Reliability and Validity of PIRLS and TIMSS: Does the Response Format Matter?

Johannes Schult* and Jörn R. Sparfeldt

Saarland University

Erschienen im European Journal of Psychological Assessment, published online August 3, 2016, © 2016 by Hogrefe Verlag, doi:10.1027/1015-5759/a000338

Author Note

This research was prepared with the support of the German funds “Bund-Länder-Programm für bessere Studienbedingungen und mehr Qualität in der Lehre (‘Qualitätspakt Lehre’)” [the joint program of the Federal and States Government for better study conditions and the quality of teaching in higher education (“the Teaching Quality Pact”)] at Saarland University (funding code: 01PL11012). The authors developed the topic and the content of this manuscript independently from this funding. We thank the Institute for School Development Research (IFS) at Technical University Dortmund / the Max Planck Institute for Human Development (MPIB) Berlin / the Standing Conference of the Ministers of Education and Cultural Affairs (KMK) as well as the Research Data Centre (FDZ) at the Institute for Educational Quality Improvement (IQB) for providing the raw data.

*Corresponding author: Johannes Schult, Bildungswissenschaften, Campus A5 4, Zi. 3.26, Saarland University, D-66123 Saarbrücken, Germany. Phone: ++49 681 302 57482. Fax: ++49 681 302 57488. Email: jutze@jutze.com.

Summary

Academic achievements are often assessed in written exams and tests using selection-type (e.g., multiple-choice; MC) and supply-type (e.g., constructed-response; CR) item response formats. The present article examines how MC items and CR items differ with regard to reliability and criterion validity in two educational large-scale assessments with fourth-graders. The reading items of PIRLS 2006 were compiled into MC scales, CR scales, and mixed scales. Scale reliabilities were estimated according to item response theory (international PIRLS sample; n = 119,413). MC showed smaller standard errors than CR around the reading proficiency mean, whereas CR was more reliable for low and high proficiency levels. In the German sample (n = 7,581), there was no format-specific differential validity (criterion: German grades, $r \approx .5$; $\Delta r = 0.01$). The mathematics items of TIMSS 2007 (n = 160,922) showed similar reliability patterns. MC validity was slightly larger than CR validity (criterion: mathematics grades; n = 5,111; $r \approx .5$, $\Delta r = –0.02$). Effects of format-specific test-extensions were very small in both studies. It seems that in PIRLS and TIMSS, reliability and validity do not depend substantially on response formats. Consequently, other response format characteristics (like the cost of development, administration, and scoring) should be considered when choosing between MC and CR.

Keywords: response format, multiple-choice, constructed-response, item response theory, validity

Reliability and Validity of PIRLS and TIMSS: Does the Response Format Matter?

Introduction

Academic achievements are often assessed in written exams and tests using selection-type and supply-type response formats (cf. Waugh & Gronlund, 2013). Selection-type items (e.g., multiple-choice [MC]) are well-suited for swift administering and automated scoring. Still, there have been doubts about the format's universal usefulness (cf. Bennett, 1993). Although MC tests can be used for various educational outcomes (besides knowledge), they are often associated with learning processes like memorization and surface learning (Martinez, 1999; Scouller, 1998). Critics argue that supply-type items (e.g., constructed-response [CR]) are needed to capture different abilities in complex domains (e.g., Martinez, 1999).

Large-scale assessments like Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Study (TIMSS) use both formats (Mullis, Martin, & Foy, 2008; Mullis, Martin, Kennedy, & Foy, 2007). The scaling procedures of the respective competence tests suggest that MC and CR are measuring one underlying construct in each study (Foy, Galia, & Li, 2007, 2009). It is not yet clear how the two formats differ in terms of measurement precision or criterion-related validity. Therefore, our aim is to examine how MC items and CR items differ with regard to reliability and criterion validity, focusing on reading competence in PIRLS (Study 1) and on mathematics competence in TIMSS (Study 2).

General Response Format Differences

MC items are the dominant subtype of selection-type items. The task is to choose the correct answer(s) from a set of responses. Besides the widespread MC-format "1 out of *x*" (i.e., items with one correct answer and $x - 1$ distractors), other MC formats include, for example, "*x* out of *y*" items (with multiple correct responses) and items with sequentially presented response options (e.g., Haladyna & Rodriguez, 2013). In contrast, supply-type items require the construction of a response by the test taker (e.g., write a short answer or an essay).

Concerning the characteristics and adequacy of different response formats, Waugh and Gronlund (2013, p. 132; see also Haladyna & Rodriguez, 2013; Martinez, 1999) provided a comprehensive overview: These authors compared selection-type items with CR items regarding measured learning outcomes ("good for measuring the recall of knowledge, understanding, and application levels of learning; inadequate for organizing and expressing ideas" vs. "inefficient for measuring the recall of knowledge; best for ability to organize, integrate, and express ideas"), content sampling (large amount of items in MC tests results in "broad coverage, which makes representative sampling of content feasible"), scoring procedure ("objective, simple, and highly reliable" vs. "subjective, difficult, and less reliable"), the validity of the interpretation of the scores (potentially biased by "reading ability and guessing" vs. "writing ability and bluffing"), and likely consequences on learning (induce "to remember, interpret, and use the ideas of others" vs. "organize, integrate, and express their own ideas"). Regarding the relations of the measured learning outcomes with the revision of Bloom's taxonomy (cf. Krathwohl, 2002), selection-type items are particularly suited to assess the recognition of knowledge, understanding, application, analysis, and in some cases evaluation (Haladyna & Rodriguez, 2013). On the other hand, CR items can in addition probe value judgments and combining of ideas (creation). In terms of item writing, the preparation of good items is a sophisticated task for both formats; while some authors argue that it is relatively easier to prepare good CR-items (cf. Gronlund & Waugh, 2013), the more expensive development of a proper scoring procedure along with the actual scoring process for CR-items should be considered. As mentioned, CR items are not necessarily more adequate for assessing more complex mental processes like understanding and application, even though they are more frequently used for that purpose and are less likely to elicit rote learning (cf. Martinez, 1999; Veeravagu, Muthusamy, Marimuthu, & Michael, 2010). In summary, MC and CR have their respective specific advantages (cf. Martinez, 1999; Waugh & Gronlund, 2013).

Concerning the dimensional structure, early studies found no substantial multidimensionality regarding MC and CR in the assessment of reading proficiency and quantitative skills (cf. Traub, 1993). A single-factor model with loadings of both item types tended to fit the data better than a two-factor model with format-specific latent factors (e.g., Bennett, Rock, & Wang, 1991; Thissen, Wainer, & Wang, 1994). In cases where two-factor models fitted better, the MC factor and the CR factor correlated substantially (e.g., $.74 \leq r \leq .94$, Lissitz, Hou, & Slater, 2012; $r \geq .94$, Wan & Henly, 2012). The mean correlation between MC responses and CR responses is particularly large (i.e., close to 1) when both formats use the same item stem (Rodriguez, 2003). In a reading test for 9th-graders a general reading ability factor based on all items correlated substantially with a nested factor for supply-type items ($r = .44$; Rauch & Hartig, 2010).

Response formats can also be compared in an experimental setting with constant item stems and randomly presented response formats. Recent studies on language (Hohensinn & Kubinger, 2011) and using an economics syllabus (Kastner & Stangl, 2011) reported that format-specific measurement models are dispensable and concluded that MC and CR can be used interchangeably in that respect, although MC items tend to be easier than “equivalent” CR items (Chan & Kennedy, 2002; Hohensinn & Kubinger, 2011). Nevertheless, even if the scaling procedure supports a unidimensional assessment, differing response formats may correspond with, besides different item difficulties, differential discrimination parameters and (although less likely with increasing support for a one-factor solution) differential criterion-related validity coefficients.

Format-Specific Reliability Differences

Reliability relates to the extent to which a measurement procedure yields identical results on consistent assessment conditions. An easy but costly way to obtain a more reliable test is to add similar items (Carmines & Zeller, 1979). Adaptive testing procedures are an alternative to test extensions; by choosing items that match each participant's ability level, shorter adaptive tests are usually more reliable than longer traditional tests (cf. Embretson & Reise, 2000).

If we assume (based on the above) that MC items and CR items related to one content domain measure one underlying construct, the question remains whether one format yields more reliable estimates. In unidimensional item response theory (IRT) models (cf. de Ayala, 2009; Embretson & Reise, 2000), reliability corresponds to the test information function $I(\theta)$, which can be written as the inverse of the squared standard measurement error ($I(\theta) = 1/SE(\theta)^2$). The relative efficiency $RE(MC,CR) = I(\theta,MC)/I(\theta,CR)$ compares format-specific scales (de Ayala, 2009). Values larger than 1 suggest that MC is more reliable, values below 1 indicate smaller standard errors for CR. Previous findings regarding format-specific reliability are contradictory. In a study with the Advanced Placement (AP) Chemistry Test (*n* = 18,462), MC items were more efficient than CR items for all proficiency levels (Lukhele, Thissen, & Wainer, 1994). The relative efficiency lay between 8 and 19 in terms of information per minute of testing; suggesting that "for a middle level examinee ($\theta = 0$), one would need about 20 times as much examination time to get the same amount of information with an essay as one would obtain with multiple-choice items" (Lukhele et al., 1994, p. 244). The difference was less pronounced in a study of 4th-, 8th-, and 12th-graders, where the MC items of a computerized K–12 science test provided up to twice as much information per minute of testing as the CR items; only one out of six comparisons showed an advantage of CR over MC (i.e., in grade 8, CR items rated from 0–3 were more efficient than MC items; Wan & Henly, 2012, p. 69). Based on one booklet (#2) of the TIMSS 2007 mathematics test and data from 320 8th-graders, CR items appeared to be more reliable than MC items: RE(MC,CR) = 0.69 (Gültekin & Demirtaşlı, 2012). Still, MC was more reliable for average proficiency levels: RE(MC,CR) > 1 for $-1 \leq \theta \leq 0.5$. Therefore, Gültekin and Demirtaşlı recommended using MC items along with (at least) 40% CR items to achieve a relatively reliable assessment. This advantage may depend on the lack of very easy and very difficult MC items in such assessments (Lee, Liu, & Linn, 2011).

Format-Specific Validity Differences

Validity concerns the adequate and appropriate interpretation and use of test scores with regard to a particular setting (e.g, Messick, 1995; see also American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, pp. 9–24). The correlation between MC scales and CR scales indicates the amount of construct equivalence: If both scales assess the same construct, their correlation (corrected for non-perfect reliability) should approach unity. The findings presented above evidence convergent validity for format-specific factors (e.g., Lissitz et al., 2012; Rodriguez, 2003; Wan & Henly, 2012). Educational large-scale assessments commonly use unidimensional scaling procedures, thereby establishing construct validity across response formats (e.g., Foy et al., 2007, 2009). Further validity evidence is based on the correlations of the scores of a test with external criteria. In educational assessments, corresponding course grades are adequate criteria (e.g., Arnold, Bos, Richert, & Stubbe, 2007; Phan, 2008).

Reliability is a necessary, but not a sufficient prerequisite for validity (Carmines & Zeller, 1979); so format-specific validities do not necessarily reflect the findings regarding reliability. Perhaps a combination of item formats is necessary to cover a sufficiently broad range of a particular cognitive domain (cf. Gültekin & Demirtaşlı, 2012; Martinez, 1999), or CR items might yield a more valid assessment by being more suited to capture higher cognitive processes; of course, this also depends on the item construction rationale (Rauch & Hartig, 2010). According to one study, the correlation of AP test scores with college grade point average (GPA) was larger for MC items than for CR items, but only in American History and Biology ($0.06 \leq \Delta r \leq 0.09$), not in European History and English Language (Bridgeman & Lewis, 1994, p. 42). As with reliability, validity seems to occasionally benefit more from MC items than from CR items. These mixed findings are unsatisfactory and should be substantiated by further research.

PIRLS and TIMSS

Large-scale assessments like PIRLS and TIMSS offer a suitable setting to compare the reliability and the validity of MC and CR formats. They share many methodological features, for example comparable theoretical frameworks, similar sampling strategies, and identical scaling procedures (Foy et al., 2007, 2009). They provide carefully constructed items that measure reading (PIRLS) and mathematics competencies (TIMSS). MC items are used in both studies not only for knowledge reproduction but, like CR items, also for application and problem solving (Mullis et al., 2007, 2008). The CR coders were trained resulting in high inter-rater reliabilities (e.g., 93 % of agreement on average across countries in PIRLS 2006; Mullis et al., 2007, pp. 301–302). Unlike previously described scales (e.g., Lissitz et al., 2012; Lukhele et al., 1994), both response formats make up a comparable proportion of each assessment, facilitating the comparison of potential response format effects. Finally, both studies comprise a sufficient sample size to allow the detection of small effect sizes and the application of IRT methods.

In large-scale assessments, item construction often aims at unidimensional measurements. Poorly fitting items are often eliminated during pretests. The remaining items of both formats are loading on one factor. This homogeneity does not preclude response format-specific differences in reliability and validity, but any differences are probably rather small. If there were consistent format differences within and across studies, developers of large-scale assessments might be interested in resorting to the more reliable and valid format. Additionally, one should look at specific ability ranges. For example, if MC was less reliable than CR for very competent children (cf. Wan & Henly, 2012), future assessments should be complemented by rather difficult items.

Research Questions

The present paper scrutinizes the assessment of reading (PIRLS) and mathematics (TIMSS) in fourth-graders. MC items and CR items were compared in terms of reliability and validity (using school grades as criterion[1]). Additionally, the relationship between test extensions and format-specific increases in reliability and validity were explored: Is it better to stick to one response format or to add CR items to an MC-only scale (and, vice versa, MC items to a CR-only scale)? A combination of both response formats appears to yield the highest reliabilities, albeit only for higher-than-average proficiency levels (Gültekin & Demirtaşlı, 2012). Therefore, the following research questions were analyzed:

Q1: Do MC-only scales and CR-only scales differ with regard to reliability?

[1] Although grades are supposed to reflect scholastic achievements with regard to a specific curriculum, which varies across and within countries, and competence scales in large-scale assessment are supposed to assess the corresponding competence (e.g., reading literacy; not directly related to a specific curriculum), there is considerable conceptual and empirical overlap to expect substantial validity coefficients (cf. Phan, Sentovich, Kromrey, Dedrick, & Ferron, 2010).

Q2: Is there a differential increase in reliability when a MC-only scale is extended by either more MC items or by CR items? Is there a differential increase in reliability when a CR-only scale is extended by either more CR items or by MC items?

Q3: Do MC-only scales and CR-only scales show differential validity regarding school grades?

Q4: Is there a differential increase in validity when an MC-only scale is extended by either more MC items or by CR items? Is there a differential increase in validity when a CR-only scale is extended by either more CR items or by MC items?

To explore whether format-specific differences were moderated by the level of proficiency, we used an IRT framework.

Study 1

PIRLS focused on the assessment of reading in 4th-graders. The assessment of reading proficiency consisted of reading passages and corresponding questions (Mullis et al., 2007).

Method

Sample.

The PIRLS reading items had been calibrated using responses from 119,413 children from 26 countries (Foy et al., 2007). For the comparison of format-specific reliabilities, we used the official item parameter estimates[2]. The analyses of differential criterion validity were based on the German subsample (*n* = 7,899), because the data set contains German course grades as indicators of academic achievement. These grades are regarded as homogeneous in terms of grading format, language, and educational system. Forty-two children without responses to any items pertaining to one reading passage were excluded. Individuals without grade points were also dropped from the analyses (*n* = 286; ten cases met both exclusion criteria).

[2] http://timss.bc.edu/pirls2011/downloads/P11_ItemParameters.zip

Instruments and scale composition.

The reading test in PIRLS 2006 consisted of 10 passages (literary or informational texts), each of which contained a text passage to be read, questions related to the text passage, and the answers, which differ in both item types: In CR items, the answers had to be written down by the examinee in his or her own words, whereas in MC items, the examinees had to choose one answer out of four options accompanying each question. The items are supposed to cover four different reading processes (Mullis et al., 2007, pp. 55–62, p. 285): focus on and retrieve explicitly stated information and ideas (19 MC items, 12 CR items); make straightforward inferences (29 MC, 14 CR); interpret and integrate ideas and information (6 MC, 28 CR); and examine and evaluate content, language, and textual elements (10 MC, 8 CR). One item was excluded from the scaling procedure by the PIRLS team (R021S08M), leaving 125 items for our analyses.

We divided the MC and the CR items of the PIRLS booklets into two respective halves and reassembled them into a set of new scales in order to compare response formats: (1) an MC&MC scale containing both MC halves, (2) a CR&CR scale containing all CR items, (3) an MC&CR scale consisting of half the MC items with half the CR items, and (4) a CR&MC scale containing the other half of the MC items with the other half of the CR items.

For the analyses of reliabilities, the MC items of the five texts C, A, F, Y, and K (odd/even split of the 10 text passages) constituted the first MC half. The MC items of the remaining five texts U, S, E, L, and N formed the second MC half. The CR halves were built on the same two sets of texts. Thus, the new MC&CR scale consisted of the first MC half (MC items of texts C, A, F, Y, and K) and the second CR half (CR items of texts U, S, E, L, and N).

For the analyses of validities, there were by design only responses to one booklet (i.e., two texts; Mullis et al., 2007, p. 286) from each participant. The MC items pertaining to the first text passage (i.e., text C for booklet 1, text F for booklet 2, etc.) were used as the first MC half. The MC items pertaining to the second passage (i.e., text F for booklet 1, text Y for booklet 2, etc.) were used as the second MC half. So for the children who worked on booklet 1, MC&MC was the extension of the MC items of text C with MC items from text F whereas MC&CR was the extension of the MC items of text C with CR items from text F. Table 1 shows the item frequencies for the new scales.

---Table 1---

Data analyses.

The official parameter estimates for PIRLS 2006 were based on a unidimensional IRT model. Three-parameter logistic models (3-PL) were estimated for the MC items. Dichotomous CR items were scaled with two-parameter logistic models (2-PL). Generalized partial credit models (GPCM) were used for the CR items with more than two possible score levels (cf. Foy et al., 2007).

For the analyses of reliability, we followed the procedure of Lukhele and colleagues (1994) as well as Gültekin and Demirtaşlı (2012; see also Embretson & Reise, 2000, pp. 183–186) by plotting test information curves, I(θ), and by calculating the area under the curve (AUC). Three comparisons were made: (1) MC&MC vs. CR&CR, (2) MC&MC vs. MC&CR, and (3) CR&CR vs. CR&MC; the first comparison relates to Q1 (overall format effects) whereas the other two comparisons relate to Q2 (test extension). The relative efficiency RE = I(θ,first scale)/I(θ,second scale) was calculated for proficiency levels of $-3 \le \theta \le 3$ (> 99.5% of children fall within this range) and for $-2 \leq \theta \leq 2$ (> 95% of the children). A slightly adapted version of PlotIRT (Hill & Langer, 2005) provided the test information curves and corresponding AUCs.

In IRT analyses, reliability is partly determined by the location of the item difficulty parameters *b*. The *b* values are higher for easier items than for more difficult ones (Foy et al., 2007). Mean differences between the location of *b* of MC items and CR items were tested using *t*-tests. We tested the equality of standard deviations using variance ratio tests. All significance tests used $\alpha = .05$.

For the analyses of validity, we used EAP(θ) ability estimates ("expected a posteriori"; Thissen & Orlando, 2001, p. 112) that were calculated for each of the four new scales for each child. The few skipped or uncompleted items were regarded as not solved ($\bar{k} = 1.3$, $SD_{\bar{k}} = 2.4$). In PIRLS, 1/4 of these uncompleted items were MC items. Not reached items constituted about 15% of item nonresponse. Self-reported grades in German (cf. Schneider & Sparfeldt, 2016) served as criterion; please note that in Germany, 1 is the best grade and 6 the worst. As with the analyses of reliability, three pairs of criterion-related validity coefficients were compared: (1) MC&MC vs. CR&CR, (2) MC&MC vs. MC&CR, and (3) CR&CR vs. CR&MC. The first comparison relates to Q3 (comparable criterion validity) whereas the other two comparisons relate to Q4 (differential validity of test extensions). The correlations in each pair were compared with the *t*-test for non-independent correlations because they were both derived from the same sample and contained the same criterion. Validity coefficients were also calculated and compared within two subgroups of children, the lower quartile (Q25) and the upper quartile (Q75), in order to explore whether validity differences depended on the ability level. The first set of plausible values was used to identify the groups.

Results

Reliability.

The test information curves for the four new scales are shown in Figure 1. As expected and intended, the reading assessment in PIRLS was most reliable for an average ability level ($\theta \approx 0$). The overall precision of assessment (for $-3 \leq \theta \leq 3$) was similar for MC and CR (RE ≈ 1, see Table 2). MC-only showed superior reliability in the range θ ± 1 *SD* whereas CR-only worked better than MC for very high and very low ability levels ($|\theta| \geq 2$). This is illustrated in Figure 2 (RE > 1; for $-2 \leq \theta \leq 2$, see also Table 2). With regard to reliability for the two test extensions, MC was superior to CR around the mean of θ. Adding CR items rather than MC items lead to higher test reliabilities at both ends of the proficiency scale, but not in the middle.

---Figure 1, Table 2, Figure 2---

The mean difference between the item difficulty parameters was small. MC items were on average easier ($d = 0.33$, $t(163) = 2.05$, $p = .042$)[3]. The variance of item difficulties was larger for CR items (Var(CR)/Var(MC) = 1.61, $p = .044$). This is reflected in the more narrow peaks of the MC test information curves in Figure 1.

Validity.

The descriptive statistics of the German subsample are presented in Table 3. The correlation between the MC-only scale and the CR-only scale was large ($r > .60$), but not perfect. The validity coefficients were large, as well ($|r| \approx .50$; see Table 3). Additionally, there was no significant difference between the validity of MC&MC and CR&CR ($\Delta r = 0.01$, $p = .09$; see Table 4). Extending the first MC half with CR items (rather than MC items) lead to a significant albeit negligible increase of validity ($\Delta r = 0.01$, $p = .030$). This effect was more pronounced in

[3] The number of difficulty parameters exceeds the number of items because GPCM items contribute multiple thresholds.

the upper quartile of the sample (Δr = 0.06, p = .012), but not in the lower quartile. There were no differential validity coefficients for the extension of the first CR half.

--- Tables 3, 4---

Summary of Findings

Overall, there were no substantial differences between the reliability of MC-only scales and CR-only scales (Q1). MC was superior to CR around the mean of θ whereas CR was more reliable at the extrema. The MC-only scale became more reliable when CR items were added rather than MC items (Q2). The reliability of the CR-only scale benefitted more from additional MC items than from additional CR items. Likewise, an MC extension is preferable over a CR extension for minimizing the standard measurement error for individuals within 2 *SD* around $\bar{\theta}$.

The effects of format-specific differential validity were very small (Q3). These validity-related findings mirror the reliability results. This differential validity should be regarded as a lower bound estimate, because the PIRLS-procedure to develop unidimensional scales attenuates response format-specific effects. Additionally, the MC-only scale became slightly more valid when CR items were added rather than MC items, especially for good readers (Q4).

Study 2

The aim of the second study was to replicate the findings presented above in a different content domain, specifically mathematics. Differential reliability and differential validity might vary across different fields (Bennett, 1993; Bridgeman & Lewis, 1994). Unlike in PIRLS, the TIMSS mathematics items do not pertain to one or two main stimuli (i.e., text passages in PIRLS). For example, in TIMSS 2007 (4th grade), most items stand alone (151 out of 179). This could be relevant, because the questions and the set of choices in reading items seem to provide contextual information that sometimes facilitates finding the correct option (e.g., Sparfeldt, Kimmel, Löwenkamp, Steingräber, & Rost, 2012).

About half of the TIMSS-items (96 out of 179) feature an MC response format with one correct answer and three distractor responses. The remaining items display various CR formats, for example, completing a graph or writing a short answer (Mullis et al., 2008). In previous analyses of TIMSS (1995) mathematics items MC items were on average slightly easier than CR items ($\bar{d} = 0.02$, $SD(d) = 0.12$; $d = 0.06$ in the German subsample; Hastedt & Sibberns, 2005). As with PIRLS, there has not been a comparison of response formats with regard to criterion validity so far. In general, previous dimensional analyses suggest at most small format-specific differences in reading comprehension and quantitative tasks (cf. Traub, 1993).

Method

Sample.

We used data from TIMSS 2007. The calibration of the TIMSS mathematics items was based on 160,922 4th-graders from 36 countries (Foy et al., 2009). The official item parameter estimates[4] were used to analyze format-specific reliabilities. The analyses of differential validity was based on the German subsample ($n = 5{,}200$). This data set contained mathematics grades as indicators of mathematical achievement. Six children had no responses to (at least) one whole block of items. Students with missing mathematics grades were also excluded from the analyses ($n = 85$).

Instruments and scale composition.

The assessment of mathematics performance in TIMSS 2007 covered three cognitive domains (Mullis et al., 2008, p. 374): knowing (45 MC items, 24 CR items), applying (37 MC,

[4] http://timssandpirls.bc.edu/timss2011/downloads/T11_ItemParameters.zip

33 CR), and reasoning (14 MC, 26 CR). Two items were excluded from the TIMSS scaling procedure (M031223, M031002), leaving 177 items for our analyses.

We divided the MC items and the CR items of the TIMSS booklets into two respective halves and reassembled them into four new scales analogous to those described in Study 1: MC&MC, CR&CR, MC&CR, and CR&MC. Each of the 14 TIMSS booklets was divided into two blocks. The first block contained the items that overlap with the previous booklet. The second block contained the items that overlap with the following booklet. There were 14 blocks in total. For the analyses of reliabilities, the MC items of odd blocks made up the first MC half, the MC items of even blocks constituted the second half. Again, the CR halves were built in the same manner. Subsequently, the four new scales mentioned above were assembled.

For the analyses of validities, there were by design only two blocks per respondent/booklet. Each block appeared in two booklets, once as the first block and once as the second block. This overlap allowed us to use each block as the first half for the group working on the first of these two booklets, and as the second half for the group working on the other booklet. So for example, the MC&CR scale for the children who worked on booklet 1 contained the same CR items as the CR&MC scale for the children who worked on booklet 2. Table 1 shows the item frequencies for these scales.

Data analyses.

Self-reported mathematics grades (cf. Schneider & Sparfeldt, 2016) served as criterion in the analyses of validities. The official parameter estimates for the mathematics assessment in TIMSS 2007 were based on a unidimensional IRT model (3-PL for the MC items, 2-PL for dichotomous CR items, and GPCM for the CR items with more than two possible score levels; see Foy et al., 2009 for details). Statistical analyses of reliabilities and validities were run analog to Study 1. Again, skipped or uncompleted items were regarded as not solved ($\bar{k}$ = 2.2, $SD_{\bar{k}}$ = 3.1). In TIMSS, 1/3 of these few missing items were MC items. Uncompleted items constituted about 15% of item nonresponse.

Results

Reliability.

The mathematics assessment in TIMSS was most reliable around the mean of θ (see Figure 1). The MC-only scale was more reliable than the CR-only scale for $0.1 < \theta < 1.8$. The relative efficiency favored CR, even for the restricted range $–2 \leq \theta \leq 2$ (RE < 1, see Table 2). Regarding test extensions, MC was for the most part preferable.

There was no significant mean difference between the item difficulty parameters in TIMSS (d = –0.07, t(186) = –0.47, p = .64), but the variance of item difficulties was larger for CR items (Var(CR)/Var(MC) = 1.69, p = .012).

Validity.

The descriptive statistics of the German subsample are presented in Table 3. The correlation between MC-only and CR-only was large (r > .60). The validity coefficients were also large ($|r| \approx .50$; see Table 3). The MC-only scale had a slightly larger validity coefficient than the CR-only scale (Δr = –0.02, p = .017). The effect was slightly larger, but not significant for the lower quartile and the upper quartile, respectively. There was no differential validity for format-specific test extensions (see Table 4).

Summary of Findings

In TIMSS 2007, CR yielded a more reliable assessment of the mathematics performance than MC (Q1). MC was slightly superior to CR around the mean of θ. CR was more reliable for low and (very) high proficiency levels (possibly due to MC-related ceiling effects; cf. Hastedt &

Sibberns, 2005). In TIMSS, the response format of the test extension did not moderate the scale's overall reliability substantially (Q2).

The effects of format-specific differential validity were very small, although there was a significant difference in favor of MC (Q3). The response format of the test extension did not moderate the scale's criterion validity (Q4). Therefore, the reliability differences in favor of CR were not reflected in superior validity coefficients.

General Discussion

Response Format Differences

We studied response format effects in educational assessments by reassembling the items from PIRLS 2006 and TIMSS 2007 into format-specific scales. The comparison of these new scales shows that the reliabilities of select (MC) and supply items (CR) differ slightly for different θ-segments: MC offers more precision around the mean whereas CR is more reliable at the extrema. The advantages of MC appear more pronounced in the PIRLS results. The pattern of proficiency-specific reliabilities in both studies suggests a lack of very easy and very difficult MC items with sufficient discriminatory power. Improved distractors might make very difficult MC items more reliable by attenuating the guessing parameters and increasing the discrimination parameters. Unfortunately, very easy and very difficult MC items were scarce in both studies, although such items can be created (Waugh & Gronlund, 2013).

The differential validity effects are small ($|\Delta r| \leq 0.02$ for the comparisons of MC-only and CR-only) and in line with previous findings from Advanced Placement tests, which showed similar response format effects when the criterion corresponded closely to the assessment instrument (Bridgeman & Lewis, 1994). Given the lack of differential validities, other format-specific strengths could be considered. Wainer and Thissen (1993, p. 111) suggested that scoring one CR item costs substantially more than scoring an MC item (or MC items) providing comparable test information. Scoring MC items is also less error-prone than coding constructed responses. Despite advances in the automated scoring of constructed responses, MC items remain ideally suited for computer-based adaptive testing (CAT). Reading competences and math competences are still often assessed with paper-pencil-test formats, at least in PIRLS and TIMSS. Hopefully, the important improvements in computer-based assessments (e.g., regarding scoring, adaptive testing) will find their way into large-scale assessments, whenever useful and adequate. By relying on MC items, the duration of the assessment can be shortened – or additional items can be included to increase the reliability and to reach a broader coverage (Wainer & Thissen, 1993). Unfortunately, the present data do not provide information regarding the time spent on each item. Previous research suggests that the reliability per minute spent on a test is larger for MC items than for CR items (Lukhele et al., 1994; cf. Wan & Henly, 2012, Table 4).

Response Format Similarities

As mentioned, there were no substantial reliability differences between MC and CR beyond the moderation of relative efficiency by ability. This lack of substantial differential psychometric properties is in line with the objective of the scaling process in both studies, which aimed at a unidimensional assessment. The slightly lower reliability of MC items (compared to CR items) in TIMSS 2007 could be used to argue against the use of MC items. Still, there is no corresponding attenuation of validity. Neither did the slightly larger variances of CR item difficulty parameters result in improved validities in the lower and upper quartiles in either of the two studies. It appears that the constructs which are relevant for educational achievement are still sufficiently reflected in the IRT scores of the MC scales. Neither the inherent limitations of MC items (e.g., MC is not suited for the assessment and discussion of opinions; cf. Haladyna & Rodriguez, 2013), nor possible distractor-related biases (Sparfeldt et al., 2012) seem to affect the corresponding aspects of a valid competence assessment.

Using a broad set of items that cover two main competence domains, we found converging results for reading and mathematics. They can be interpreted as a compromise of the conflicting results of earlier reliability studies (e.g., Gültekin & Demirtaşlı, 2012; Lukhele et al., 1994). MC and CR tend to show very similar psychometric properties; sometimes (e.g., identification of gifted students) a combination of formats might work best (cf. Gültekin & Demirtaşlı, 2012).

The present studies benefit from the matrix-sampling approach of PIRLS and TIMSS, which made it possible to assemble new scales with mostly balanced item frequencies. Consequently, the validity results are based on a set of different blocks and do not depend on a singular booklet. In both studies, MC items and CR items were presented side by side. It could be argued that the scale composition affects student behavior. Students might choose to spend more time on MC items than on CR items. The small item nonresponse differences are in line with this hypothesis. The small proportion of missing data, however, suggests that such an effect would be marginal. The test taking behavior might be different in assessments that feature only one response format. Additionally, expectations regarding the response format can influence the learning process (e.g., using more surface learning when preparing for an MC-only test; cf. Martinez, 1999). Another potential issue regarding the joint administration of MC and CR is test motivation. Reduced motivation is associated with lower scores in low-stakes tests; this might lead to inflated validity coefficients when the criterion in question contains motivational aspects (Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011). Further research, preferably not only on the student level but also on the item level, might determine whether this effect is moderated by response format differences.

Limitations

Local dependence can occur in reading assessments that administer several questions related to the same text (Hartig & Höhler, 2009). But, the previous analysis of booklet effects in PIRLS 2006 (Chang & Wang, 2010) and the similarity of PIRLS results and TIMSS results (a conceptual replication in a different domain with different samples) suggest that this is not an important issue in our study. Any violation of the assumption of local independence would presumably affect items regardless of their response format.

There were fewer CR items than MC items in TIMSS. Still, the number of difficulty parameters was similar for both item types, because some CR items were graded from 0–2 or from 0–3. Therefore, we did not adjust the relative efficiencies presented in Figure 2 for the difference in item numbers. Although PIRLS and TIMSS cover important competences, both studies are limited in terms of content (reading and mathematics) and target populations (4th-graders). The findings of Gültekin and Demirtaşlı (2012) suggest that CR-based scales in TIMSS might be more efficient for 8th-graders. Further research with more than just one booklet is needed to corroborate this notion. Future studies might also investigate whether differential validity is more pronounced for criteria other than course grades.

Conclusion

The comparison of CR and MC is not only confined to purely diagnostic aspects; for example, the MC format is often met with reservations due to political rather than psychometric reasons. Despite the social aspects of testing culture that might favor the CR format, economic and diagnostic properties should be considered. Although MC does not always yield the most reliable assessment, it is certainly suitable for large-scale assessments. The converging results of PIRLS and TIMSS highlight MC scales and CR scales both have desirable psychometric properties. A reliable and valid measurement is possible with either response format, although the assessment of high levels of reading proficiency may be more valid when MC items and CR items are combined.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (4th ed.). Washington, DC: AERA.

Arnold, K.-H., Bos, W., Richert, P., & Stubbe, T. C. (2007). Schullaufbahnpräferenzen am Ende der vierten Klassenstufe [School career preferences at the end of fourth grade]. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, … R. Valtin (Eds.): *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich* (pp. 271–297). Münster, Germany: Waxmann.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–27). Hillsdale, NJ: Erlbaum.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28,* 77–92. doi:10.1111/j.1745-3984.1991.tb00345.x

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*, 37–50. doi:10.1111/j.1745-3984.1994.tb00433.x

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage. doi:10.4135/9781412985642

Chan, N., & Kennedy, P. E. (2002). Are multiple-choice exams easier for economics students? A comparison of multiple-choice and “equivalent” constructed-response exam questions. *Southern Economic Journal, 68*, 957–971.

Chang, Y., & Wang, J. (2010, July). *Examining testlet effects on the PIRLS 2006 assessment*. Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *PNAS, 108*, 7716–7720. 10.1073/pnas.1018601108

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Foy, P., Galia, J., & Li, I. (2007). Scaling the PIRLS 2006 reading assessment data. In M. O. Martin, I. V. Mullis & A. M. Kennedy (Eds.), *PIRLS 2006 Technical Report* (pp. 149–172). Boston, MA: IEA.

Foy, P., Galia, J., & Li, I. (2009). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin & I. V. Mullis (Eds.), *TIMSS 2007 Technical Report* (revised edition, pp. 225–280). Boston, MA: TIMSS & PIRLS International Study Center.

Gültekin, S., & Demirtaşlı, N. Ç. (2012). Comparing the test information obtained through multiple-choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online, 11*, 251–263.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*, 57–63. doi:10.1016/j.stueduc.2009.10.002

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*, 145–161.

Hill, C. D., & Langer, M. (2005). *PlotIRT: A collection of R functions to plot curves associated with item response theory*. R functions version 1.03. Retrieved from http://www.unc.edu/~dthissen/dl.html

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732–746. doi:10.1177/0013164410390032

Kastner, M., & Stangl, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia-Social and Behavioral Sciences*, *12*, 263–273. doi:10.1016/j.sbspro.2011.02.035

Krathwohl, D. R. (2002). A revision of Bloom’s taxonomy: An overview. *Theory Into Practice, 41*, 212–218. doi:10.1207/s15430421tip4104_2

Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*, 115–136. doi:10.1080/08957347.2011.554604

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13*(3). Retrieved from http://www.jattjournal.com/index.php/atp/article/view/48366

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234–250. doi:10.1111/j.1745-3984.1994.tb00445.x

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207–218. doi:10.1207/s15326985ep3404_2

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *IEA's Progress in International Reading Literacy Study in primary school in 40 countries*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Phan, H. T. (2008). *Correlates of mathematics achievement in developed and developing countries: An HLM analysis of TIMSS 2003 eighth-grade mathematics scores* (Doctoral dissertation). Retrieved from http://scholarcommons.usf.edu/etd/452

Phan, H., Sentovich, C., Kromrey, J., Dedrick, R., & Ferron, J. (2010, April). *Correlates of mathematics achievement in developed and developing countries: An HLM analysis of TIMSS 2003 eighth-grade mathematics scores*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Rauch, D., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*, 354–379.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40,* 163–184. doi:10.1111/j.1745-3984.2003.tb01102.x

Schneider, R., & Sparfeldt, J.R. (2016). Zur (Un-)Genauigkeit selbstberichteter Zensuren bei Grundschulkindern [The accuracy of self-reported grades in elementary school]. *Psychologie in Erziehung und Unterricht, 63,* 48–59. doi:10.2378/peu2016.art05d

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*, 453–472.

Sparfeldt, J. R., Kimmel, R., Löwenkamp, L., Steingräber, A., & Rost, D. H. (2012). Not read, but nevertheless solved? Three experiments on PIRLS multiple choice reading comprehension test items. *Educational Assessment, 17*, 214–232. doi: 10.1080/10627197.2012.735921

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140), Mahwah, NJ: Erlbaum.

Thissen, D., Wainer, H., & Wang, X.-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*, 113–123. doi:10.1111/j.1745-3984.1994.tb00437.x

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29–44). Hillsdale, NJ: Erlbaum.

Veeravagu, J., Muthusamy, C., Marimuthu, R., & Michael, A. S. (2010). Using Bloom's taxonomy to gauge students' reading comprehension performance. *Canadian Social Science, 6*, 205–212.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103–118. doi:10.1207/s15324818ame0602_1

Wan, L., & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education, 25*, 58–78. doi:10.1080/08957347.2012.635507

Waugh, C. K., & Gronlund, N. E. (2013). *Assessment of student achievement* (10th ed.). Boston, MA: Pearson.

Table 1

Item frequencies.

Study	Analysis	Half	Texts/Blocks	MC	CR-2	CR-3	CR-4	Total
PIRLS	Reliability	1	C, A, F, Y, K	31	14	14	4	63
		2	U, S, E, L, N	32	14	14	2	62
	Validity	1		6.3	2.5	2.8	0.7	12.3
		2		6.4	2.6	3.0	0.5	12.5
TIMSS	Reliability	1	Uneven	39	38	4		81
		2	Even	55	34	7		96
	Validity	1		6.7	5.1	0.8		12.6
		2		6.7	5.1	0.8		12.6

Notes: CR-2 = CR items with dichotomous scoring, CR-3 = CR items with graded responses (0,1,2), CR-4 = CR items with graded responses (0,1,2,3)

Table 2

Relative efficiency of the new scales.

	PIRLS		TIMSS	
Scales	$-3 \leq \theta \leq 3$	$-2 \leq \theta \leq 2$	$-3 \leq \theta \leq 3$	$-2 \leq \theta \leq 2$
MC&MC vs. CR&CR	0.99	1.15	0.85	0.90
MC&MC vs. MC&CR	0.95	1.04	1.00	1.04
CR&MC vs. CR&CR	1.06	1.13	1.05	1.07

Table 3

Descriptive statistics and intercorrelations of the new scales and grades.

Study	Scale	*M* (*SD*)	1.	2.	3.	4.
PIRLS (*n* = 7,581)	1. MC&MC	0.3 (0.7)				
	2. MC&CR	0.1 (0.7)	.80			
	3. CR&MC	0.2 (0.8)	.81	.67		
	4. CR&CR	0.0 (0.8)	.62	.83	.84	
	5. Grade (German)	2.6 (0.9)	–.50	–.51	–.52	–.52
TIMSS (*n* = 5,111)	1. MC&MC	0.1 (0.7)				
	2. MC&CR	0.0 (0.7)	.79			
	3. CR&MC	0.1 (0.7)	.79	.62		
	4. CR&CR	0.0 (0.7)	.61	.79	.80	
	5. Grade (Mathematics)	2.7 (1.0)	–.55	–.54	–.54	–.53

Table 4

Comparison of validity coefficients.

	PIRLS (*n* = 7,581)			TIMSS (*n* = 5,111)		
	Δ*r*	Δ*r* (Q25)	Δ*r* (Q75)	Δ*r*	Δ*r* (Q25)	Δ*r* (Q75)
MC&MC vs. CR&CR	0.01	–0.02	0.06	–0.02*	–0.04	–0.04
MC&MC vs. MC&CR	0.01*	0.00	0.06*	–0.01	–0.03	–0.01
CR&MC vs. CR&CR	0.00	–0.01	0.03	–0.01	0.00	–0.02

Notes: Q25 = lower quartile, Q75 = upper quartile; * *p* < .05

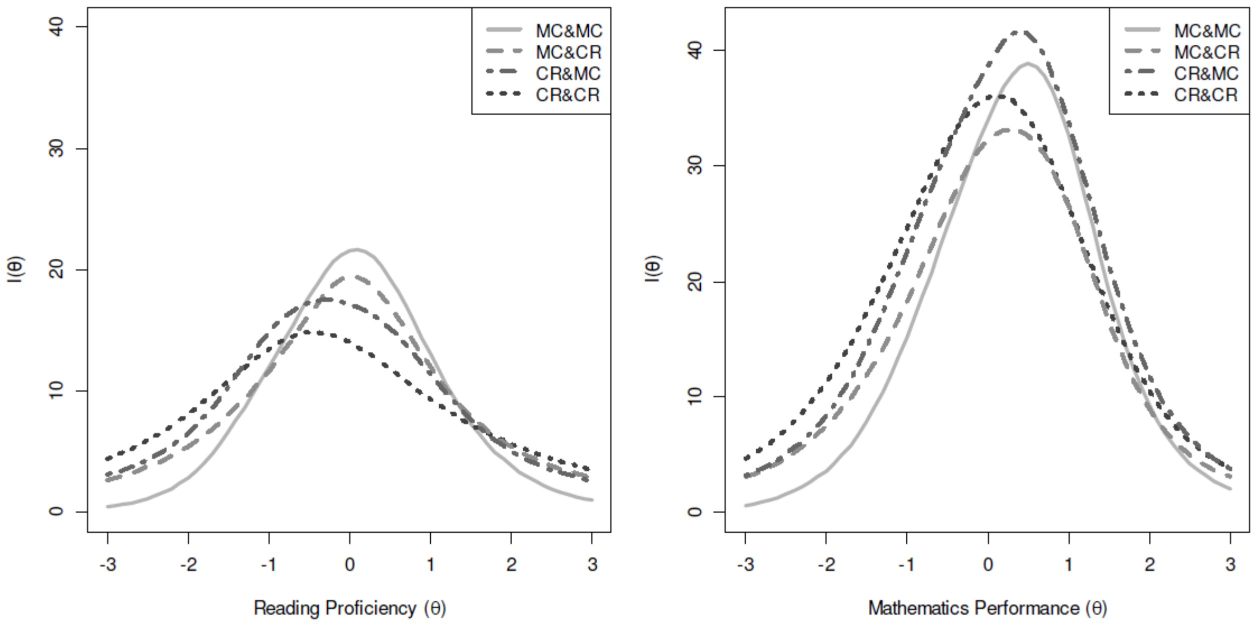

Figure 1. Test information curves of the four new scales for PIRLS (left) and TIMSS (right).

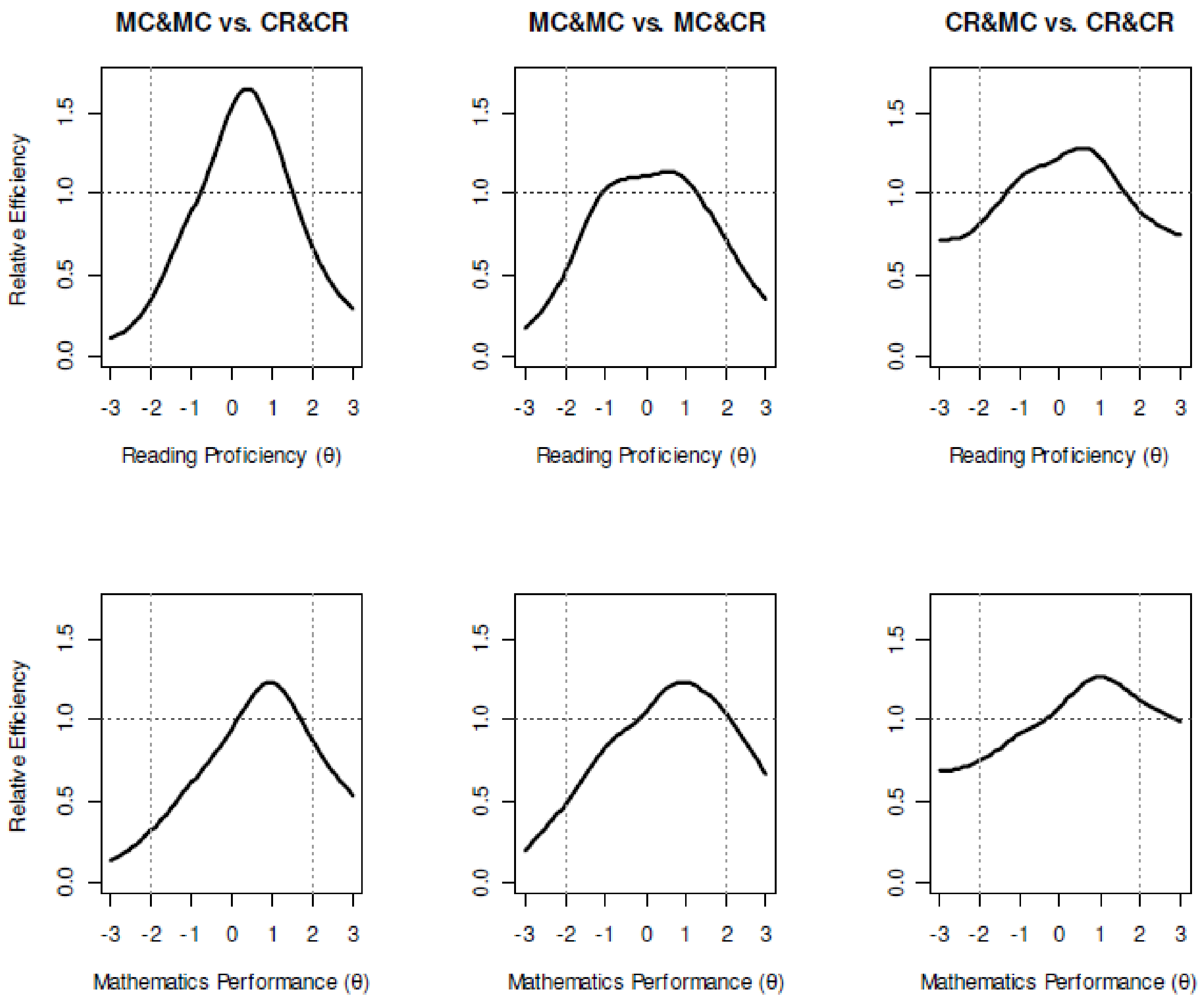

Figure 2. Relative efficiency of format-specific scales for PIRLS (top) and TIMSS (bottom).