Psychological Test and Assessment Modeling, Volume 58, 2016 (4), 597-616

The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes

Sarah Buerger[1], *Ulf Kroehne*[1] *& Frank Goldhammer*[1,2]

Abstract

This paper provides an overview and recommendations on how to conduct a mode effect study in large-scale assessments by addressing criteria of equivalence between paper-based and computer-based tests. These criteria are selected according to the intended use of test scores and test score interpretations. A mode effect study can be implemented using experimental designs. The major benefit of combining experimental design considerations with the IRT methodology of mode effects is the possibility to investigate partial measurement invariance. This allows test scores from different modes to be used interchangeably and means of latent variables or mean differences and correlations to be compared on the population level even if some items differ in difficulty between modes. For this purpose, a multiple-group IRT model approach for analyzing mode effects on the test and item levels is presented. Instances where partial measurement invariance suffices to combine item parameters into one metric are reviewed in this paper. Furthermore, relevant study design requirements and potential sources of mode effects are discussed. Finally, an extension of the modelling approach to explain mode effects by means of item properties such as response format is presented.

Keywords: mode effect, equivalence, computer-based assessment, partial measurement invariance, anchor items

[1] *Correspondence concerning this article should be addressed to:* Sarah Buerger, PhD, German Institute for International Educational Research (DIPF), Solmsstraße 73-75, 60486 Frankfurt am Main, Germany; email: buerger@dipf.de

[2] Centre for International Student Assessment (ZIB)

If different versions of a test are used and test scores are meant to be comparable between examinees, an investigation of the differences between those test versions is required (Parshall, Spray, Kalohn, & Davey, 2002; Wang & Kolen, 2001). Currently, a frequent case in which two versions of test instruments come into play in large-scale assessments involves a change in the administration mode from paper-based (PBA) to computer-based assessment (CBA; e.g., PIAAC, OECD, 2015; PISA, OECD, 2014a). Potential differences between scores on tests administered in different modes might be the result of mode effects, even if the computerization was conducted in such a way so as to make the test versions in each mode as comparable as possible. As a result, individuals with the same ability level completing the test in different modes may not obtain the same test score, meaning that scores cannot be used interchangeably (Raju, Laffitte, & Byrne, 2002; Van den Noortgate & De Boeck, 2005). Paper-based and computer-based assessments differ in measurement properties (e.g., test layout, navigation of the test and the handling of input devices, see Kroehne & Martens, 2011), which may in turn affect the comparability of their psychometric properties. Thus, differences between modes might not only be caused by a single measurement property but also by an amalgamation of the properties described as potential sources of mode effects in the next section. The probability of mode effects is assumed to increase "the more complicated it is to present or take the test on computer" (Pommerich, 2004, pp.3-4).

A change in administration mode can have an effect on psychometric properties like construct validity and the difficulty and discrimination of the test or single items (Mead & Drasgow, 1993; Puhan, Boughton, & Kim, 2007). The intended use of test scores and test score interpretations determine which psychometric properties have to be equivalent across modes. Those properties serve as criteria of equivalence, and the goal of a mode effect study is to try to falsify their equivalence.

Reviews of the literature reveal inconsistent findings regarding the equivalence of computer- and paper-based tests, meaning that some studies have falsified the equivalence hypothesis, whereas others have not (see Wang, Jiao, Young, Brooks, & Olson, 2008 for an extensive overview). One reason for these heterogeneous findings may be that there are a wide variety of methods used in mode effect studies (cf. Schroeders & Wilhelm, 2011; Wang, et al., 2008). These range from approaches based on *classical test theory* (CTT) to those based on *item response theory* (IRT). In addition, differences in sample sizes and thus the power of statistical tests allow for the detection of some effects of the administration mode, while others remain hidden. Although many studies have found no significant mode effects, the heterogeneity of results indicates that, in general, the existence of mode effects cannot be ruled out, and there are no reliable computerization rules to prevent mode effects. Thus, the appropriateness of comparisons has to be investigated in equivalence studies, whose findings must be documented as required by testing standards (e.g., AERA, APA, & NCME, 2014; American Psychological Association, Committee on Professional Standards [COPS] and Committee on Psychological Tests and Assessments [CPTA], 1986; Association of Test Publishers [ATP], 2000; International Test Commission [ITC], 2005).

In the context of current international and national large-scale assessments, computer-based testing is becoming more and more common. Thus, the comparability of comput-

er-based and paper-based tests is highly important, since it is a prerequisite for comparing more recent computerized scores with scores from previous cycles or with participants or countries in the same cycle completing the traditional paper-and-pencil form (OECD, 2014a). Evidence of comparability between different administration modes is necessary to ensure conditions that enable stable trend measures in large-scale assessments (Mazzeo & von Davier, 2008). In PISA (Programme for International Student Assessment) 2015, the administration mode shifted completely from paper-based to computer-based assessment (a move which is also planned for the NAEP [National Assessment of Educational Progress] in the U.S. in 2017), although participating countries had the option of implementing PISA as a paper-based survey in the Main Study. The comparability of computer-based and paper-based items also had to be addressed in PIAAC (Programme for the International Assessment of Adult Competencies) 2012 both because some participants lacking computer skills took the paper version and also to link scores back to previous paper-based adult literacy studies such as ALL (Adult Literacy and Lifeskills Survey, OECD, 2013) and IALS (International Adult Literacy Survey, OECD, 2013). For NEPS (National Educational Panel Study; Blossfeld, Roßbach, & von Maurice, 2011) in Germany, computer-based testing was introduced in 2012 as an alternative to the paper-based assessment, and mode effect studies have become crucial in linking different modes over time points. What all of these large scale assessments have in common is that, contrary to instruments for individual diagnostics, the focus is mainly on comparing means across populations such as schools or countries, and on correlations with other performance-related variables. This intended use of test scores in large-scale assessments leads to specific equivalence criteria such as construct equivalence that should be investigated in order to ensure the required level of measurement invariance between tests administered in different modes.

In this paper, we propose a comprehensive multiple-group IRT (Item Response Theory) model approach to assess different levels of measurement invariance for categorical dependent variables. This general approach of testing hypotheses for equivalence with regard to relevant criteria also remedies shortcomings of previous studies, which have used diverse and sometimes inappropriate methods. This model is suitable for analyzing mode effects in experimental designs, where randomization is conducted, and aims to investigate partial measurement invariance, meaning that not all items have to be invariant between the modes. Its purpose is to ensure that the prerequisites for valid comparisons of results from different administration modes hold even if some or all items are unequal between modes, because mode effects can be represented by mode-specific item parameters that account for differences (e.g. in difficulty).

Sources of mode effects

Drawing on empirical evidence from previous studies, the following section presents measurement properties that might be potential sources of mode effects. With regard to the modelling approach proposed in this paper, these properties can be used to explain mode effects.

Input devices

Different input devices like a pen on paper, a mouse on the computer or the touchscreen on a tablet might interfere and interact with the experience of the test-takers in different ways (see Bennett, 2003; Parshall, Harmes, Davey, & Pashley, 2010; Schroeders & Wilhelm, 2010).

Test and item layout

The page or screen size and orientation – typically portrait on paper and landscape on the computer – differ between administration modes, which has an effect on the number of text pages as well as the size and placement of the text (e.g., column and line breaks). The presentation of multiple items on a page, as is commonly done on paper, versus one item at a time on the screen of a computer may also cause mode effects (Schroeders & Wilhelm, 2010).

Scrolling

If the amount of information on a page is larger than the screen, scrolling or paging with a mouse or touchpad to read a text and associated items is another potential source of mode effects. Scrolling has repeatedly been shown to be more difficult than paging (Bridgeman, Lennon, & Jackenthal, 2001; Higgins, Russell, & Hoffmann, 2005; Kim & Huynh, 2008; Kingston, 2009; Mazzeo & Harvey, 1988; Pearson Educational Measurement, 2005; Poggio, Glasnapp, Yang, & Poggio, 2005; Pommerich, 2004; Schwarz, Rich, & Podrabsky, 2003; Wang et al., 2008). However, findings from the large-scale assessment PIAAC suggest that the extent of scrolling had no significant impact on the difficulty of the items in the computer-based version (Yamamoto, 2012).

Item review

Item review, that is, whether the test-taker can go back to an item and change their answer, is often prohibited on the computer due to the prevention of backward navigation. On paper, navigation between items is typically not restricted (Pommerich & Burden, 2000; Vispoel, 2000). This aspect of test-taking flexibility may comprise a difference between modes (Bodmann & Robinson, 2004). However, Vispoel (2000) found in a low-stakes testing context that preventing item review did not affect average scores or psychometric properties, as only a very small percentage of test-takers used the opportunity to change their answers. In such cases, the individual benefit resulting from using item review increased with test-takers' ability level. Vispoel (2000) also showed that test-takers expressed a strong desire for item review opportunities, especially those exhibiting test anxiety.

Item response format

Computerized response formats often look only slightly different from the paper version, but show greater differences for more complex response formats such as assignment tasks or constructed responses (e.g., Heerwegh & Loosveldt, 2002; Parshall et al., 2010; Parshall et al., 2002; Sireci & Zenisky, 2006). The complexity of an item, that is, "the number and type of examinee interactions within a given task or item" (Parshall et al., 2002, p.9) plays an essential role in mode effects. Studies have shown that computerized multiple-choice items are less prone to mode effects because they are of lower item complexity (Bennett et al., 2008; Bodmann & Robinson, 2004; Parshall et al., 2002). Items requiring constructed responses have been shown to be more difficult on the computer than on paper (Bennett et al., 2008). The response format of drop-down boxes, which are often used for assignment tasks, also turns out to be more difficult when implemented on the computer (Heerwegh & Loosveldt, 2002). Mode effect studies for NEPS (Buerger, Kroehne, & Goldhammer, 2015) and PIAAC (Yamamoto, 2012) also focused on item response formats as possible sources of differences between modes. In NEPS, items with drop-down boxes on the computer showed higher difficulty (Buerger et al., 2015). The computer-based response formats used in PIAAC, such as highlighting, clicking, or scrolling, had no effect on item difficulty (Yamamoto, 2012).

Interaction of mode with test-takers' characteristics

Mode effects might be influenced by (computer-related) test-taker characteristics, which might interact with properties of the test administration and affect test-takers' performance (Kroehne & Martens, 2011). For instance, a person's general familiarity with computers has been shown to interact with the mode of administration: Students with a high degree of familiarity had an advantage in a computer-based test over students with a low degree of familiarity (Bennett et al., 2008; Clariana & Wallace, 2002; Wang et al., 2008). However, other studies have found no interaction with test-takers' computer literacy and computer use (Bennett, 2002; Higgins et al., 2005). In large-scale assessments, where populations and sub-populations are the primary focus of test score interpretations, it is crucial to consider that mode effects may vary across countries if countries differ, for instance, in the overall accessibility and usage of computers.

Design requirements and identification of groups

When comparing different modes, the assignment of persons to mode represents the main relevant aspect of the study design. The investigation of criteria (e.g. construct equivalence) for falsifying the equivalence hypothesis as well as related inferences are primarily dependent upon the design of the mode effect study (Wang et al., 2008). Therefore, when planning a mode effect study, the decision regarding mode assignment (see Figure 1) needs thorough consideration: The mode can vary between persons (between-

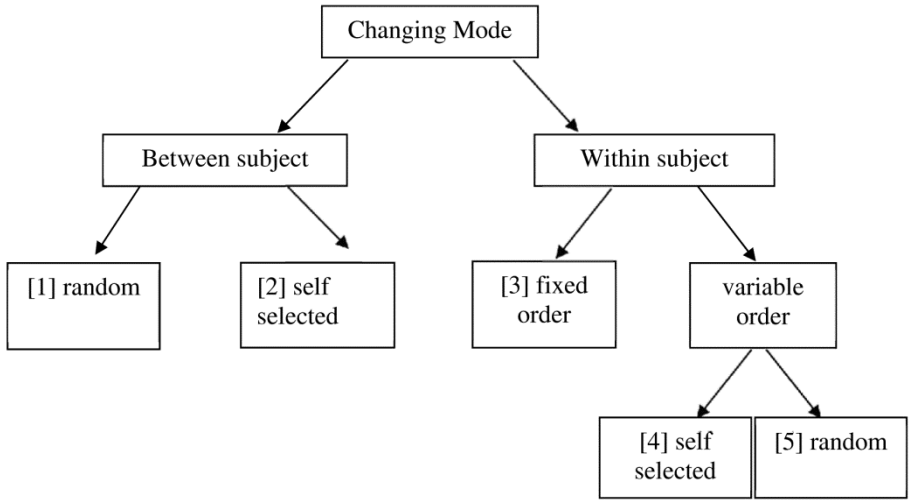

Figure 1:
Design for Mode Effect Study. Mode changing between subjects with randomization to the mode or by self-selection; or within subjects with fixed or variable order of the test parts.

subject design), meaning that each test-taker takes the test in only one mode, or it can vary within persons (within-subject design), meaning that the test is split into two parts and both parts are administered to every person in different modes. In the case of a between-subject design, there is the further distinction of whether the persons are randomly assigned to a mode [1] or get to choose the mode themselves [2] (see below). In a within-subject design, the order of the modes has to be considered. A fixed order refers to the situation when all persons take both test parts in the same order [3] (e.g., Pomplun, Frey, & Becker, 2002). With a variable test order, the question arises whether persons are allowed to choose the order in which they will take the test parts themselves [4] (e.g., Puhan et al., 2007), or whether the order is balanced and persons are randomly assigned to start with a certain mode [5] (e.g., Kim & Huynh, 2008). As the position of the test parts (modes) could have an effect, a balanced order of the test parts [5] is preferable.

The within-subject design is advantageous in several respects. It has higher statistical power because the error variance associated with individual differences is reduced and results do not depend on the assignment of persons to mode groups (Schroeders & Wilhelm, 2011). To overcome the disadvantages of a between-group design, one has to enlarge samples and ensure comparability between groups. Thus, the random assignment of test-takers to mode groups (cf. Holland & Dorans, 2006) is crucial for the interpretation of test score differences between modes. Only a design with randomly equivalent groups [1, 3, 5], ensures that the differences between modes do not occur simply due to non-random ability differences (e.g., Osterlind & Everson, 2009). Differences are then a direct result of the change in administration mode.

Groups are non-randomly equivalent if assignment to a certain mode depends, for instance, on the test-taker's decision [2, 4], meaning that essential person-related variables are not automatically balanced (Kingston, 2009). Then, comparability needs to be ensured in a different way. One possibility to make the ability distribution as comparable as possible is to use person-level covariates. In the case of self-selection of the mode, additional data on the decision-making process is necessary to build equivalent groups, for instance, with the help of matching techniques such as propensity score matching (e.g., Hox, de Leeuw, & Zijlmans, 2015). If no information about the decision-making process is available, items that can be assumed to be equal between modes are required in order to create a common ability metric.

When planning the design, how the criterion of construct equivalence is to be investigated also plays a role. The extension of the randomly equivalent group design to an order-balanced within-subject design [5] allows for the estimation of (latent) correlations between the modes and is another advantage of this design. In this case, the question of construct equivalence can be addressed by investigating whether the cross-mode correlation is not significantly different from 1. This is impossible in between-group designs, which are frequently used in large-scale assessments such as PISA, PIAAC and NEPS (Rutkowski, von Davier, & Rutkowski, 2013). Here, external criteria as suggested by a nomological network of the construct need to be measured in both groups to analyze the relationship between the test in both modes and these external criteria. For instance, for a reading comprehension test, tests of basic reading skills, such as a lexical decision task or a sentence verification task, can serve as sources of convergent evidence, while a test from a different domain (e.g. science) can be used as a source of divergent evidence. In any case, if the results from the paper-based and the computer-based assessments are equally correlated with the external criteria, the investigated hypothesis of construct equivalence is not falsified. A difference in correlations suggests that another (mode-specific) construct is also being assessed. The administration mode of criterion variables needs to be decided as well, considering that according to the *multi-trait multi-method* perspective (e.g., Eid, 2006), the correlations of tests in the same mode might be higher.

Note that an experimental mode effect study, where randomly equivalent groups complete paper and computer versions of an instrument, differs from studies investigating differential item functioning (DIF). In DIF studies, groups differ in person-level variables such as gender or mother tongue, meaning that groups cannot be randomly equivalent, whereas the instruments they complete are identical. In DIF studies, the usual way to create a common metric and thus ensure the comparability of test scores across groups is to use so-called anchor items, which are assumed to be invariant between groups and thus show no differential functioning (e.g., Reise, Widaman, & Pugh, 1993). Regarding the assessment of mode effects, items not affected by the mode change and therefore forming an anchor can be used to identify a common metric with respect to the targeted construct, allowing differences between non-anchor items to be identified as mode effects. Note that this is possible even for non-equivalent groups and not necessary for random equivalent groups, where the common metric is ensured with equal ability distributions that are guaranteed if randomization was successful.

Multiple-group IRT model for analyzing mode effects on equivalence criteria

As shown above, a broad range of methods have been used to test the existence of mode effects in previous studies. The choice of the design and the related question of whether all or at least some items should be invariant between modes is not addressed explicitly in most approaches (as for instance in multiple group confirmatory factor analysis, suggested by Schroeders, 2009 and Schroeders & Wilhelm, 2011, 2010). If the assumption of measurement invariance does not hold, it remains unclear whether this results from only some items showing a mode effect or from a general shift in difficulty. Manifest approaches, such as comparisons of means (e.g., Alexander, Bartlett, Truell, & Ouwenga, 2001; Bodmann & Robinson, 2004; Pommerich, 2004; Pomplun & Custer, 2005; Pomplun et al., 2002; Puhan et al., 2007; Schwarz et al., 2003; Wang, 2004) and cross-mode correlations of item parameters (Bennett et al., 2008; Pommerich, 2007), which are frequently described as ways of identifying differences between modes, are not sufficient to falsify hypotheses on equivalence without evidence of equal latent constructs, a fact which is often ignored.

The multiple-group IRT modelling approach proposed in this paper provides insights into item-specific mode effects by introducing a mode effect parameter. This parameter can apply either to all items or vary across items. When it varies across items, it is possible to find a selection of items that are invariant between modes. Such items can be used as anchor items in non-randomly equivalent groups. Furthermore, our approach helps to identify item properties that may increase the probability of mode effects.

In the next section, criteria for falsifying equivalence are presented, followed by the latent variable modeling approach that is illustrated for a within-subject and a between-subject design. This approach investigates mode effects regarding item parameters on both the test and item levels. Thereby, the criteria of equivalence related to item parameters, that is, (partial) measurement invariance, can be tested systematically.

Criteria of equivalence

A mode effect study attempts to provide empirical evidence justifying cross-mode comparisons. Criteria reflecting cross-mode equivalence should be specified on the basis of the intended use of test scores and related inferences. Thus, a mode effect study can be understood as part of the validation of the test score interpretation (cf. AERA, APA, & NCME, 2014). Intended comparisons differ for large-scale assessments, individual assessments and high-stakes tests. In large-scale assessments, typical uses of scores primarily include the comparison of means and correlations at the level of populations or subpopulations (Oliveri & von Davier, 2011; Rutkowski et al., 2013), whereas assessments on the individual level, including high-stakes testing, compare individual scores and often have relevant consequences for the test-taker.

The first criterion of equivalence is construct equivalence, that is, whether the construct measured by the test is the same in both modes. Despite its importance, this criterion has

received little consideration up to this point. The second and rather minor criterion of equivalence in some contexts is equal test reliability in both modes, which ensures that test score comparisons are not affected by differences in measurement accuracy. The equality of item parameters can be considered as a third criterion, and its investigation depends on the measurement model, which is described in the next section.

Measurement model and construct equivalence. A first step of a mode effect analysis is to determine an appropriate measurement model that fits simultaneously for both modes. In a within-subject design, a multi-dimensional IRT model has to be tested, while in the case of a between-subject design, a multiple-group IRT model needs to be tested and compared with respect to information criterion and used to investigate item fit. A measurement model that fits data from both modes simultaneously implies that the mode effects can be described as differences in the set of item parameters included in this measurement model (e.g., item difficulty and discrimination). Thus, the determination of a measurement model for both modes is prerequisite to absorb mode effects on IRT item parameters using latent variable modelling.

The question of construct equivalence – that is, whether the test captures the same latent variable in both modes is the first equivalence criterion. Construct equivalence is related to the step of determining a common measurement model: A latent variable model that includes responses from both modes enables testing construct equivalence by estimating latent correlations. When switching to another mode, construct-irrelevant individual differences or even another construct may be tapped, meaning that construct equivalence has to be tested when comparing results assessed in different modes (AERA, APA, & NCME, 2014; Huff & Sireci, 2001; ITC, 2005; Parshall et al., 2002; Penfield & Camilli, 2007; Puhan et al., 2007; Russell, Goldberg, & O'Connor, 2003). The technical implementation of hypotheses regarding construct equivalence depends on the specific study design (see section of design considerations). In large-scale assessments, where interest centers on comparisons of means and correlations with regard to a certain construct at the level of (sub-)populations, construct equivalence of the test versions is critically important.

Since most data in educational measurement are categorical, we restrict this step of determining an appropriate measurement model to IRT-based approaches, although categorical data can also be modelled within the framework of structural equation modeling (e.g., for mode effect analysis with CFA, Schroeders & Wilhelm, 2011). The IRT model to be chosen depends on whether item scoring is dichotomous or polytomous. The selected model should be as liberal in terms of item parameters so as to describe data from both modes appropriately. If the data from both modes do not fit to the same IRT model, a combined and more complex and thus liberal IRT model (e.g., integrating those items conforming to the Rasch model (Rasch, 1960) and those with deviating discriminations) can be used, as was, for instance, done in PISA (OECD, 2014b). In order to find an appropriate measurement model, another and frequently described possibility is excluding items that lead to worse model adjustment due to item misfit. Although this is one way of improving model fit, it limits the results of a mode effect study because the mode difference is only investigated for a subset of items (it could be the case that items with a large mode effect were excluded in pre-analysis).

To test for construct equivalence, latent or manifest cross-mode correlations can be used in a within-subject design (e.g., Mead & Drasgow, 1993). Manifest correlations are attenuated and can underestimate the true linear relationship if the constructs are measured with error. Therefore, latent correlations are expected to be not significantly different from 1 (r_1, Figure 2) if the order of the test parts is balanced, whereas manifest correlations are expected to be as high as predicted by the test's reliability in order to support the hypothesis of construct equivalence. An additional approach besides cross-mode correlations is to investigate the relation to external criteria. This is the method of choice in the case of a between-subject design, where cross-mode correlations between the latent variables are not possible due to different examinees responding to items in only one mode. Figures 2 and 3 show the correlations with an external criterion in a within-subject and between-subject design, respectively. The criterion of construct equivalence requires equal correlations (latent or manifest) of PBA with an external criterion (r_2, either latent or manifest) and CBA with the same external criterion (r_3, either latent or manifest). When estimating latent correlations, item parameters are freely estimated between modes because measurement invariance is not a precondition for construct equivalence. Modeling data with one IRT model and testing construct equivalence is necessary condition for subsequent steps such as concurrent calibration, which aligns item parameters along a common metric and thus allows scores to be used interchangeably.

Reliability. That both tests have same level of reliability is another criterion that needs to be ensured if results from different modes are to be compared (AERA, APA, & NCME, 2014; Holland & Dorans, 2006; ICT, 2005; Kolen & Brennan, 2004). This is more important in individual assessments, and can be considered optional in large-scale studies, as larger sample sizes may compensate for a decrease in reliability (see Adams, 2005, for reliability as a measurement design effect). In addition to individual assessments, equal reliabilities also become important when equating or linking between modes (Dorans, Moses, & Eignor, 2010) and when modes are changed in longitudinal studies (Buerger et al., 2015). In the context of IRT modeling, reliability is a function of item difficulty and item discrimination parameters, which represent another criterion for evaluating mode differences.

Item parameters. A third criterion according to which mode differences can be analyzed is item parameters, i.e. difficulty and discrimination (depending on the measurement model). Mode-related differences reflected in item parameters can be classified as a) homogeneous effects on the test level that affect all items in a similar manner or b) heterogeneous effects on the item level. Mode effects that vary across items might be systematic and depend on specific item properties, for example (Green, Bock, Humphreys, Linn, & Reckase, 1984). Analyzing mode effects on item parameters on the test and item level is described in the next section.

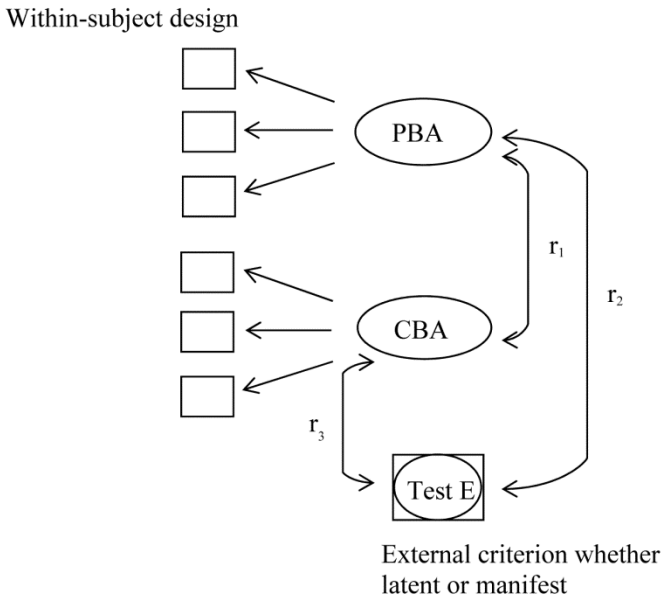

Figure 2:
Testing for construct equivalence in a within-subject design; r_2 is the correlation of PBA and the external criterion (Test E), r_3 is the correlation of CBA and the same external criterion and r_1 is the cross-mode-correlation

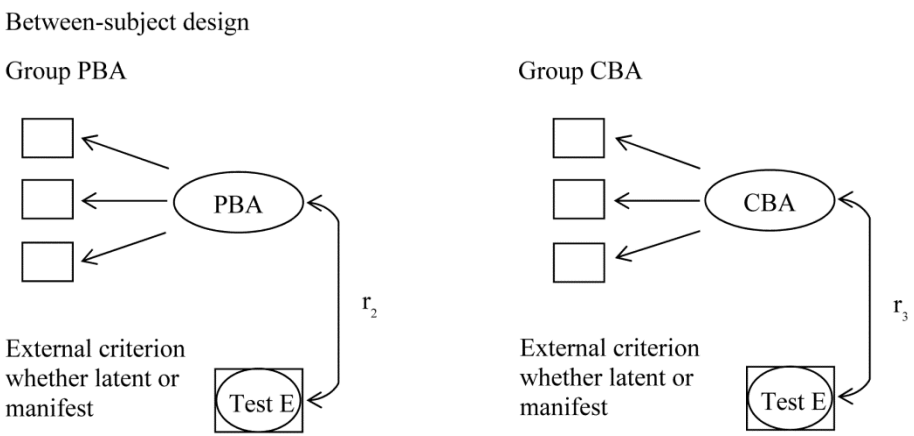

Figure 3:
Testing for construct equivalence in a between-subject design; r_2 is the correlation of PBA and the external criterion (Test E) and r_3 is the correlation of CBA and the same external criterion

The modeling approach

After investigating the common measurement model as a precondition for comparing item parameters between modes, measurement invariance has to be tested, that is, whether the relationship between the observed responses and the underlying latent variable is identical in both modes.

(Partial) Measurement invariance. Measurement invariance (i.e., equality of item parameters between groups) is typically addressed in studies in which groups are not randomly equivalent because an experimental assignment of persons to groups (defined by person-level variables) is not possible. For instance, when a PISA test is translated into multiple languages and differential item functioning between those test versions is investigated, persons cannot be arbitrarily assigned to language versions regardless of the person-level variable mother tongue. In mode effect studies, however, assignment to mode groups proceed randomly using the experimental design described above. This offers the advantage that measurement invariance (i.e., item parameters are invariant across both modes) is not needed to align items from both modes to one common metric, because performance differences cannot be the result of group differences but rather can be clearly attributed to differences in the administration mode.

The multiple-group IRT model approach allows for the investigation of partial measurement invariance. Specifically, partial measurement invariance is needed when the assumption of random equivalent groups does not hold (e.g., when persons are allowed to choose the mode themselves). In such cases, items that turn out to be invariant between modes can serve as anchor items. Kolen and Brennan (2004) describe requirements for common items to serve as an anchor. They have to represent the measured construct and must be representative of the test's specifications. That is, the more heterogeneous the test content is, the more invariant items are needed to capture variety in content. In general, the random equating error decreases with an increasing number of invariant items. In order to avoid a distortion in equating due to position effects, the position of the anchor items should be the same in both test versions.

Model. The common measurement model that fits the CBA and PBA data determines the item parameters and thereby where differences between administration modes can be observed. If the Rasch model applies, discrimination is assumed to be equal for all items within groups, so that only item difficulty is investigated in equivalence analysis. In a 2 PL model, item discrimination also needs to be examined for equivalence, while in a 3 PL model the guessing parameter needs additional consideration (Birnbaum, 1968). In this paper, we illustrate the investigation of partial measurement invariance for tests that fit the Rasch model, meaning that a potential mode effect only affects item difficulties. However, the described modelling approach can be easily generalized to more complex IRT models.

In the Rasch model, the probability P of a correct response of person *i* on item *j* in mode (group) *m* is defined as:

$$logit\left[\mathrm{P}\left(Y_{ijm}=1\right)\right]=\theta_i-\beta_{jm} \tag{1}$$

In the case of a between-subject design, the mode *m* is the mode group to which test-takers are randomly assigned. In a within-subject design, the mode *m* stands for the order-balanced test parts in PBA and CBA.

Under PBA (mode = 0), item *j* has the difficulty of $\beta_{j,PBA}$:

$$logit\left[\mathrm{P}(Y_{ijm}=1|G=0)\right]=\theta_i-\beta_{j,PBA} \quad (2)$$

If there is a mode effect at the item level, for item *j* the difficulty is changed to the amount of a new introduced mode parameter ME_j under CBA (mode = 1):

$$logit\left[\mathrm{P}(Y_{ijm}=1|G=1\right]=\theta_i-(\beta_{j,PBA}+ME_j) \quad (3)$$

If there is a systematic mode effect across all CBA items (i.e., mode effect at the test level), all items show the same shift in difficulty, that is, ME_j is equal for all items *j*:

$$logit\left[\mathrm{P}(Y_{ijm}=1|G=1)\right]=\theta_i-(\beta_{j,PBA}+ME) \quad (4)$$

With this multiple-group model approach introduced, invariant items are identified by testing differences in item parameters between modes. To do this, software is required that allows for the modeling of multiple groups and the estimation of standard errors of the difference between item parameters from different groups. This can be done, for example, in Mplus (Muthén & Muthén, 1998-2015) using the MLR estimator and mixture type of analysis, and by introducing a new parameter (ME_j) representing the difference in the item parameters for all item pairs from PBA and CBA. To test mode effects, constraints are imposed on this new parameter ME_j. Basically, if the constraints worsen the model fit, meaning that equality assumptions do not hold, a mode effect is indicated.

Testing constraints. To test for measurement invariance using our modelling approach, three hypotheses on the test and item levels are inspected. *Hypothesis A* is that no mode effect exists at the test level. In *Hypothesis B,* the mode effect is expected to be equal for all items, whereas in *Hypothesis C* it varies across items. *Hypothesis A* is tested by the constraint $\sum ME_j = 0$, which means that across all items *j= 1...J* there is no mode effect.

To test this hypothesis, the Wald test statistic (Wald, 1943) can be used. A significant Wald test statistic for *Hypothesis A* means that there is a mode effect on item difficulties. Here, it is important to note that the sum of all item differences might also be zero and the Wald test insignificant if there are some mode effects in opposite directions that cancel each other out. Finding a significant mode effect in this step means there has been a shift in average item difficulty from one mode to another. Note that for the identification of the mode specific parameter ME_j, we either need the assumption of random equivalent groups, which allows us to fix the mean of the latent variable in both mode groups to zero, or anchor items must be defined to allow for the estimation of latent mean differences. In this illustration of the modelling approach, we assume random equivalent groups. Furthermore, as a consequence of the assumption of the Rasch model

as the IRT model that fits the data in both modes, the variances of the latent variable can be constrained to be equal.

The next step on the test level is to analyze whether this mode effect is homogeneous over all items. Therefore, in *Hypothesis B,* all differences between item parameters are constrained to be equal across items, $ME = ME_j$. If *Hypothesis B* holds, all items on the computer get a CBA-specific item parameter by adding the *ME* component to the PBA-specific item parameter $\beta_{j,CBA} = \beta_{j,PBA} + ME$. In the case where *Hypothesis B* is rejected, the mode effect cannot be simplified to a general shift in item difficulties, and the model with item-specific mode effects fits the data better. Accordingly, *Hypothesis C* tests mode effects on the item level for each item *j*: $ME_j = 0$. For each item, the mode effect is represented by the difference in PBA and CBA item parameters that is tested to be different from zero. Given the estimated standard error for the difference between the item difficulties in both modes, a t-test can be conducted for each item *j*, testing whether or not ME_j is different from zero. This test can be used to identify single items showing a mode effect, thus also identifying anchor items. If there is a mode effect for only some items in the test, those items get a specific item parameter for CBA, $\beta_{j,CBA} = \beta_{j,PBA} + ME_j$. Partial measurement invariance can be assumed for those items for which *Hypothesis C* holds and that are thus not affected by the mode change.

Furthermore, the multiple-group IRT model approach allows for an investigation of the mode effect at the item level by relating it to item properties (as previously described in the section on sources of mode effects). To do this, a new parameter is created, representing whether a given item exhibits such a property. For instance, all computerized items with the need for scrolling could be assumed to be more difficult under CBA than PBA. Thus, the mode effect at the item level can be explained by a set of item properties that are assumed to induce a mode effect: $ME_j = \gamma_1 \bullet \mathrm{x}_{j1} + \gamma_2 \bullet \mathrm{x}_{j2} + \ldots + \gamma_k \bullet \mathrm{x}_{jk}$. For each item property *k*, x_{jk} indicates whether this property is given for item *j* with $\mathrm{x}_{jk} = 1$ and $\mathrm{x}_{jk} = 0$ otherwise. If this decomposition can be identified, the weight γ_k indicates how strong the property contributes to the mode effect. For instance, if the property "scrolling" is given for an item, and it shows a significant effect on the logit scale, this effect can be translated into a change in the probability of completing the item successfully.

Discussion

Since a change from paper-based to computer-based assessment has taken place in many large-scale assessments, including, for instance, PISA, PIAAC, and NEPS, mode effect studies are highly relevant. In mode effect studies between 2000 and today, both the way tests are presented on the computer and the methods used to check equivalence vary considerably. Some researchers made decisions about equivalence solely by comparing mean scores for mode groups with no regard to items, whereas others did an extensive

investigation of item parameters, the order of test versions and characteristics of test-takers or subgroups. In addition, differences in sample size and thus the power of statistical tests have let some effects of the administration mode be detected, while others probably could not be discovered. Thus, empirical evidence about the equivalence of modes cannot be compared easily. For this reason and to propose a standard tool, respectively, we presented a multiple-group IRT model approach for investigating differences between paper-and-pencil and computer tests in item parameters. Moreover, we discussed one major equivalence criteria, that is, the issue of construct equivalence.

Defining the measurement model and checking for construct equivalence has seldom been examined in previous studies, although it is critical to the interpretation of data from different administration modes. The step of defining a measurement model determines the parameters in which mode differences may be observed. Regarding differences in item parameters between modes, we revert to the terminology of measurement invariance testing (where groups usually are non-randomly equivalent). Transferring this approach to a mode effect analysis assuming randomly equivalent groups generated via experimental designs offers an opportunity to investigate partial measurement invariance. If groups are non-randomly equivalent for some reason, invariant (anchor) items are needed to put scores on a common metric and use scores interchangeably.

Examining each item for mode effects allows the model to be expanded easily to investigate whether the mode effect depends on item properties (e.g., response format). As part of this process, item properties that increase the difficulty of an item on the computer can be identified. Knowing about those properties provides an opportunity to construct test items that will be less prone to mode effects. To better understand and prevent differences between modes, further research on how item properties are related to mode effects is required (Buerger et al., 2015). Characteristics of subgroups or countries disadvantaged by a particular administration mode should also be considered in future mode effect analyses of large-scale assessments.

If significant mode effects are found for specific items, the question arises of how big the effect size is and whether this effect is of practical relevance. The increase (or decrease) of probability of success for test-takers with the same ability but taking the test in different modes may help to illustrate the relevance of an effect. However, the literature on DIF could be consulted to evaluate effect sizes more clearly, and techniques proposed there could be adapted to the analysis of mode effects (see Magis, Béland, Tuerlinckx, & de Boeck, 2010, for a detailed overview of appropriate methods as well as Zieky, 1993 for a classification scheme).

Some research has shown that computer-based tests have a faster completion time (e.g., Alexander et al., 2001; Bodmann & Robinson, 2004), which means that more items can be presented to test-takers on the computer than on paper. This should be considered if reliability differs between modes because the additional items might compensate for reliability differences by increasing the reliability of the computer test. Differences in time intensity can be regarded as the result of a mode effect, or as a variable mediating an effect on other psychometric properties such as difficulty. Here, further analysis in-

vestigating test-taking time for computer-based assessments compared to paper-based tests is required.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31,* 162-172.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington: AERA, APA, NCME.

Alexander, M. W., Bartlett, J. E., Truell, A. D., & Ouwenga, K. (2001). Testing in a Computer Technology Course: An Investigation of Equivalency in Performance Between Online and Paper and Pencil Methods. *Journal of Career and Technical Education*, *18* (1). Retrieved from http://scholar.lib.vt.edu/ejournals/JCTE/v18n1/alexander.html

American Psychological Association, Committee on Professional Standards (COPS), & Committee on Psychological Tests and Assessments (CPTA). (1986). *Guidelines for computer-based tests and interpretations*. Washington: American Psychological Association, Inc.

Association of Test Publishers (ATP). (2000). *Guidelines for computer-based testing*. Washington DC: Association of Test Publishers.

Bennett, R.E. (2002). *Using electronic assessment to measure student performance*. (Issue Brief). Washington, DC: NGA Center for Best Practices. Retrieved from http://www.nasbe.org/Standard/10_Summer2002/bennett.pdf

Bennett, R. E. (2003). Online Assessment and the Comparability of Score Meaning. *Research Memorandum 3* (5), 1–19. Retrieved from http://www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *The Journal of Technology, Learning, and Assessment, 6* (9).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examiner's ability. In F. M. Lord & M. R. Novick (Ed.) *Statistical theories of mental test scores,*17–20.

Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J. (Eds.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). [Special Issue] *Zeitschrift für Erziehungswissenschaft, 14*.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31* (1), 51–60.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (ETS-RR-01-23). Princeton, NJ: Educational Testing Service.

Buerger, S., Kroehne, U., & Goldhammer, F. (2015, July). *Investigating Mode Effects in Reading Assessments*. Paper presented at the 6th Conference of the European Survey Research Association, Reykjavik.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with the test mode effect. *British Journal of Educational Technology, 33* (5), 593–602.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and Practices of Test Score Equating* (ETS Research Rep. No. RR-10-29). Princeton, NJ: ETS.

Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 223–230). Washington, DC: American Psychological Association.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical Guidelines for Assessing Computerized Adaptive Tests. *Journal of Educational Measurement, 21* (4), 347–360. Retrieved from http://www.jstor.org/stable/1434586

Heerwegh, D., & Loosveldt, G. (2002). An Evaluation of the Effect of Response Formats on Data Quality in Web Surveys. *Social Science Computer Review, 20* (4), 471–484.

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the Effect of Computer-Based Passage Presentation of Reading Test Performance. *The Journal of Technology, Learning, and Assessment, 3* (4).

Holland, P. W., & Dorans, N. J. (2006). Linking and Equating. In R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 187–220). Westport, CT: Praeger.

Hox, J.J., de Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology, 6*:87. doi: 10.3389/fpsyg.2015.00087

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practices, 20* (3), 16–25.

International Test Commission (ITC). (2005). *International Guidelines on Computer-Based and Internet Delivered Testing.* Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Kim, D.-H., & Huynh, H. (2008). Computer-Based and Paper-and-Pencil Administration Mode Effects on a Statewide End-of-Course English Test. *Educational and Psychological Measurement, 68* (4), 554–570.

Kingston, N. M. (2009). Comparability of Computer- and Paper-Administered Multiple-Choice Tests for K–12 Populations: A Synthesis. *Applied Measurement in Education, 22,* 22–37.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer-Verlag.

Kroehne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft, 14* (2), 169–186.

Magis, D., Béland, S., Tuerlinckx, F, & de Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42* (3), 847-862. doi:10.3758/BRM.42.3.847

Mazzeo, J., & Harvey, A. L. (1988). *The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests. A Review of the Literature* (College Board Report 88-8). New York: College Entrance Examination Board.

Mazzeo, J., & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Education Working Papers EDU/PISA/GB (2008)*, 28, 23–24.

Mead, A. D., & Drasgow, F. (1993). Equivalence of Computerized and Paper-and-Pencil Cognitive Ability Tests: A Meta-Analysis. *Psychological Bulletin, 114* (3), 449-458.

Muthén, L.K., & Muthén, B.O. (1998-2015). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

OECD (2013). The Survey of Adult Skills: Reader's Companion, OECD Publishing. http://dx.doi.org/10.1787/9789264204027-en

OECD (2014a). PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014), PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

OECD (2014b). Pisa 2015 Field Trial Goals, Assessment Design, and Analysis Plan for Cognitive Assessment, PISA, OECD Publishing.

OECD. (2015). Adults, Computers and Problem Solving: What's the Problem? OECD, Publishing. http://dx.doi.org/10.1787/9789264236844-en

Oliveri, M. E., & von Davier, M. (2011). Investigating of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53* (3), 315-333.

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning*. Thousand Oaks: SAGE Publications, Inc.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. J. (2010). Innovative Items for Computerized Testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of Adaptive Testing, Statistics for Social and Behavorial Sciences* (pp. 215–230).

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Pearson Educational Measurement. (2005). *Recent Trends in Comparability Studies*. PEM Research Report 05-05. Austin, TX: Author

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26*. *Psychometrics,* (pp.125–167). New York, NY: Elsevier.

Poggio, J., Glasnapp, D. R., Yang, X., & Poggio, A. J. (2005). A Comparative Evaluation of Score Results from Computerized and Paper & Pencil Mathematics Testing in a Large Scale State Assessment Program. *The Journal of Technology, Learning, and Assessment, 3* (6).

Pommerich, M, & Burden, T. (2000, April). *From Simulation to Application: Examinees React to Computerized Testing.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning, and Assessment, 2* (6), 3-44.

Pommerich, M. (2007). The effect of using item parameters calibrated from paper administrations in computer adaptive test administrations. *The Journal of Technology, Learning, and Assessment, 5* (7). Retrieved from http://files.eric.ed.gov/fulltext/EJ838609.pdf

Pomplun, M., & Custer, M. (2005). The Score Comparability of Computerized and Paper-and-Pencil Formats for K-3 Reading Tests. Journal of Educational Computing Research, 32 (2), 153-166.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The Score Equivalence of Paper-and-Pencil and Computerized Versions of a Speeded Test of Reading Comprehension. *Educational and Psychological Measurement, 62,* 337–354.

Puhan, G., Boughton, K., & Kim, S. (2007). Examining Differences in Examinee Performance in Paper and Pencil and Computerized Testing. *The Journal of Technology, Learning, and Assessment, 6* (3). Retrieved from http://www.jtla.org

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87* (3), 517–529.

Rasch, G. (1960). *Probabilistic Models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114* (3), 552–566.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of International Large-Scale Assessment: Backround, Technical Issues, and Methods of Data Analysis*. CRC Press, Boca Raton.

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-Based Testing and Validity: A Look Back and Into the Future. *Assessment in Education, 10* (3), 279-293.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of Reading and Listening Comprehension Across Test Media. *Educational and Psychological Measurement, 71* (5), 849–869. Retrieved from http://epm.sagepub.com/content/71/5/849

Schroeders, U., & Wilhelm, O. (2010). Testing Reasoning Ability with Handheld Computers, Notebooks, and Paper and Pencil. *European Journal of Psychological Assessment, 26* (4), 284–292.

Schroeders, U. (2009). Testing for equivalence of test data across media. In F. Scheuermann, & J. Björnsson (Eds.), *The transition to computer-based assessment. Lessons learned from the PISA 2006 computer-based assessment of science (CBAS) and implications for large scale testing* (pp. 164–170).

Schwarz, R. D., Rich, C., & Podrabsky, T. (2003, April). *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests*. Paper presented at Annual Meeting of the National Council on Measurement in Education, Chicago.

Sireci, S.G., & Zenisky, A. L. (2006). Innovative Item Formats in Computer-Based Testing: In Pursuit of Improved Construct Representation. In Downing, S. M. & Haladyna, T. M. (Eds.), *Handbook of test development* (pp. 329–347).

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30* (4), 443–464.

Vispoel, W. P. (2000). Reviewing and Changing Answers on Computerized Fixed-Item Vocabulary Tests. *Educational and Psychological Measurement, 60* (3), 371–384.

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.

Wang, S. (2004). *Online or Paper: Does Delivery Affect Results? Administration Mode Comparability Study for Stanford Diagnostic Reading and Mathematics Tests*, San Antonio, Texas: Harcourt.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of Computer-Based and Paper-and-Pencil Testing in K 12 Reading Assessments: A Meta-Analysis of Testing Mode Effects. *Educational and Psychological Measurement, 68* (1).

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria, and an example. *Journal of Educational Measurement, 38,* 19-49.

Yamamoto, K. (2012). *Outgrowing the Mode Effect Study of Paper and Computer Based Testing.* Retrieved from http://www.umdcipe.org/conferences/EducationEvaluationItaly/COMPLETE_PAPERS/Yamamoto/YAMAMOTO.pdf

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.