned out to be too easy (e.g. see figure 2).

By using the package `pairwise`, all statistical analyses were based on parameter estimates which parallel the principle of limited information estimation (e. g. Maydeu-Olivares, 2001; Bolt, 2005; Maydeu-Olivares & Joe, 2006; Maydeu-Olivares, 2006; Joe & Maydeu-Olivares, 2010). By doing so, it is possible to apply wide spread model fit-statistics with small datasets at the level of global model testing (i.e., Andersen likelihood tests and graphical model tests) as well as tests at the item level (i.e., Fischer-Scheiblechner tests and root-mean-square statistics). In this way, the fit-statistics could be used as evidence of the psychometric properties of the CFT 1-R in this study. Measurement invariance across students from inclusive settings and from special schools was fulfilled for our sample in this study. Such measurement equivalence is a requirement for the interpretation of latent group differences. Thus, the findings of the present research justify the use of the CFT 1-R to compare special student samples as was done in previous research (e. g. Hövel et al., 2015; Gebhardt et al., 2012; Sonntag, 2010; Voß et al., 2014) Certainly, we did not directly test the measurement invariance for students with and without LD in this research, but it can be assumed that such invariance is highly likely because the test was originally constructed to measure students without LD. However, the question of whether the measurement concept is valid for students with LD was tested in this study. As recommended by (Heydrich et al., 2013), some further analysis for students with SEN / LD in the regular age range and in comparison to regular classes might be necessary and thus should be subject of further research.

References

Aitkin, I., & Aitkin, M. (2011). *Statistical Modeling of the National Assessment of Educational Progress*. New York, NY: Springer New York.

Akaike, H. (1974, December). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatik Control*, *AC-19*(6), 716–723.

Andersen, E. B. (1973, March). A goodness of fit test for the rasch model. *Psychometrika*, *38*(1), 123–140. doi: 10.1007/BF02291180

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. *Journal of Applied Measurement*, *4*(3), 205–221.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi: 10.1007/BF02293801

Bock, R. D., Gibbons, R., & Muraki, E. (1988, January). Full-Information Item Factor Analysis. *Applied Psychological Measurement*, *12*(3), 261–280. doi: 10.1177/014662168801200305

Bollen, K. A. (1996). A Limited-Information Estimator for LISREL Models With or Without Heteroscedastic Errors. In G. A. Marcoulides (Ed.), *Advanced structural equation modeling* (pp. 227–242). Mahwah, NJ: Erlbaum.

Bolt, D. M. (2005). Limited- and Full-Information Estimation of Item Response Theory Models. In R. P. McDonald, A. Maydeu-Olivares, & J. J. McArdle (Eds.), *Contemporary psychometrics: a festschrift for Roderick P. McDonald.* Mahwah, N.J: Lawrence Erlbaum Associates.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: University Press.

Box, G. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics: proceedings of a workshop* (pp. 201–236). New York: Academic Press.

Bundschuh, K., & Winkler, C. (2014). *Einführung in die sonderpädagogische Diagnostik* (8th ed.). München: UTB GmbH.

Büttner, M. (1984). Diagnostik der intellektuellen Minderbegabung Untersuchung über die Zuverlässigkeit von Testbefunden. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *33*, 123–133.

Cattell, R. B. (1950). *Culture Fair (or Free) Intelligence Test (A Measure of „g" Skala 1. Handbook for the Individual or Groups*. Champaign, IL US: IPAT.

Cattell, R. B. (1963, February). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. doi: 10.1037/h0046743

Cattell, R. B., Cattell, R. B., & Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*(592).

Choppin, B. (1968). Item Bank using Sample-free Calibration. *Nature*, *219*(5156), 870–872. doi: 10.1038/219870a0

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, *62*(2), 269–276. doi: 10.2307/2335362

Cox, D. R., & Reid, N. (2004). A Note on Pseudolikelihood Constructed from Marginal Densities. *Biometrika*, *91*(3), 729–737.

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017, March). Investigating the Practical Consequences of Model Misfit in Unidimensional IRT Models. *Applied Psychological Measurement*, *41*(6), 439–455. doi: 10.1177/0146621617695522

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

Fischer, G. H. (1970, November). *A further note on estimation in Rasch's measurement model with two categories of answers* (Research Bulletin No. 3). Vienna: University of Vienna.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G. H. (1988). Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells. In K. D. Kubinger (Ed.), *Moderne Test Theorie.* Weinheim ; München: Psychologie Verlags Union.

Fischer, G. H., & Scheiblechner, H. (1970). Algorithmen und Programme fuer das probabilistische Testmodell von Rasch. *Psychologische Beitrage*(12), 23–51.

Fisher, W. P. (2010). IRT and Confusion about Rasch Measurement. *Rasch Measurement Transactions*, *24*(2), 1288–1288.

Forero, C. G., & Maydeu-Olivares, A. (2009, September). Estimation of IRT Graded Response Models: Limited Versus Full Information Methods. *Psychological Methods*, *14*(3), 275–299. doi: 10.1037/a0015825

Gebhardt, M. (2015). Gemeinsamer Unterricht von Schülerinnen und Schülern mit und ohne sonderpädagogischen Förderbedarf. Ein empirischer Überblick. In E. Kiel (Ed.), *Inklusion im Sekundarbereich* (pp. 39–52). Stuttgart: Kohlhammer.

Gebhardt, M., Schwab, S., Krammer, M., & Gasteiger, K. (2012, July). Achievement and Integration of Students with and without Special Educational Needs (SEN) in the Fifth Grade. *Journal of Special Education & Rehabilitation*, *13*(3/4), 7–19. doi: 10.2478/v10215-011-0022-6

Gebhardt, M., Zehner, F., & Hessels, M. G. P. (2014, April). Basic Arithmetical Skills of Students with Learning Disabilities in the Secondary Special Schools: An Exploratory Study covering Fifth to Ninth Grade. *Frontline Learning Research*, *2*(1), 50–63. doi: 10.14786/flr.v2i1.73

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch Model. In G. Fischer & I. Molenaar

(Eds.), *Rasch models: Foundations, recent developments, and applications.* New York: Springer.

Grünke, M. (2004). Lernbehinderung. In G. W. Lauth & M. Grünke (Eds.), *Interventionen bei Lernstörungen: Förderung, Training und Therapie in der Praxis* (pp. 65–77). Göttingen: Hogrefe.

Haeberlin, U., Bless, G., Moser, U., & Klaghofer, R. (1998). *Die Integration von Lernbehinderten. Versuche, Theorien, Forschungen, Enttäuschungen, Hoffnungen* (3rd ed.). Bern: Haupt Verlag AG.

Hasselhorn, M., & Autorengruppe Bildungsberichterstattung. (2014). *Bildung in Deutschland 2014: ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen.* Bielefeld: wbv, Bertelsmann.

Heimlich, U., Lotter, U., & März, M. (2005). *Diagnose und Förderung im Förderschwerpunkt Lernen. Eine Handreichung für die Praxis.* Donauwörth: Auer.

Heine, J.-H. (2017). *pairwise: Rasch Model Parameters by Pairwise Algorithm.* (R package version 0.4.2-1)

Heine, J.-H., Sälzer, C., Borchert, L., Siberns, H., & Mang, J. (2013). Technische Grundlagen des fünften internationalen Vergleichs. In M. Prenzel, C. Sälzer, E. Klieme, & O. Köller (Eds.), *PISA 2012 - Fortschritte und Herausforderungen in Deutschland.* Münster: Waxmann.

Heine, J.-H., & Tarnai, C. (2015). Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling*, *57*(1), 3–36.

Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013, September). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online / Journal für Bildungsforschung Online*, *5*(2), 217–240.

Holland, P. W. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hövel, D. C., Hennemann, T., Casale, G., & Hillenbrand, C. (2015). Das erweiterte LUBO-Schultraining in der Förderschule: Evaluation einer indizierten Präventionsmaßnahme in der Primarstufe der Förderschule. *Empirische Sonderpädagogik*, *7*(2), 117–134.

Irtel, H. (1987). On Specific Objectivity as a Concept in Measurement. In E. E. C. I. Roskam & R. Suck (Eds.), *Progress in Mathematical Psychology 1* (pp. 35–45). Amsterdam: North-Holland.

Joe, H., & Maydeu-Olivares, A. (2010, September). A General Family of Limited Information Goodness-of-Fit Statistics for Multinomial Data. *Psychometrika*, *75*(3), 393–419. doi: 10.1007/s11336-010-9165-5

Khalid, M. N., & Glas, C. A. W. (2014, April). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197. doi: 10.1016/j.measurement.2013.12.019

Kocaj, A., Kuhl, P., Kroth, A. J., Pant, H. A., & Stanat, P. (2014, June). Wo lernen Kinder mit sonderpädagogischem Förderbedarf besser? Ein Vergleich schulischer Kompetenzen zwischen Regel- und Förderschulen in der Primarstufe. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *66*(2), 165–191. doi: 10.1007/s11577-014-0253-x

Kubinger, K. D. (2005, December). Psychological Test Calibration Using the Rasch Model—Some Critical Suggestions on Traditional Approaches. *International Journal of Testing*, *5*(4), 377–394. doi: 10.1207/s15327574ijt0504_3

Kuhn, J.-T., Holling, H., & Freund, P. A. (2008, October). Begabungsdiagnostik mit dem Grundintelligenztest (CFT 20-R): Psychometrische Eigenschaften und Messäquivalenz. *Diagnostica*, *54*(4), 184–192. doi: 10.1026/0012-1924.54.4.184

Lee, H., & Geisinger, K. F. (2015, May). The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment. *Educational and Psychological Measurement*, 1–23. doi: 10.1177/0013164415585166

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement*, *2*(3), 266–283.

Linacre, J. M. (1999). Understanding Rasch Measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement*, *3*(4), 382–405.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, *5*(1), 95–110.

Lindsay, B. G. (1988). Composite Likelihood Methods. *Contempory Mathmatics*, *80*(80).

Lindsay, G. (2007, March). Educational psychology and the effectiveness of inclusive education/mainstreaming. *British Journal of Educational Psychology*, *77*(1), 1–24. doi: 10.1348/000709906X156881

Lloyd, J. W., Keller, C., & Hung, L.-y. (2007, August). International Understanding of Learning Disabilities. *Learning Disabilities Research and Practice*, *22*(3), 159–160. doi: 10.1111/j.1540-5826.2007.00240.x

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgment sampling. *Psychometrika*, *66*(2), 209–227.

Maydeu-Olivares, A. (2006, March). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*(1), 57–77. doi: 10.1007/s11336-005-0773-4

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111–127). New York: Routledge.

Maydeu-Olivares, A., & Joe, H. (2005, September). Limited- and full-information estimation and goodness-of-fit testing in 2(n) contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020. doi: 10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006, November). Limited Information Goodness-of-fit Testing in Multidimensional Contingency Tables. *Psychometrika*, *71*(4), 713–732. doi: 10.1007/

s11336-005-1295-9

Mellenbergh, G. J. (1982). Contingency Table Models for Assessing Item Bias. *Journal of Educational Statistics*, *7*(2), 105–118. doi: 10.2307/1164960

Millsap, R. E., Gunn, H., Everson, H., & Zautra, A. J. (2015). Using Item Response Theory to Evaluate Measurement Invariance in Health-Related Measures. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment.* New York & London: Routledge.

Millsap, R. E., & Maydeu-Olivares, A. (Eds.). (2009). *The SAGE handbook of quantitative methods in psychology*. Los Angeles ; London: SAGE.

Myklebust, J. O. (2002). Inclusion or exclusion? Transitions among special needs students in upper secondary education in Norway. *European Journal of Special Needs Education*, *17*, 251–264.

Osterlind, S. J. (1983). *Test Item Bias*. Newbury Park, California: SAGE Publications, Inc.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* (No. 1). Copenhagen: Danmarks pædagogiske Institut.

Rasch, G. (1964). *Objective Comparisons* [Lectures given at the UNESCO Seminar]. Oslo, Voksenåsen.

Rasch, G. (1977). On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements. *Danish Yearbook of Philosophy*, *14*, 58–93.

Robitzsch, A., Kiefer, T., & Wu, M. (2017, August). *TAM: Test Analysis Modules.*

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2., vollst. überarb. u. erw. Aufl. 2004 ed.). Bern: Huber.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased Item Detection Techniques. *Journal of Educational Statistics*, *5*(3), 213. doi: 10.2307/1164965

Scheiblechner, H. H. (2009). Rasch and pseudo-Rasch models: suitableness for practical test applications. *Psychology Science Quarterly*, *51*(2), 181–194.

Schuck, K. D. (2011). Die Bedeutung diagnostischer Daten im Prozess der Förderung durch Integrative Förderzentren in Hamburg. *Empirische Sonderpädagogik*, *3*(3), 188–206.

Schwab, S., & Helm, C. (2015). Überprüfung von Messinvarianz mittels CFA und DIF-Analysen. *Empirische Sonderpädagogik*, *7*(3), 175–193.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Sijtsma, K., & Hemker, B. T. (2000, December). A Taxonomy of IRT Models for Ordering Persons and Items Using Simple Sum Scores. *Journal of Educational and Behavioral Statistics*, *25*(4), 391–415. doi: 10.3102/10769986025004391

Sonntag, W. (2010). Fördert induktives Denken die Gedächtnisstrategie des Kategorisierens bei lernbehinderten Sonderschülern? *Empirische Sonderpädagogik*, *2*(1), 5–21.

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien; New York: Springer.

Tent, L., Witt, M., Bürger, W., & Zschoche-Lieberum, C. (1991). Ist die Schule für Lernbehinderte überholt. *Heilpädagogische Forschung*, *17*(1), 3–13.

van den Wollenberg, A. (1982, June). Two new test statistics for the rasch model. *Psychometrika*, *47*(2), 123–140. doi: 10.1007/BF02296270

Varin, C. (2008, February). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, *92*(1), 1–28. doi: 10.1007/s10182-008-0060-7

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5–42.

Voß, S., Blumenthal, Y., Sikora, S., Mahlau, K., Diehl, K., Hartke, B., & others. (2014). Rügener Inklusionsmodell (RIM)-Effekte eines Beschulungsansatzes nach dem Response to Intervention-Ansatz auf die Rechen-und Leseleistungen von Grundschulkindern. *Empirische Sonderpädagogik*, *6*(2), 114–132.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. doi: 10.1090/S0002-9947-1943-0012401-3

Weiß, R. H. (2008). *CFT 20-R Grundintelligenztest Skala 2 - Revision*. Göttingen: Hogrefe.

Weiß, R. H., & Osterland, J. (2013). *CFT 1-R Grundintelligenztest Skala 1 - Revision*. Göttingen: Hogrefe.

Wild, E., Lütje-Klose, B., Schwinger, M., Gorges, J., & Neumann, P. (2017). *BiLieF- Bielefeld longitudinal study of learning in inclusive and exclusive support arrangements.* IQB - Institute for Educational Quality Improvement.

Wright, B. D. (1977, July). Solving Measurement Problems with the Rasch Model. *Journal of Educational Measurement*, *14*(2), 97–116.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 3–24. doi: 10.1080/10705519609540026

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Zwick, R. (2012, June). A Review of ETS Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. *ETS Research Report Series*, *2012*(1), i–30. doi: 10.1002/j.2333-8504.2012.tb02290.x

Zwick, R., Donoghue, J. R., & Grima, A. (1993, September). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement*, *30*(3), 233–251. doi: 10.1111/j.1745-3984.1993.tb00425.x