Educational and Psychological
Measurement
71(5) 849–869
© The Author(s) 2011
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164410391468
http://epm.sagepub.com
SAGE

Equivalence of Reading and Listening Comprehension Across Test Media

Ulrich Schroeders[1] and Oliver Wilhelm[1]

Abstract

Whether an ability test delivered on either paper or computer provides the same information is an important question in applied psychometrics. Besides the validity, it is also the fairness of a measure that is at stake if the test medium affects performance. This study provides a comprehensive review of existing equivalence research in the field of reading and listening comprehension in English as a foreign language and specifies factors that are likely to have an impact on equivalence. Taking into account these factors, comprehension measures were developed and tested with $N = 442$ high school students. Using multigroup confirmatory factor analysis, it is shown that reading and listening comprehension both were measurement invariant across test media. Nevertheless, it is argued that equivalence of data gathered on paper and computer depends on the specific measure or construct, the participants or the recruitment mechanisms, and the software and hardware realizations. Therefore, equivalence research is required for specific instantiations unless generalizable knowledge about factors affecting equivalence is available. Multigroup confirmatory factor analysis is an appropriate and effective tool for the assessment of the comparability of test scores across test media.

Keywords

equivalence, comparability, MGCFA, CBA

Introduction

In the past decade, the measurement of language proficiency has changed considerably. Many commercial vendors have transposed their measurement instruments from paper-and-pencil to computerized tests. In 1998, the paper-based TOEFL (Test of

[1] Humboldt-Universität zu Berlin, Berlin, Germany

Corresponding Author:
Ulrich Schroeders, Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
Email: ulrich.schroeders@iqb.hu-berlin.de

English as a Foreign Language) was replaced by a computerized version (Educational Testing Service, 2001) that is now exclusively offered via the Internet (Chapelle, Enright, & Jamieson, 2008). In 2005, a computerized version of the academic part of the IELTS (the CB IELTS, Computer-Based International English Language Testing System) was introduced, and despite low correlations between both forms, test scores are used interchangeably (Blackhurst, 2005). The process of transition to another test medium is mainly motivated by anticipated financial and administrative advantages. Even though precise cost–benefit calculations are more complex than it seems at first glance (Farcot & Latour, 2009), the mid- and long-term costs can be reduced when considering the whole testing cycle from item development and revision, over compilation and administration of the test, up to scoring and reporting test takers' performance. In national and international large-scale assessment, stakeholders' attitudes toward transition of assessment to computers seem to be more restrained and cautious. Apart from some piloting efforts (e.g., Computer-Based Assessment of Science; Halldórsson, McKelvie, & Björnsson, 2009), the transition to technology-based testing is everything but swift. Compared with commercial agencies offering language tests under prespecified hardware and software instantiations, educational researchers usually have to revert to the preexisting diverse information technology (IT) infrastructure in schools. In the case of the "National Assessment of Educational Progress" (NAEP) program, such organizational factors lead to the evaluation that transition and short-term operating costs for electronic assessment are substantial (Sandene et al., 2005). Nonetheless, two arguments may outweigh the skepticism evoked by the required effort and initial costs. First, the way in which we gather, process, and store knowledge has changed with the development of IT. In a school context, teachers make use of multimedia to impart knowledge in classrooms, and students use computers and the Internet for various purposes such as completion of their homework. And because computers become increasingly important in daily life and academic success, it is reasonable to measure and document student achievement on computers. Thus, to stay abreast of social and technological changes, it may be inevitable to implement computers in the near future into educational large-scale assessment. Second, the use of computers often embed the hope to measure constructs that are hard to access (e.g., problem-solving skills) or that are not adequately accessible on paper (e.g., IT literacy). This would constitute a significant gain in psychological and educational assessment rather than improving already existing measures. Although it seems to be a matter of time until computers are used extensively as a test medium, paper-and-pencil and computer tests will likely coexist for a long time. Therefore, it is vital to understand the comparability of test scores across test media more profoundly and to consider the implications caused by a lack of equivalence on diagnostic decisions (e.g., admission into a naturalization and immigration process). There is a plethora of studies evaluating the comparability of data gathered on paper versus computer. Table 1 summarizes some recent comparability studies in the field of reading comprehension (RC) and listening comprehension (LC).

Table 1. An Overview of Recent Empirical Studies Comparing Paper-Based Versus Computer-Based Measures

Authors	Year[a]	Sample	Design	Method	Key findings
Reading comprehension					
Choi, Kim, and Boo	2003	258 university students	Within	Mean group comparison (t tests) ANOVA Invariance testing with CFA	Strong main effect for mode in ANOVA, $F(1, 256) = 156.62$, $p < .01$, $ES = 0.69$, $N = 258$, higher scores for PBT r (PBT – CBT) = .63 ($N = 132$; $N = 126$) Equivalence of factor loadings (weak invariance)
Higgins, Russell, and Hoffmann	2005	219 fourth grade students	Between	Mean group comparison Mean differences across media for groups stratified by hand-on skills, computer literacy, and computer use	No significant differences between scores for PBT and both CBT (with and without scrolling) No interaction between test medium and hand-on skills, computer literacy
Kim and Huynh	2008	439 middle or high school students	Within	Comparison of mean scale scores Comparison of item bank parameter and 1-PL ability estimates Construct equivalence Invariance testing with CFA	Significant mode effect in ANOVA for the complete language test, $F(1, 437) = 65.49$, $p < .01$, $ES = 0.23$, higher scores for PBT Some items showed a mode effect that have no impact at the overall test level 1-PL ability estimates higher for PBT in the reading comprehension domain, $t(438) = 12.99$, $p < .01$, $ES = 0.66$ Equivalence of factor loadings and residual variances (strict invariance)
Neuman and Baydoun	1998	411 undergraduate students	Within	Mean group comparison (t tests) Invariance testing with CFA	Equivalence of factor loading pattern (configural invariance) for the whole test battery No significant mean differences across modes r(PBT – CBT) = .84

(continued)

Table 1. (continued)

Authors	Year[a]	Sample	Design	Method	Key findings
Overton, Taylor, Zickar, and Harms	1996	607 and 484 employees (two test batteries)	Within	Correlations between (a) PBT, (b) pen-based computer condition, and (c) keyboard-based computer condition	Disattenuated correlations between PBT and pen-based computer condition were equal to unity for all four power tests, but only in three for keyboard-based condition Disattenuated correlations between PBT and both CBT were equal to unity
Pommerich	2004	1,893 and 3,171 eleventh and twelfth grade students (two studies)	Between	Mean group comparison Comparison of item difficulty across test media	Significant mean differences for Study 1, higher scores for PBT ($t(1892) = 3.14$, $p < .01$, $ES = 0.14$) but not for Study 2; changes made from Study 1 to 2: (a) line break was aligned to the PBT, (b) scrolling speed was increased, and (c) training of scrolling Number of items with differences in difficulty higher than expected but negligible effect sizes
Pomplun, Frey, and Becker	2002	121 and 92 college students (two parallel test forms)	Within/ between	Comparison of means and variances Raw score and attenuated correlations Equating conversions Predictive validity (with course grades)	Ambiguous results concerning mean differences for the two test forms: significant mean difference across modes for only one test form ($t(120) = -2.19$, $ES = -0.21$), higher scores for CBT No differences in variances for reading comprehension scales $r(\text{PBT} - \text{CBT}) = .69$ and $r(\text{PBT} - \text{CBT}) = .76$ Ambiguous results regarding equating No significant differences in predictive validity

(continued)

Table 1. (continued)

Authors	Year[a]	Sample	Design	Method	Key findings
Listening comprehension					
Choi, Kim, and Boo	2003	258 university students	Within	Mean group comparison (t tests) ANOVA Invariance testing with CFA	Main effect for mode in ANOVA, $F(1, 256) = 57.72$, $p < .01$, $ES = -0.35$, $N = 258$, higher scores for CBT r(PBT – CBT) = .76 (N = 132) and r(PBT – CBT) = .72 (N = 126) Equivalence of factor loadings (weak invariance)
Coniam[b]	2006	115 high school students	Within	Mean group comparison Rank order comparison	Significant mean differences, higher scores for CBT, $t(114) = -5.16$[c], $p < .01$, $ES = 0.46$ r(PBT–CBT) = .76 Inconclusive results regarding comparability across test media
Neuman and Baydoun	1998	411 undergraduate students	Within	Mean group comparison (t tests) Invariance testing with CFA	Equivalence of factor loading pattern (configural invariance) for the whole test battery No significant mean differences across modes r(PBT – CBT) = .91 (subtest: *Oral Direction*)

Note. PBT = paper-based testing; CBT = computer-based testing; ANOVA = analysis of variance; 1-PL = one-parameter logistic; CFA = confirmatory factor analysis; ES = effect size.

a. Year of publication.

b. In this overview, Coniam's (2006) study was the only one using adaptive item presentation.

c. The degrees of freedom have been corrected to 114. Inferential statistics are provided for significant differences in means or variances. Nonsignificant results are reported in text form only. To foster comparability of the studies, t-test statistics and *ES* are reported consistently in the form that positive values indicate higher scores for PBT.

Research assessing the equivalence of LC tasks is sparse and inconclusive (Choi, Kim, & Boo, 2003; Coniam, 2006). Despite the fact that studies assessing RC outnumber those measuring LC, results are inconsistent in both domains of comprehension. A recent meta-analysis for English RC covers 36 data sets (6 data sets were excluded to eliminate effect size heterogeneity) from 11 primary studies of the past 25 years (Wang, Jiao, Young, Brooks, & Olson, 2008). The weighted mean effect size of all studies was not statistically different from zero. Three postulated moderator variables (grade level, type of test, and whether the test was delivered via the Internet) had no statistically meaningful influence, whereas four other moderator variables (study design, sample size, computer delivery algorithm, and computer practice) affected the differences in RC scores between test media. Obviously, the small number and the selection of studies included in this meta-analysis limit the significance and generalizability of the results.

In specific cases, substantial differences between paper- and computer-based testing may occur depending on the specific measure, the participants, and the soft- and hardware realizations. For instance, Bridgeman, Lennon, and Jackenthal (2003) compared students' reading scores in a computer-based test as a function of different screen sizes and resolutions. They found that small screens at low screen resolution impair reading performance and reasoned that scrolling caused the differences in performance. This substantiated the assumption of a similar study that did not reach significance, supposedly because of the small sample size (Higgins, Russell, & Hoffmann, 2005). The mean performance gap between a paper- and computer-based reading condition can even be removed by carefully aligning the typeset between both conditions, increasing scrolling speed, and explicitly instructing test takers how to use the sliding scroll bar (Pommerich, 2004). Other researchers have suggested that the response procedures, and not the characteristics of the presentation (e.g., screen resolution), are decisive for differences in reading performance across media (Pomplun, Frey, & Becker, 2002). Clicking the correct answer with a mouse is more time-consuming than ticking the solution on a sheet with a pen. Especially with speeded measures, this extra time is a disadvantage for participants completing a computerized test version unless scores are corrected for speededness. Accordingly, Overton, Taylor, Zickar, and Harms (1996) showed higher disattenuated cross-mode correlations between a paper-and-pencil RC task and its computerized pen-based counterpart than between the paper test and a computerized version that used keyboard entries. Using a keyboard requires motor skills that are different from the normal answer format (Neuman & Baydoun, 1998). This motor restriction applies to both RC and LC.

Additionally, computer familiarity is often discussed as a factor affecting test scores (Leeson, 2006). However, the construct of computer familiarity is not clearly defined and may reflect the ability to cope with and handle the perceptual and motor skill limitations imposed by computerization. For instance, in a recent NAEP study, writing performance was compared on paper versus computers. Computer familiarity was assessed as (a) hands-on computer proficiency, (b) extent of computer use in general, and (c) computer use for writing in particular was significantly related to computer-based writing performance after controlling for the paper-based performance (Horkay, Bennett, Allen, & Kaplan, 2005). For such tasks that afford entering text, test media–related differences in performance are more conceivable than for a conventional RC task. Studies varying line length, foreground and background color, and contrast found little to no significant influence on reading rate and comprehension (e.g., Clausing & Schmitt, 1990).

The results we reported so far are predominantly focusing on the level of mean differences and differential validity in between-group designs. Note that this is based on a specific notion of the term equivalence. Questions concerning equivalence have been answered heterogeneously because the methods used address different facets of the term. For example, in the 'Standards for Educational And Psychological Testing' (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the issue of equivalence is addressed in sections on score comparability (p. 49), test administration (p. 61), and fairness of testing (p. 71). We argue that scores of measures are equivalent if they capture the same construct with the same measurement precision. Therefore, in unbiased measurement, apart from random errors, test scores are completely dependent on the ability to be measured and unaffected by the means of measurement. According to this definition, evaluating any potential test medium effect solely on the basis of mean comparisons is insufficient. From this perspective, the common procedure of analyzing mean differences (see Table 1) is based on the untested assumption that sources of variances are the same within and across media. Obviously, it is preferable to apply methods that enable adequate testing of these assumptions (Lubke, Dolan, Kelderman, & Mellenbergh, 2003). A more exhaustive test of equivalence is achieved if additionally bi- or multivariate relations are investigated. This can be accomplished by applying confirmatory factor analysis (CFA) in either within-subject or between-subject designs (see Schroeders, 2009).

Equivalence research in the field of language testing often used within-subject designs where test forms with different items—supposedly drawn at random from an item sample—are compared with each other (Choi et al., 2003; Kim & Huynh, 2008; Neuman & Baydoun, 1998). Here, test media effects can be modeled directly as nested method factors (e.g., with the correlated-trait–correlated-method minus one model; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Schroeders & Wilhelm, 2010). In a between-subject design, data gathered with the same test form in distinct groups are analyzed with multigroup confirmatory factor analysis (MGCFA) constituting an extension of CFA (Meredith, 1993). MGCFA is a suitable method to provide the necessary within- and between-group comparisons (Lubke et al., 2003). Relative to between-subject designs, within-subject designs have two advantages. First, their statistical power is (other things being equal) higher because each subject serves as his or her control, which reduces error variance (Venter & Maxwell, 1999). Second, within-subject designs are not dependent on unbiased assignments of persons to groups. To compensate for these disadvantages in a between-subject design, it is ceteris paribus necessary to test larger samples and to ensure that the groups are comparable with respect to variables associated with the construct in question. For example, when measuring English proficiency, the groups have to be comparable with respect to the number of years of English education. On the other hand, the threat of undesired learning or sequence effects—especially in settings akin to a retest situation—is stronger in within-subject than in between-subject designs. In this study, we used a between-subject design, where the required homogenization across groups was achieved by randomly assigning students within school classes to test conditions (see Design section).

Research Questions

To convert traditional paper-based comprehension tasks to a computerized form and to guarantee the generalizability of information on validity and the comparability of test performance across test media, it is pivotal to investigate their measurement equivalence. The question of equivalence is not restricted to the transition period in which tests will be delivered on both media. Furthermore, knowledge about a potential effect of test media will allow for long-term comparisons across different modes of administration and, thus, ensure continuity of ongoing research. In most large-scale studies, measurement instruments are compiled by drawing items from a large item database. Besides linguistic information (e.g., genre and topic of texts) and information concerning the layout (e.g., text length, size of a table, or diagram), such an item database contains statistical information about the items (difficulty and discrimination parameter). All this information is necessary for test compilation. However, if a measure has originally been developed for administration on paper, the statistical characteristics are bound to this test medium, and it remains questionable whether the parameters, including norm data, can be transferred one-to-one if the test is adapted for computers.

In this article, we want to examine whether the measurement of comprehension skills in English as a foreign language is affected by test medium. We assess both RC and LC on paper and with computers in a between-subject design. The computer-based test was designed with the intention to minimize sources of potential differences between the conditions. For instance, long text passages that would require scrolling were avoided and, to prevent differential speededness of diverging response modes, only multiple-choice (MC) format was implemented. The main objective of this article is to establish which psychometric aspects are affected to what degree by transferring measures for testing English as a foreign language comprehension from paper to computerized administration. Therefore, the focus of the feasibility study is on assessing the equivalence of the measurement across media with an appropriate statistical method, that is, invariance testing by means of MGCFA.

Methods

Participants

A total of 442 German high school students of intermediate-track *Realschule* (n = 195) and academic-track *Gymnasium* (n = 247) participated in this study. The sample was range-restricted to a certain degree because only educational institutions at the upper end of the ability spectrum were included. Students were in the fifth or sixth year of their foreign language education. About 73% of the sample attended ninth grade (*n* = 324), and the remaining quarter attended 10th grade (*n* = 118). In all, 41% of the participants were female (*n* = 181). Mean age was about 16 years (*M* = 15.9, *SD* = 0.70, range 14.7-18.1 years). Participation was mandatory for all students.

Measures

Participants worked on measures of RC and LC that are based on the national educational standards for English as the first foreign language. For this study, items were drawn from a larger database of field-tested items (Rupp, Vock, Harsch, & Köller, 2008). The item database did not have a sufficient number of items that we considered suitable for computerization—that is, items that fitted on a single computer screen without scrolling and had a MC response format. Therefore, we had to adapt some of the items with respect to these constraints (e.g., change the response format). As a consequence of this modification, no prior information about psychometrics of modified items was available. Therefore, we deemed it acceptable to search, identify, and, if necessary, exclude items with inadequate statistical properties from further analysis. All items were scored dichotomously. Following the ability tests, participants completed a sociodemographic background questionnaire and a computer usage questionnaire.

Reading comprehension. The RC measure consisted of 12 short texts (between 76 and 224 words, average of 128 words) that fitted on a computer screen. The text passages described various topics, for instance, one text listed the places in Vancouver city and the surrounding area where different sporting events of the Olympic Winter Games took place. One half of the items were followed by only 1 question and the other half had a testlet structure offering between 3 and 5 questions. In total, the RC task consisted of 24 MC questions. Participants were given the opportunity to go back and forth within the RC part to review and change previous answers, although there is good empirical evidence that the majority of examinees will change only a few responses (Revuelta, Ximénez, & Olea, 2003). The functionality was mainly added to align the computer condition as closely as possible to the paper condition. Test time added up to 25 minutes, including a sample item and instructions.

Listening comprehension. The LC task consisted of 10 audio tracks—lasting between 20 seconds and 2 minutes 40 seconds—covering real-life scenarios. For example, among the 33 items were a telephone call between two friends and a radio commercial soliciting donations for local libraries. A sample item made participants familiar with the task and served to adjust the volume. Prior to presenting the audio tracks, participants were given time to read the questions, thus minimizing the effect of memory on comprehension ability. While listening to the audio tracks, participants saw all MC questions belonging to a specific item (between 1 and 6). In the computerized version, all questions fitted on one screen, so that neither scrolling nor paging was necessary to answer an item. After the scheduled time for each item elapsed, the computer program automatically jumped to the next item. In the paper-based condition, audio recordings were presented via a portable CD player and participants had to mark the correct answer in a booklet. Overall, the LC task took 25 minutes, including the sample item and instructions.

Design

The study was embedded in the process of assessing educational progress in Germany and establishing national performance scales in English as a foreign language. The Data Processing Center of the International Association for the Evaluation of Educational Achievement selected the schools participating in the study and proctored the tests in schools according to guidelines provided by the authors. To control for any hardware or software effects, identical laptops (Fujitsu Siemens Esprimo mobile U9200 with a 12.1-inch liquid crystal display) with the same test environment were brought into school. For reasons of feasibility and affordability, only schools from one northern state of Germany were recruited. To minimize cluster effects and to account for specificities of the between-subject design, one randomly assigned half of each class worked first on paper and then on computer ($n = 234$). The other half completed the test conditions in reverse order ($n = 208$). Within the two groups, half of the participants started with the RC task ($n = 221$), whereas the other half started with the LC task ($n = 221$). The resulting four groups are balanced with respect to order of test media and order of comprehension skill.

Method of Item Selection

Item selection was necessary because the response format of some items was changed from short-answer or mapping to MC format (see Measures section), and these modified parts of the item database were not sufficiently field tested. Two item characteristics were considered in the item selection process: Either the item showed extreme difficulty (i.e., $p < .04$ or $p > .96$) or the fit of a single-factor model considerably improved after excluding a specific item. All measurement models were computed with Mplus 5.21 (L. K. Muthén & Muthén, 2009) and were based on the weighted least squares mean and variance adjusted (WLSMV) estimator. The WLSMV estimator was chosen for the analyses because the data of the present study were categorical, and simulation studies have shown the superiority of WLSMV estimator compared with maximum likelihood estimator both in terms of model rejection rates and appropriate estimation of factor loadings for this type of data (Beauducel & Herzberg, 2006).

Following a recommendation by Hu and Bentler (1999), we used a two-index strategy that combines an absolute fit index such as the root mean square error of approximation (RMSEA) and the weighted root mean square residual (WRMR), respectively, with an incremental fit index such as the comparative fit index (CFI) to evaluate model fit. For categorical data, Yu (2002) reported the following cutoff values as indicators of good model fit: CFI ≤ .96, RMSEA ≤ .05, and WRMR ≤ .95. In contrast to these goodness-of-fit statistics, neither χ^2 statistics nor the degrees of freedom (*df*) are decisive indicators of model fit because in WLSMV estimation the degrees of freedom are estimated rather than computed.

In the first step, item selection was conducted separately for each task and each medium, resulting in four independent processes of item selection. In the second step, all items that had to be excluded for one test form were also removed from the counterpart. That is, if an item of the paper-based LC task led to model misfit, it was also deleted from the computerized counterpart of the task. Because MGCFA presupposes a common item pool, it was necessary to combine the selection processes that were conducted independently across media.

Method of Invariance Testing

We tested for mean differences with a *t* test for independent samples and for differences in variances with Levene's test of homogeneity of variances. However, to guarantee comparability of test scores across test media, it is insufficient to compare mean scores and dispersions. Two tests can possess the same score distribution characteristics and measure different constructs and, vice versa, two measures can assess the same construct and show different means. A suitable method to check for such differences in measurement is MGCFA. In this framework, item responses are conceptualized as a function of three parameters: (a) factor loadings, (b) intercepts/thresholds (continuous/categorical variables), and (c) residuals variances. By constraining parameters across test media to equality, different forms of invariance can be assessed that allow for different forms of comparisons. If the constraints that are imposed on the parameters are violated in terms of deterioration in model fit relative to less constrained models, various forms of invariance can be inferred.

Different levels of invariance can be assessed with a straightforward procedure of comparing measurement models in a fixed order, from the least to the most restrictive model (see Table 2). Each model is nested within the previous ones, for example, Model A3 can be derived from Model A2 by imposing additional constraints on the intercepts. Because of this nested character, a potential deterioration in model fit is testable through a χ difference test (Bollen, 1989; Mplus offers a special DIFFTEST option that allows for testing nested model fit with WLSMV estimator; L. K. Muthén & Muthén, 2007).

It should be noted that equality of item parameters can also be tested with other methods such as differential item functioning. Compared with this item response theory framework, the CFA approach seems to be stricter in rejecting measurement invariance (for a detailed comparison, see Raju, Laffitte, & Byrne, 2002). Moreover, CFA seems to provide further extensions of measurement and structural models—for example, when it comes to account for nonequivalence or analyzing change in latent variables.

Table 2. Testing for Measurement Invariance With Continuous and Categorical Data

(A)	Continuous variables	Factor loadings	Intercepts	Residual variances	Factor means
(A1)	Configural invariance	*	*	*	Fixed at 0
(A2)	Weak invariance	Fixed	*	*	Fixed at 0
(A3)	Strong invariance	Fixed	Fixed	*	Fixed at 0/*
(A4)	Strict invariance	Fixed	Fixed	Fixed	Fixed at 0/*
(B)	**Categorical variables**	**Factor loadings**	**Thresholds**	**Residual variances**	**Factor means**
(B1)	Configural invariance	(*	*)	Fixed at 1	Fixed at 0
(B2)	Strong invariance	(Fixed	Fixed)	Fixed at 1/*	Fixed at 0/*
(B3)	Strict invariance	(Fixed	Fixed)	Fixed at 1	Fixed at 0/*

Note. The asterisk (*) indicates that the parameter is freely estimated. Fixed = the parameter denominated in the title of the column is fixed to equity across groups; Fixed at 1 = the residual variances are fixed at 1 in both groups; Fixed at 1/* = the residual variance is fixed at 1 in one group whereas freed in the other group; Fixed at 0 = factor means are fixed at 0 in both groups. Fixed at 0/* = the factor mean is fixed at 0 in one group and freed in the other. Parameters in parentheses need to be varied in tandem (for additional remarks see text).

Invariance Testing With Categorical Data

The steps of invariance testing differ from the more familiar case of invariance testing with continuous variables (B. O. Muthén & Christoffersson, 1981; B. O. Muthén & Asparouhov, 2002). To illustrate the differences in invariance testing, both are listed in Table 2. Assessing invariance with categorical variables requires varying factor loadings and thresholds in tandem because item characteristic curves are based on both parameters. Therefore, the second step of invariance testing for continuous indicators is skipped in invariance testing for categorical indicators (L. K. Muthén & Muthén, 2009). The invariance testing with categorical data comprises three steps: In Step B1, all measurement parameters (factor loadings, thresholds, residual variances) are freely estimated in both conditions (paper-and-pencil and computer). Model B1 checks for configural invariance where the pattern of loadings is more decisive than their actual magnitude. In Step B2, strong invariance, the models are invariant with respect to both their factor loadings and thresholds, whereas residual variances are fixed at 1 in one group and freed in the other. A prerequisite for meaningful cross-group comparisons (Bollen, 1989) is that the rank order of individuals is not affected by the mode of administration. With continuous variables, this holds if measurement invariance is established on the level of weak (or metric) invariance (Step A2). However, for categorical data there is no direct equivalent, so that the standard is not met until the stage of strong invariance. If strong invariance holds, it is commonly assumed that the ability scores can be compared directly because they are corrected for measurement error. According to the rationale, the most restrictive model (Step B3) in which all measurement parameters are held to be equal is considered unduly strict and inappropriate for many practical scenarios (e.g., Vandenberg & Lance, 2000). However, Deshon (2004) indicated that there are two possible sources of group difference in the residual variances—"random noise" and variance of unintentionally measured variables that are not (yet) covered by the model. The presumption that diverging residual variances are only indicative for diverging reliabilities of the observed scores is only true if there are neither correlations among the item residuals nor correlations between the item residuals and the common factors (for a more detailed discussion on this topic, see Wu, Li, & Zumbo, 2007). If there is no so-called conditional independence across groups, the measure will be biased in some respect.

Delta Versus Theta Parameterization

The software package Mplus (L. K. Muthén & Muthén, 2009) offers two parameterizations that impose different constraints to identify factor model parameters—that is, the Delta and the Theta parameterization (for a detailed description of the two approaches, see Millsap & Yun-Tein, 2004; B. O. Muthén & Asparouhov, 2002). To align the steps of invariance testing for categorical data to those for continuous data and to omit some interpretational problems (i.e., the interpretation of scale factors that are functions of factor loadings, factor variances, and residual variances), we concentrate on the Theta parameterization. We also controlled the findings with the alternative parameterization—the Mplus default.

Results

All measurement models listed in Table 3 are single-factor two-parameter logistic models. Neither correlations between the residual terms nor negative residual variances were allowed. After removing one invalid observation in the computer-based LC task that was considerably contributing to the skewness of score's distribution, all measurement models yielded good fit as indicated by the *p* value, CFI, RMSEA, and WRMR. Model fit was good even without explicitly accounting for the testlet structure of the measures. Local dependencies, therefore, only played a minor role. After combined item selection, the RC task consisted of 21 items and the LC task 23 items. It should be noted that all items that had to be removed because of extreme item difficulty were very simple items (i.e., $p > .96$).

Descriptive Statistics

Table 4 summarizes descriptive statistics of the final test forms.

The range restriction of the sample (see Participants section) was partly responsible for the high rates of correct responses. The average item difficulty was higher for paper-based RC (PB-RC) than for the corresponding computerized version (CB-RC); p_{mean}(PB-RC) = .78 versus p_{mean}(CB-RC) = .74; $t(440) = 2.33$, $p = .02$. For LC, there was no difference in mean item difficulty across test media; p_{mean}(PB-LC) = .78 versus

Table 3. Measurement Models With Separated and Combined Item Selection Process

			Number of items excluded because of							
	n	n_{item}[a]	$p(x) > .96$[b]	Misfit[c]	χ^2	df	p	CFI	RMSEA	WRMR
(A) Separated item selection process										
PB-RC	221	21/24	3	—	68.8	64	.32	.974	.018	.851
CB-RC	221	22/24	2	—	90.0	74	.10	.930	.031	.892
PB-LC	221	26/33	4	3	74.1	70	.35	.929	.016	.887
CB-LC	220	25/33	3	5	78.3	70	.23	.930	.023	.896
(B) Combined item selection process										
PB-RC	221	21/24	3[d]	—	68.8	64	.32	.974	.018	.851
CB-RC	221	21/24	3	—	89.3	76	.14	.944	.028	.868
PB-LC	221	23/33	5	5[e]	68.4	68	.46	.994	.005	.858
CB-LC	220	23/33	5	5	70.6	66	.33	.960	.018	.865

Note. All models used WLSMV (weighted least squares mean and variance adjusted) estimator for categorical data (see additional remarks in the text). PB-RC = paper-based reading comprehension; CB-RC = computer-based reading comprehension; PB-LC = paper-based listening comprehension; CB-LC = computer-based listening; CFI = comparative fit index; RMSEA = root mean square error of approximation; WRMR = weighted root mean square residual.

a. Number of items after/before item selection.

b. Number of items excluded because of extreme difficulty, $p(x) > .96$.

c. Number of items excluded because of model misfit.

d. Number of items excluded because of extreme difficulty, $p(x) > .96$ in at least one condition.

e. Number of items excluded because of model misfit in at least one condition.

p_{mean}(CB-LC) = .77; $t(439) = –0.32$, $p = .75$. Variances were found to be equal across test medium; RC, $F(1, 439) = 1.76$, $p = .19$; LC, $F(1, 438) = 1.31$, $p = .25$. Because of the dichotomous character of the data, we computed Cronbach's alphas that were based on the tetrachoric correlation matrix. Cronbach's alpha marks a lower bound of a reliability estimate (for a detailed discussion, see Zinbarg, Revelle, Yovel, & Li, 2005) and ranged between $.74 \le \alpha \le .85$. To evaluate the reliabilities of the latent scales, we estimated the reliabilities for the latent scales using McDonald's omega (ω; McDonald, 1999). These reliability estimates varied between $.80 \le \omega \le .87$ and turned out to be higher for RC than LC. Item difficulties correlated highly between paper- and computer-based RC task ($r = .91$) and LC task ($r = .92$). Because the design of this study included a change of both medium and task (see the Design section), it is not possible to estimate the correlation between RC and LC within a test medium but only across media. The cross-media correlations between RC and LC are ρ(CB-RC, PB-LC) = .75 and ρ(PB-RC, CB-LC) = .80.

Invariance Testing

Table 5 shows the results of the invariance testing for RC and LC. In the first step (RC1/LC1), all measurement parameters (factor loadings, thresholds, residual

Table 4. Descriptive Statistics for the Test Forms After Item Selection

	n	Mean	Median	*SD*	*SE*	Minimum	Maximum	Range	SIQR	$M_\lambda(SD_\lambda)$[a]	Skewness	Kurtosis	α[b]	ω[c]
PB-RC	221	16.3	17	3.2	.22	7	21	14	2.5	.47 (.14)	-0.58	-0.20	.83	.86
CB-RC	221	15.5	16	3.6	.24	4	21	17	2.5	.48 (.14)	-0.76	0.54	.85	.87
PB-LC	221	17.8	18	2.7	.18	8	23	15	2	.38 (.16)	-0.55	0.77	.74	.80
CB-LC	220	17.9	18	3.0	.20	8	23	15	2	.43 (.16)	-0.89	0.84	.81	.85

Note. PB-RC = paper-based reading comprehension; CB-RC = computer-based reading comprehension; PB-LC = paper-based listening comprehension; CB-LC = computer-based listening comprehension; *SE* = standard error of mean; SIQR = semi-interquartile range.

a. Mean and standard deviation of the factor loadings.

b. Cronbach's alpha.

c. McDonald's omega.

Table 5. Testing for Measurement Invariance for Reading and Listening Comprehension

		χ^2/df	p	CFI	RMSEA	$\Delta\chi^2/df$	p
(RC)	Reading comprehension						
(RC1)	Configural invariance	155.1/138	.15	.960	.024	—	—
(RC2)	Strong invariance	150.1/135	.17	.963	.023	14.9/14	.38
(RC3)	Strict invariance	154.1/138	.17	.962	.023	20.1/18	.33
(LC)	Listening comprehension						
(LC1)	Configural invariance	139.0/134	.37	.972	.013	—	—
(LC2)	Strong invariance	134.3/129	.36	.970	.014	15.5/14	.34
(LC3)	Strict invariance	159.3/134	.07	.859	.029	73.5/21	.00

Note. n(PB-RC/CB-RC) = 221/221, *n*(PB-LC/CB-LC) = 221/220. *df* = degrees of freedom; CFI = comparative fit index; RMSEA = root mean square error of approximation; PB = paper based; CB = computer based.

variances) were freely estimated, assuming configural invariance. In the second step (RC2/LC2), this is comparable with strong invariance, factor loadings and thresholds were set to be equal in tandem across test medium. In addition to the constraints of Step 2, the residual variances are constrained to equity in the last step. The χ^2 difference test is calculated between two consecutive models (e.g., RC1 vs. RC2, RC2 vs. RC3) and denotes the loss in model fit in comparison with the gain in degrees of freedom. For RC, even imposing the most restrictive constraints does not lead to deterioration in model fit. For LC, holding factor loadings and thresholds equal across test media does not lead to a meaningful reduction in model fit, whereas a model in which the residual variances are additionally assumed to be equal is not supported by the data. Scrutinizing the residual correlations contributing to model misfit, deviations were not systematic and almost negligible in number and magnitude. Therefore, residual item variance can be attributed to random error. Ability estimates that are corrected for this random error are comparable across groups. Both measurement instruments are at least strongly invariant across test media. Delta parameterization and Theta parameterization showed nearly identical measurement parameters and model fits for all models.

Discussion

To compare test scores across test media, it is an essential precondition to guarantee that the measurement is not distorted by test medium. Concepts of equivalence diverge widely, and accordingly, different statistical methods have been proposed to test for equivalence. In this article, we argue that comparisons of univariate statistics across test media are insufficient and that it is pivotal to consider the bivariate relations in order to evaluate the comparability of test scores. MGCFA provides a convenient and effective way to test whether a certain measure is invariant across test media. The kind of comparisons and inferences that are admissible depends on the highest level of measurement invariance that can be achieved (see Cheung & Rensvold, 2002). For example, strong invariance is needed to assume that the correlations with external criteria are equal. The advantage of the MGCFA approach in comparison with an approach that considers means and dispersions becomes evident through the data at hand: We found statistically significant mean differences for the RC task in favor of the paper-based version, which is line with findings of other studies reported in the Introduction section (see Table 1). Evaluating the comparability of test scores on the base of means and dispersions is, however, insufficient, because such comparisons do not address the variance–covariance structure. A method tapping this structure, and which is therefore qualified in assessing interindividual differences, is invariance testing by means of MGCFA. Thus, we could establish measurement invariance. This means that the relative standing of any two individuals is independent of test media. As the residual variances are equal across test media, even test raw scores of the RC task can be converted directly into each other by adding or subtracting the mean difference as a constant term. Compared with previous LC studies (Choi et al., 2003; Coniam, 2006; Neuman & Baydoun, 1998; see Table 1), we found no higher scores for the computerized version—at least on a manifest level. In addition to identical means, the LC measure used in this study showed identical standard deviations across media. However, this is no definite evidence as to the equivalence of test scores. It is possible that two measurement instruments possess identical parameters concerning the score distribution and, nevertheless, do not deliver identical test scores. This was the case for the LC task in the present study that was strongly invariant across test media. That means, the rank order of participants was not affected by test media but both instruments measured with different reliabilities. The ability estimates that are corrected for measurement error are equivalent for the paper- and the computer-based LC task. However, it is not possible to interchange raw test scores, because the measurement error (or residual term) is affected by test medium. Nevertheless, factor scores that are corrected for measurement errors are comparable. Using comparisons of means and variances, the comparability of latent ability scores would have gone unnoticed because they do not take the residual variances into account. Gauging the effect of measurement error on test scores by considering the reliability of the measures often leads to uncertain statements. A final advantage of the MGCFA approach that could not be demonstrated with the present data is its ability to reflect more complex structures such as higher order or nested factor models.

There is a huge research literature comparing different measures across different software and hardware realizations for different groups of participants. The cross-mode differences mostly found in these studies are small to negligible. Nevertheless, significant differences do occur in some studies (e.g., Choi et al., 2003, see Table 1). In this study, we found two specific measures assessing comprehension skills in English as a foreign language to be at least measurement invariant across test media. This level of invariance was achieved because in the process of transferring a paper-and-pencil test to the computer, we tried to keep differences between the test forms at a minimum by carefully aligning the demands to each other. First, we only used MC items for which the motor skill requirements are comparable, that is, ticking or clicking the correct answer. Second, we deliberately selected short text passages to avoid scrolling that was found to compromise reading performance. If we had chosen items demanding scrolling or text input as a response, we might have obtained different results. More specifically, we would expect an increase in reaction times for the computerized part and worse performance compared with the paper-and-pencil condition. Furthermore, it should be noted that measurement invariance was only achieved after a two-stage item selection process. In the first step, we conducted item selection for each test medium separately. In the second step, we selected only such items for the final item pool that were part of both item sets. This selection process creates the necessary conditions to adequately check for measurement invariance for different instantiations of the same measure. However, the item selection process and the testing for measurement invariance are both based on the same sample. To use the present item set in high-stakes testing, the model testing should be replicated with an independent student sample of high school students. Both the item selection process and the restrictions of items can be seen as limitations of this study. But even if these precautions are followed, we can see no theoretical or empirical framework that guarantees that measures would be invariant across test media. We also caution about generalizing our findings to other tests within the domain of RC and LC without testing. It is likely that two specific realizations will differ one way or the other in terms of the variance–covariance structure and that the cause for such discrepancy will not be obvious. The psychometric information collected with one specific instantiation of a measure cannot be generalized across a substantial range of other instantiations (van Lent, 2008). One conclusion readers might want to derive from this state of affairs is that it is a substantive problem to choose the most adequate instantiation of a specific measure. Accordingly, statements about the comparability of test data are never final and are made in reference to specific instantiations of a measure. It is, therefore, essential to use an appropriate statistical method to evaluate which comparisons are allowed and which are flawed. If comparisons of data across media are intended, it is crucial to go beyond insufficient mean comparisons and tackle the variance–covariance structure. In this article, we demonstrated how different levels of measurement invariance can be assessed with MGCFA and showed which comparisons are feasible depending on the level of measurement invariance that is achieved.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

This work was supported by the Stifterverband für die deutsche Wissenschaft and the Institute for Educational Quality Improvement.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least square estimation in confirmatory factor analysis. *Structural Equation Modeling, 13*, 186-203.

Blackhurst, A. (2005). Listening, reading and writing on computer-based and paper-based versions of IELTS. *Cambridge ESOL Research Notes, 21*, 14-17. Retrieved from http://www.cambridgeesol.org/rs_notes/rs_nts21.pdf

Bollen, K. A. (1989). *Structural equations with latent variables.* Oxford, England: John Wiley.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution and display rate on computer-based test performance. *Applied Measurement in Education, 16*, 191-205.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Choi, I.-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295-320.

Clausing, C. S., & Schmitt, D. R. (1990). Paper versus CRT: Are reading rate and comprehension affected? In *Proceedings of selected paper presentations at the convention of the Association for Educational Communications and Technology*. (ERIC Document Reproduction Service No. ED323924)

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English language listening test. *ReCALL Journal, 18*, 193-211.

Deshon, R. P. (2004). Measures are not invariant across groups with error variance homogeneity. *Psychology Science, 46*, 137-149.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8*, 38-60.

Educational Testing Service. (2001). *Computer-based TOEFL. Score user guide.* Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/989551.pdf

Farcot, M., & Latour, T. (2009). Transitioning to computer-based assessments: A question of costs. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 108-116). JRC Scientific and Technical Reports. Retrieved from http://crell.jrc.it/RP/reporttransition.pdf

Halldórsson, A. M., McKelvie, P., & Björnsson, J. K. (2009). Are Icelandic boys really better on computerized tests than conventional ones? Interaction between gender, test modality and test performance. In F. Scheuermann & J. K. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 178-193). JRC Scientific and Technical Reports. Retrieved from http://crell.jrc.it/RP/reporttransition.pdf

Higgins, J., Russell, M., & Hoffmann, T. (2005). Examining the effect of computer-based passage presentation on reading test performance. *Journal of Technology, Learning, and Assessment, 3*(4). Retrieved from http://escholarship.bc.edu/jtla/

Horkay, N., Bennett, R. E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R. E. Bennett, N. Allen, J. Braswell, B. Kaplan, & A. Oranje (Eds.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005-457). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

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*, 554-570.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*, 1-24.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543-566.

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

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Researc*h*, 39*, 479-515.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Notes: No. 4). Retrieved from http://www.statmodel.com/examples/webnote.shtm

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407–419.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2009). *Mplus* (Version 5.21) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement, 22*, 71-83.

Overton, R. C., Taylor, L. R., Zickar, M. J., & Harms, H. J. (1996). The pen-based computer as an alternative platform for test administration. *Personnel Psychology, 49*, 455-464.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *Journal of Technology, Learning, and Assessment, 2*(6). Retrieved from http://escholarship.bc.edu/jtla/

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.

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*, 517-529.

Revuelta, J., Ximénez, M. C., & Olea, J. (2003). Psychometric and psychological effects of item selection and review on computerized testing. *Educational and Psychological Measurement, 63*, 791-808.

Rupp, A. A., Vock, M., Harsch, C., & Köller, O. (2008). *Developing standards-based assessment tasks for English as a first foreign language. Context, processes, and outcomes in Germany.* Münster, Germany: Waxmann.

Sandene, B., Horkay, N., Bennett, R. E., Allen, N. Braswell, J., Kaplan, B., & Oranje, A. (Eds.). (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project* (NCES 2005-457). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Schroeders, U. (2009). Testing for equivalence of test data across media. In: F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment* (pp. 164-170). JRC Scientific and Technical Reports. Retrieved from http://crell.jrc.it/RP/reporttransition.pdf

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*, 284-292.

van Lent, G. (2008). Important considerations in e-assessment. In F. Scheuermann & A. Guimarães Pereira (Eds.), *Towards a research agenda on computer-based assessment* (pp. 97-103). Ispra, Italy: European Commission, Joint Research Centre.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Venter, A., & Maxwell, S. E. (1999). Maximizing power in randomized designs when sample size is small. In R. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 31-58). Thousand Oaks, CA: Sage.

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*, 5-24.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research, and Evaluation, 12*(3). Retrieved from http://pareonline.net/pdf/v12n3.pdf

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles.

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.