

The School Age Gender Gap in Reading Achievement: Examining the Influences of Item Format and Intrinsic Reading Motivation

Franziska Schwabe

Nele McElvany

*Institute for School Development
Research (IFS), TU Dortmund University,
Germany*

Matthias Trendtel

*Federal Institute for Education Research,
Innovation and Development of the
Austrian School System, Salzburg*

ABSTRACT

The importance of reading competence for both individuals and society underlines the strong need to understand the gender gap in reading achievement. Beyond mean differences in reading comprehension, research has indicated that girls possess specific advantages on constructed-response items compared with boys of the same reading ability. Moreover, it was shown that differences in intrinsic motivation in the tested domain might affect test scores. Differential item functioning was used to analyze the complex relationships among gender, item format, and intrinsic reading motivation in two samples taken from large-scale German assessments of reading comprehension (PIRLS 2011 and PISA 2009). In line with prior research, results showed that compared with equally skilled boys, both 10- and 15-year-old girls performed better on constructed-response items. Furthermore, findings suggest an advantage of 15-year-old but not 10-year-old students with high levels of intrinsic reading motivation when responding to constructed-response items. Results are discussed in relation to the design and interpretation of (large-scale) assessments, the increasing use of constructed-response items in new assessments in response to the Common Core State Standards, and gender-sensitive educational practice.

School achievement studies consistently demonstrate higher levels of reading achievement and intrinsic reading motivation in female students than males. Because reading is a fundamental prerequisite for academic success and participation in society (Connor et al., 2011; Pfof, Dörfler, & Artelt, 2012; Snow, Burns, & Griffin, 1998), the lower performance of boys is a crucial issue in educational research, administration, and practice. Beyond the girls' advantage in reading in general, research suggests that girls additionally exhibit specific advantages in regard to several requirements of reading tests (e.g., Lafontaine & Monseur, 2009). Specific advantages can be identified and illustrated by investigating features of test items.

Considering that reading tests are commonly composed of constructed-response (CR) and multiple-choice (MC) items (Haladyna & Rodriguez, 2013), and taking into account the current progression toward increasing use of CR items in newly developed assessments in response to the Common Core State Standards (e.g., as developed by the Partnership for Assessment of Readiness for College and Careers [PARCC] and Smarter Balance in the United States), the item format is a core feature to investigate. The main difference between the two formats is that in contrast to MC items, CR items require not only receptive but also productive language skills (Haladyna & Rodriguez,

2013). Because girls are more likely to possess the productive language skills necessary to cope with CR items, this specific requirement is assumed to impede boys from fully demonstrating their reading competence (Simkin & Kuechler, 2005). Thus, CR reading test items could be disproportionately more difficult for boys.

Another characteristic in distinguishing between the two item formats used in large-scale reading assessments is that CR items require greater effort than MC items (Rodriguez, 2003). One example is that typically responding to a CR item takes more time than responding to a MC item. Thus, assuming that students who exhibit higher levels of intrinsic reading motivation may try harder in testing situations (Guthrie & Wigfield, 2005), highly motivated students might do relatively better on CR items than less motivated students.

Advantages of subgroups of test takers, which exist in addition to and independently from mean differences in ability between the groups, can be detected with analyses of differential item functioning (DIF; Angoff, 1993). *DIF* refers to a psychometric difference in how an item functions across groups. DIF exists when examinees who belong to different groups, such as male versus female or young versus old, vary in their probabilities of responding successfully to an item, despite being equally skilled in the ability that the item is supposed to assess. The focus on relative differences is an advantage compared with, for example, an ANCOVA because it investigates student groups of equal ability but with different profiles of competencies.

To gain insights into gender-related specific advantages in reading, this study investigates the interaction between gender and item format in both 10- and 15-year-old students by conducting DIF analyses based on two German samples of large-scale assessments of reading comprehension. Moreover, the current study also explores interaction effects between intrinsic reading motivation and item format for both age groups. Finally, it investigates possible influences of intrinsic reading motivation on the interaction between gender and item format.

Theoretical Background

Gender Differences in Reading Achievement and Reading Motivation

Reading comprehension is broadly defined as the active extraction and construction of meaning from various text types. Hence, it is very important for educational success (see Connor et al., 2011; Snow, 2002). Reading comprehension is assumed to be influenced by reading behavior, which itself is substantially affected by reading motivation

(Becker, McElvany, & Kortenbruck, 2010; Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). One important dimension of the multifaceted construct of reading motivation is intrinsic reading motivation. This can be defined as “the drive to read for internal purposes, such as deriving pleasure, attaining personal goals, or satisfying curiosity” and is also assumed to be multidimensional (Conradi, Jang, & McKenna, 2014, p. 154; see also Pintrich & Schunk, 2002; Schaffner, Phillip, & Schiefele, 2014; Schiefele, Schaffner, Möller, & Wigfield, 2012). For example, Wigfield and Guthrie (1997) found intrinsic reading motivation to comprise involvement, reading curiosity, and perceived value of reading. Research identified personal and situational key factors that elicit and sustain intrinsic reading motivation. For example, factors found to enhance intrinsic motivation, especially in the context of schooling, are positive performance feedback and perceived competence (Conradi et al., 2014).

Although no gender differences in regard to reading achievement and motivation were found for adult readers (e.g., International Adult Literacy Survey, Programme for the International Assessment of Adult Competencies; Hannon, 2014), findings regarding school-age girls and boys are less conclusive. International large-scale assessment studies—one implemented at the end of grade 4 (Progress in International Reading Literacy Study [PIRLS]; Mullis, Martin, Foy, & Drucker, 2012), one for students at the age of 15 (Programme for International Student Assessment [PISA]; OECD, 2010), and one for the U.S. context (National Assessment of Educational Progress [NAEP]; National Center for Education Statistics, 2011)—have revealed significant reading achievement gaps between boys and girls, demonstrating a slight advantage for younger girls (e.g., PIRLS 2006 [10-year-olds]: Cohen’s $d = 0.25$; NAEP 2011: 0.2 standard deviation [SD] at grade 4) and a medium advantage for older girls (PISA 2006 [15-year-olds]: Cohen’s $d = 0.49$; Lynn & Mikk, 2009; NAEP 2011: 0.3 SD at grade 8; National Center for Education Statistics, 2011).

Nevertheless, in meta-analyses of studies on reading achievement, no gender gap was observed (mean of all age groups [6 to >25]: Cohen’s $d = 0.03$; for 6–10-year-olds: Cohen’s $d = 0.09$; for 11–18-year-olds: Cohen’s $d = 0.02$; Hyde & Linn, 1988). Small to negligible gender differences were also replicated in 16 German studies of students in elementary school (Cohen’s $d = 0.11$; Mücke, 2009) and for U.S. students in grades 6–11, demonstrating no variation across age groups (studies 1970–1991: Cohen’s $d = 0.06$; studies 1992–2002: Cohen’s $d = 0.24$; Lietz, 2006).

More recent studies concerned with specific aspects of reading revealed that gender differences cannot be considered as being equivalent across all requirements of reading tests, but differ in regard to specific aspects of reading competence. Clinton et al. (2012) found differences in

specific cognitive components of reading comprehension in fourth-grade children that were related in a subsequent study to different memory performances (semantically vs. episodically based; Seipel, Clinton, & Carlson, 2012). Moreover, it was found that boys' reading comprehension depends largely on the text content, whereas girls' achievement is hardly influenced by this factor (Oakhill & Petrides, 2007). This finding implies the relevance of taking motivational variables into account when studying the gender gap in reading.

Considering intrinsic reading motivation, there is a considerable gender gap in favor of girls across all ages (United Kingdom: McGeown, Goodwin, Henderson, & Wright, 2012; United States: McKenna, Conradi, Lawrence, Jang, & Meyer, 2012; Mucherah & Yoder, 2008; international: Mullis et al., 2012 [PIRLS]; OECD, 2010 [PISA]). Moreover, in their longitudinal study on the development of reading motivation, Archambault, Eccles, and Vida (2010) made an important contribution to the understanding of attitudes toward reading (meta-analyses: Petscher, 2010; Schaffner et al., 2014), demonstrating that boys showed lower levels of ability self-concept and subjective value in reading both across all ages—from the beginning of elementary to the end of secondary school—and across different forms of development of reading motivation. Moreover, reading motivation in general decreased from grade 1 to grade 12. For Germany, Becker et al. (2010) demonstrated that girls and boys did not differ in the extent to which reading motivation decreased between grades 3 and 6. Nevertheless, the study revealed a gender gap in regard to the level of reading motivation at all grades investigated. Comparable patterns for reading attitudes in various age groups have been found in the United States (Logan & Johnston, 2009; McKenna et al., 2012; Swalander & Taube, 2007).

In conclusion, results based on school achievement studies and meta-analyses indicate a slight advantage, if any, of girls in reading competence in general but gender-related differences in specific components of reading comprehension, which might be traced back to differences in reading motivation. Studies addressing (intrinsic) reading motivation and its development reveal a significant gender gap in the level of motivation but no increase in this gap across time. Investigating boys' and girls' specific advantages in responding to reading test items—while taking the gender gap in intrinsic reading motivation into account—might provide a deeper understanding of the gender gap in reading achievement. However, most reading tests are composed of items that differ systematically in their format (e.g., NAEP reading test).

Item Format

This aspect, sometimes also referred to as question type, has been subject to controversial discussion ever

since standardized reading tests were developed. The two main formats of reading test items are MC and CR items (Haladyna & Rodriguez, 2013; compare the TIMSS and PIRLS databases for examples of CR items in reading, science, and mathematics). One motivation for the use of MC items is their efficiency in terms of administration and scoring (e.g., Rodriguez, 2003). However, a frequently mentioned disadvantage of this format is its distance to the demands of school teaching. The most common argument favoring CR items is the assumption that they measure some kind of deeper understanding (Bacon, 2003). Despite this advantage, CR items are nonetheless subject to some significant restrictions, which pertain mainly to the scoring of given answers. This has several drawbacks compared with the scoring of MC items. CR scoring tends to be more complex and subjective, thereby reducing the reliability of assessment (Rodriguez, 2003; Wainer & Thissen, 1993) and being more time consuming and more expensive. Moreover, immediately scoring during the assessment process as it is necessary for adaptive testing cannot or can only restrictedly be realized.

Alongside the overall discussion on the appropriateness of applying different item formats in general, one core issue is their interaction with specific characteristics of subgroups of test takers (e.g., Grisay & Monseur, 2007; Routitsky & Turner, 2003). Studies based on large-scale survey data suggest that the item format has a significant effect on the performance of different subgroups. Specific advantages in responding to CR items have been found for students with higher levels of proficiency (Routitsky & Turner, 2003), for students living in specific countries (Grisay & Monseur, 2007), and for girls (Lafontaine & Monseur, 2009).

Concerning the differences between CR and MC items, two key aspects might be responsible for these differential effects on the performance of subgroups of test takers: First, the requirement to write (long) answers may still be an obstacle for some students in elementary and even secondary schools, independent of their reading competence level (Guthrie & Wigfield, 2005), especially for those who lack productive language skills. Second, CR items seem to be more challenging for students because they frequently require the respondents to find an answer in their own words (Solheim, 2011). Furthermore, CR items generally require a greater amount of time (Rodriguez, 2003) and, therefore, may be perceived as more demanding because more persistence is needed to deal with the item in question. Whereas the first aspect, productive language skills, has been discussed and investigated in several studies (Ben-Shakhar & Sinai, 1991; Lafontaine & Monseur, 2009; Rodriguez, 2003; Simkin & Kuechler, 2005; Solheim, 2011; Willingham & Cole, 1997), the

second aspect, intrinsic motivation, has yet to be studied explicitly.

Interaction Between Gender and Item Format

The first abovementioned aspect of CR items, the way they require productive language skills, might differentially influence the performance of boys and girls on reading tests. Some researchers argue that CR items favor test takers with superior writing skills, even if grammar or spelling mistakes from test takers with less writing proficiency do not affect the mastery of the item in question (Zeidner, 1987). Moreover, research has demonstrated that girls usually do better in verbal domains than boys, particularly in productive language skills (e.g., Clinton et al., 2012; Halpern, 2000; Priess & Hyde, 2010; see DESI-Konsortium, 2008, for German grade 8). The reported advantage of girls in productive language skills can be expected to have a significantly greater influence on responses to CR items than on answers to MC items (Simkin & Kuechler, 2005).

There is empirical evidence for the superiority of girls on CR items because they perform better on tests composed of CR items than on those composed of MC items (see Bennett, 1993, for an early meta-analysis). In addition, a specific advantage of female students over males on CR items has been observed frequently for older students in verbal domains such as reading and writing (e.g., 15-year-olds on PISA 2000; Lafontaine & Monseur, 2009; Taylor & Lee, 2012; Zenisky, Hambleton, & Robin, 2004). Yet, findings for younger students point in the opposite direction: Neither the Latvian PIRLS 2006 reading items (Geske & Ozola, 2010) nor the German PIRLS 2001 reading items (Schwippert, Bos, & Lankes, 2004) revealed any interaction between gender and item format in fourth-grade students.

Furthermore, the specific advantage of girls on CR items does not seem to be stable across domains because it does not appear or even reverses in mathematics and science (Mullis, Martin, Fierros, Goldberg, & Stemler, 2000; Routitsky & Turner, 2003). In addition, a study focusing on PISA 2003 mathematics items has shown that high-achieving students do relatively better on CR items than on MC items (Routitsky & Turner, 2003).

The empirical findings and theoretical considerations lead to the conclusion that the interaction between gender and item format depends on aspects of the tested domain, the students' age, and the students' ability in this domain. In concrete terms, girls are more proficient readers and do better on CR reading items, whereas boys reveal higher levels of competence in mathematics and science and exhibit a specific advantage on CR mathematics and science items. This observation suggests that differences in language production

skills cannot solely account for the gender gap on CR reading items, and it supports the assumption of an influence of intrinsic motivation.

Interaction Between Intrinsic Motivation and Item Format

Concerning the second aspect of CR items compared with MC items, the assumption that CR items require higher intrinsic motivation, the relationship between motivation and performance needs to be clarified. First, intrinsic motivation is known to affect learning and, thus, influence achievement and test performance (McKenna et al., 2012; Pintrich & Schunk, 2002; Schiefele et al., 2012). Second, apart from this indirect effect of intrinsic motivation on test performance, there might also be additional direct effects. For example, Bandura (1997) argues that competencies might be inhibited during assessments by self-doubt as a motivational variable. Alongside the potential influence of self-doubt in testing situations, other motivational factors might impact test performance, such as a student's task value beliefs, his or her self-concept and self-efficacy, and the student's goal orientations (see Schiefele et al., 2012, for a detailed description of the constructs mentioned).

Furthermore, high levels of intrinsic motivation in a specific domain correlate with great expectations of success and, therefore, with a domain-specific willingness to make an effort (e.g., Schiefele et al., 2012). As a consequence, test scores of poorly motivated students would not (just) reflect their levels of competence but a lack of motivation caused by test characteristics (Solheim, 2011). Concerning reading assessments, Guthrie and Wigfield (2005) have argued that students with low reading motivation will encounter problems when facing complex tasks, because "if a reading assessment has a high level of complexity, students' sustained effort, avoidance of distractions, and commitment to completing tasks successfully, are likely to contribute to successful performance" (p. 201).

In conclusion, several researchers have proposed a differential effect of motivational variables on responses to test items with different formats in addition to the general influence of motivation on achievement and test performance. More precisely, compared with students who show high levels of intrinsic reading motivation, students who exhibit low levels may perceive creating and writing an answer to an CR item to be more demanding than choosing among alternatives on an MC, independent of their level of reading competence. As a consequence, there might be a substantial interaction between intrinsic reading motivation and item format in reading tests. Thus, girls' high intrinsic reading motivation might explain their specific advantage on CR reading items for both younger and older students. Moreover,

apart from its interaction with item format, motivation might also affect the reported interaction between gender and item format.

Research Questions and Hypotheses

The current study aims to relate the important issue of the gender gap in reading to the possible influences of item format and intrinsic reading motivation by analyzing (a) the interaction between gender and item format (CR vs. MC), (b) the interaction between intrinsic reading motivation and item format in reading tests, and (c) the influence of intrinsic reading motivation on the interaction between gender and item format. As previous studies have revealed different findings for younger versus older and more experienced readers, all research questions will be investigated systematically for both 10-year-old¹ and 15-year-old students. In detail, the study explores the following research questions:

1. Do girls have a specific advantage in responding to CR reading items compared with boys of equal reading ability when they are both 10 and 15 years old?

In line with prior research on different productive language skills and levels of intrinsic reading motivation in boys and girls, we expect a statistically significant interaction between gender and item format, with girls outperforming boys in responding to CR items at the ages of 10 (hypothesis 1.1) and 15 years (hypothesis 1.2), while controlling for mean differences in reading achievement.

2. Do students with high levels of intrinsic reading motivation have a specific advantage on CR reading items compared with less motivated students at the ages of 10 and 15 years?

As Guthrie and Wigfield (2005) and, even earlier, Bandura (1997) have suggested, motivational factors impact performance when dealing with items in different formats across age groups. Therefore, we expect a significant interaction between intrinsic reading motivation and performance in various item formats. Concerning the direction of the effect, we assume that highly motivated test takers will outperform less motivated test takers on CR items to a higher degree compared with their performance on the whole item set because the CR items presumably require greater effort. These theoretical assumptions and considerations apply for students at the ages of 10 (hypothesis 2.1) and 15 years (hypothesis 2.2).

3. Can the specific advantage of girls, as expected in research question 1, on CR reading items be explained

by intrinsic reading motivation in both 10- and 15-year-old readers?

Taking into account prior research on the interaction between gender and item format as well as considerations and findings on intrinsic reading motivation, we expect the interaction between gender and item format to decrease after controlling for intrinsic reading motivation. Because the gender gap in reading motivation is consistent in both age groups, we expect similar patterns for 10-year-old (hypothesis 3.1) and 15-year-old students (hypothesis 3.2).

Method

Samples

To draw on large samples of 10- and 15-year-old readers who have worked on CR and MC items in reading assessments, analyses were based on data from the 2011 PIRLS and 2009 PISA. A total of 4,000 fourth graders participated in the German PIRLS assessment in 2011. Boys were slightly in the majority (50.5%). The average age at the end of grade 4 was about 10 years for both genders ($M_{Age-F} = 10.33$, $SD_{Age-F} = 0.01$; $M_{Age-M} = 10.42$, $SD_{Age-M} = 0.016$), $t(3,998) = -196.31$, $p < .001$. Considering family background variables, 65.1% of the girls and 65.1% of the boys stated that they spoke the language of the test at home, $\chi^2(1, N = 4,000) = 6e^{-04}$, $p = .98$. Data were collected within the framework of PIRLS, which, in Germany, was conducted by the Institute for School Development Research (IFS), Dortmund.

A total of 4,979 students from the 1993 birth cohort took part in the German PISA 2009 study. Boys were again slightly in the majority (51.1%). The average age in the birth cohort was 15 years for both genders ($M_{Age-F} = 15.82$, $SD_{Age-F} = 0.28$; $M_{Age-M} = 15.84$, $SD_{Age-M} = 0.29$), $t(78) = 1.343$, $p = .18$. Regarding family background, 79.3% of the girls and 75.7% of the boys stated that they spoke the language of the test at home, $\chi^2(1, N = 4,979) = 0.254$, $p = .64$. Data were collected within the framework of PISA, which, in Germany, was conducted in April and May 2009 by a national consortium coordinated by the German Institute for International Educational Research (DIPF), Frankfurt.

Measures

Reading Competence

PIRLS 2011 administered 13 test booklets, each composed of two texts. Both fictional and nonfictional texts were given, following a multimatrix design (Martin, Mullis, & Kennedy, 2007). Four types of comprehension processes were used in the PIRLS assessment, ranging from “focus

on and retrieve explicitly stated information” to “examine and evaluate content, language, and textual elements” (see Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009, pp. 19–29, for details). Throughout the assessment, students were encouraged to demonstrate reading competence by constructing meaning from written texts. A variety of comprehension questions was administered, each dealing with one of the comprehension processes described. Of the 146 questions in total, 74 were administered in a MC format and 72 in a CR format.

For PISA 2009, the major classification of texts was between fictional or literary texts and nonfictional texts (OECD, 2009). These texts were followed by comprehension questions, each designed to assess one of five aspects of reading competence. The aspects ranged from “retrieving information” to “reflecting on and evaluating the form of a text” (see OECD, 2009, pp. 34–43, for details). In PISA 2009, a total of 100 comprehension questions were administered in 13 test booklets using a multimatrix design (OECD, 2012). The test booklets consisted of continuous, noncontinuous, mixed, and multiple texts (OECD, 2012). Forty-seven items were administered in a MC format and 53 in a CR format.

Intrinsic Reading Motivation

Intrinsic reading motivation of students in grade 4 (PIRLS) and at the age of 15 years (PISA) was measured with comparable item sets. In PIRLS, six items were administered. Four out of six were phrased positively (e.g., “I enjoy reading”), and the other two were phrased negatively (“I think reading is boring,” and “I read only if I have to”). In PISA, 11 items assessed intrinsic reading motivation. Six of these were phrased positively (e.g., “Reading is one of my favorite hobbies”), and five were phrased negatively (e.g., “I read only if I have to”). In both studies, students rated their agreement with the items on a 4-point scale ranging from 1 (disagree completely) to 4 (agree completely). The scales had a high reliability (PIRLS: $\alpha = .88$; PISA: $\alpha = .92$). A dummy variable was created that assigned a high motivation to students who reached levels equal or higher than median (0 = low motivation; 1 = high motivation).

Students’ Background

In both studies, this was assessed by student questionnaires. Gender was measured by self-reporting. To assess immigrant background, students were asked to state the language used at home. The scale offered three possible gradings (“always or almost always,” “sometimes,” and “never”). A dummy variable was created that assigned an immigrant background to students who spoke the language of the test either sometimes or never in their family (0 = no immigrant background; 1 = immigrant background).

Analyses

The clustered samples resulting from the survey design of PISA and PIRLS may lead to biased standard errors in significance tests. To avoid such bias, we computed all standard errors using the jackknife procedure—a resampling technique especially useful for variance and bias estimation—suggested by the associated technical report (PIRLS: Martin et al., 2007; PISA: OECD, 2012). We conducted *t*-tests for preliminary analyses of data and interpreted mean differences with Cohen’s *d* (Cohen, 1960).

To investigate the research questions, we conducted analyses of uniform DIF following the generalized linear mixed models (GLMM) approach proposed by De Boeck et al. (2011). We computed the analyses with the program R (R Core Team, 2013) and especially with the package TAM (Kiefer, Robitzsch, & Wu, 2013).

Following De Boeck et al. (2011), the one-parameter logistic (1PL) item response theory model can be formulated for test taker *p* and item *i* as follows:

$$\eta_{pi} = \theta_p + \beta_i,$$

with η_{pi} describing the logit of the probability of a correct response, $\theta_p \sim N(0, \sigma_\theta^2)$ the ability of the test taker, and β_i the easiness of the item. According to De Boeck et al., a DIF model can be formulated as follows:

$$\eta_{pi} = \theta_p + \beta_i + \omega_{focal} Z_{(p,i)focal} + \sum (h=1)^H \omega_{DIFh} W_{(pi)h}$$

with θ_p and β_i as above and ω_{focal} the main effect of the focal group in comparison with the reference group²: with $Z_{(p,i)focal} = 1$ for the focal group and 0 for the reference group; with $W_{(p,i)h}$ as the person-by-item covariate *h*, in such a way that $W_{(p,i)h} = 1$ if both $Z_{(p,i)focal} = 1$ and the considered person is part of the focal group (item subset DIF), and $Z_{(p,i)focal} = 1$ otherwise; and ω_{DIFh} as the corresponding DIF parameter.

Because each DIF model is a generalization of the 1PL model, models can be compared with a likelihood ratio test for imputed data (Asparouhov & Muthén, 2008).

For the first research question, we specified the main effect of gender and an interaction effect (gender [female]: item format [CR]) and referred to the resulting model as model 1 for both data sets. To investigate the second and third research questions, we included intrinsic reading motivation in the models as a main effect and specified interaction effects (motivation [high]: item format [CR]—second question referred to as model 2; gender [female]: item format [CR]—third question referred to as model 3).

Missing Data

Missing achievement data appeared because of (a) the test design (not all items were administered to all students), (b) omission of administered items by individual

students, and/or (c) failure to cope with administered items. Following Martin et al. (2007), we coded missing data of type (a) as missing in all cases and those of type (b) as an incorrect answer in all cases. We coded type (c) missing data as missing for item calibration and as an incorrect answer for the estimation of person ability.

Moreover, missing data appeared in the background information variables. One recommended solution to this problem is multiple imputations (e.g., van Buuren & Groothuis-Oudshoorn, 2011). We performed multiple imputations with the package multivariate imputation by chained equations (van Buuren & Groothuis-Oudshoorn, 2011). We generated and analyzed five complete data sets separately and conflated the results following Rubin (1987).

Results

Descriptives

The well-known gender gap, which has already been reported for the PISA 2009 and the PIRLS 2011 data, was replicated (see Table 1). Girls in Germany outperformed boys in reading achievement at the ages of 10 and 15 years, PIRLS: $t(3,998) = -142.34, p < .001$; PISA: $t(4,973) = -129.24, p < .001$. The differences revealed medium effect sizes (see Table 1). Moreover, girls showed higher levels of intrinsic reading motivation in both age groups, PIRLS: $t(125) = 7.67, p < .001$; PISA: $t(78) = -20.66, p < .001$.

Interaction Between Gender and Item Format

To investigate the assumed specific advantage of girls in responding to CR items, we specified a DIF model following the GLMM framework proposed by De Boeck et al. (2011). The model is referred to as model 1_(gender × CR) in Table 2. We studied item subset DIF; that is, we investigated the effect of all CR items simultaneously. For both the PIRLS and the

PISA data, Table 2 shows the estimated parameters of the effects of gender and of the interaction between gender and item format on reading achievement.

For 10-year-old students (PIRLS), there was no statistically significant effect of gender on reading achievement ($\omega_{\text{female}} = 0.04, p = .05$) but a significantly positive interaction between gender and item format ($\omega = 0.04, p < .05$). Additionally, model 1_(gender × CR) fitted the data better than a 1PL model, including only the main effect of gender and no interaction effect, 1PL: $p < .001$ (see Table 3). This means that 10-year-old girls did significantly better on CR items compared with equally skilled³ boys. The specific advantage of girls is illustrated by the positive sign of the significant interaction effect. The result that (a) the main effect of gender did not remain significant when the interaction effect was included in the model, and (b) the better fit of the complex model 1_(gender × CR) compared with a model without interaction effects implies that the interaction between gender and item format explains, at least in part, how the gender-associated differences in reading achievement arise: Girls have a specific advantage in mastering CR items.

As far as students at the age of 15 (PISA) are concerned, the estimated coefficient of the main effect of gender differed significantly from zero, $\omega_{\text{female}} = 0.20, p < .05$ (see model 1_(gender × CR) in Table 2). The interaction between gender and item format was also significant, $\omega = 0.06, p < .05$. Model 1_(gender × CR) fitted the data better than a 1PL model, including only the main effect of being female, $p < .001$ (see Table 3). As in the 10-year-olds, 15-year-old female students showed a specific advantage in CR reading items compared with boys of the same age and of the same level of reading competence. Nonetheless, the overall advantage of 15-year-old girls in reading achievement did not vanish in the model after controlling for their advantage in CR items. This indicates that the gender-associated differences in reading achievement of 15-year-olds cannot be traced back solely to the interaction between gender and item format as in the 10-year-olds. In conclusion, however, hypotheses 1.1 and 1.2 are supported by the data: Girls have

TABLE 1
Comparison of Means of Reading Competence and Reading Motivation

Study	M_{Age}	Gender	N	Reading competence			Reading motivation		
				M_{θ}^{***}	SD_{θ}	Cohen's d	M_{θ}^{***}	SD_{θ}	Cohen's d
PIRLS 2011	10.38	Female	1,979	-0.08	0.04	0.59	3.41	0.02	1.50
		Male	2,021	-0.04	0.04		3.18	0.03	
PISA 2009	15.83	Female	2,434	-0.08	0.04	0.63	2.66	0.03	1.49
		Male	2,545	-0.05	0.04		2.04	0.03	

Note. M = mean; SD = standard deviation. All parameters were obtained by using TAM test analysis modules^a to compute a Rasch model for dichotomous data.

^aKiefer, T., Robitzsch, A., & Wu, M. (2013). TAM: Test analysis modules (R package version 0.7-35). Retrieved from CRAN.R-project.org/package=TAM

***Significant difference ($p < .001$) using jackknife.

TABLE 2

Predicting Reading Achievement: Effects of Item Format and Gender (Model 1) and of Item Format, Gender, and Intrinsic Reading Motivation (Models 2 and 3)

Effects	PIRLS 2011 (10-year-olds): Par. (SE)			PISA 2009 (15-year-olds): Par. (SE)		
	Model 1 _(gender × CR)	Model 2 _(mot × CR)	Model 3 _(gender × CR; mot)	Model 1 _(gender × CR)	Model 2 _(mot × CR)	Model 3 _(gender × CR; mot)
<i>Main effects</i>						
Gender (female)	-0.04 (0.02)	0.01 (0.02)	-0.01 (0.02)	0.20* (0.02)	0.06* (0.03)	0.03 (0.02)
Motivation (high)		0.28* (0.03)	-0.30* (0.03)		0.47* (0.02)	0.51* (0.04)
<i>Interaction effects</i>						
Gender (female) × Item format (CR)	-0.04* (0.02)		-0.04* (0.02)	0.06* (0.02)		0.06* (0.02)
Motivation (high) × Item format (CR)		-0.03 (0.02)			0.07* (0.02)	

Note. CR = constructed response; mot = motivation; Par. = estimated parameter; SE = estimated standard error of the respective parameter. * $p < .05$.

TABLE 3
Model Comparison

Study	Model	Number of parameters	T_{imp}^a	df^a	p^a
PIRLS 2011	1PL	101			
	Model 1 _(gender × CR)	102	6.94	1	<.001
	1PL mot	102			
	Model 2 _(mot × CR)	103	2.67	1	.10
	Model 3 _(gender × CR; mot)	103	6.88	1	<.001
PISA 2009	1PL	101			
	Model 1 _(gender × CR)	102	19.99	1	<.001
	1PL mot	102			
	Model 2 _(mot × CR)	103	20.76	1	<.001
	Model 3 _(gender × CR; mot)	103	20.02	1	<.001

Note. 1PL = model including main effect of gender; 1PL mot = model including main effects of gender and intrinsic reading motivation; CR = constructed response; imp = imputation; mot = motivation.

^aModel comparison based on T_{imp} following Asparouhov and Muthén.^b

^bAsparouhov, T., & Muthén, B. (2008). *Chi-square statistics with multiple imputation: Technical appendix*. Los Angeles, CA: Muthén & Muthén.

an advantage in CR reading items at both 10 and 15 years. Girls performed significantly better in CR items compared with boys of the same level of reading competence, as measured by MC and CR items together.

Interaction Between Intrinsic Reading Motivation and Item Format

To investigate the interaction between intrinsic reading motivation and item format, we specified a second DIF model following the GLMM framework proposed by De Boeck et al. (2011): model 2_(mot × CR). Once again, we studied

item subset DIF. Table 2 shows the estimated parameters of the models for both age groups. Model 2_(mot × CR) includes both the main effect of gender and intrinsic reading motivation and an interaction effect between intrinsic reading motivation and item format on reading achievement. For 10-year-old students (PIRLS), neither the estimated coefficient of the main effect of gender ($\omega_{female} = 0.01, p = .74$) nor the interaction between intrinsic reading motivation and item format differed significantly from zero ($\omega = 0.03, p = .13$). Only the main effect of motivation differed significantly from zero, $\omega_{mot} = 0.28, p < .05$. Moreover, model 2_(mot × CR) showed a worse fit than a 1PL model, including

only main effects of gender and intrinsic reading motivation without interactions, 1PL mot: $p = .10$ (see Table 3). This pattern implies that there is a positive effect of motivation on reading achievement, but highly motivated students did not perform better on CR items compared with equally skilled but less motivated students at the age of 10.

Considering 15-year-old readers (PISA), the result differed from the PIRLS findings. The estimated coefficient of both main effects—gender: $\omega_{\text{female}} = 0.06$; and intrinsic reading motivation: $\omega_{\text{mot}} = 0.47$ —differed significantly from zero ($p < .05$), implying that both factors impact significantly on reading achievement for students at the age of 15. Contrary to a lack of any interaction between intrinsic motivation and item format in 10-year-old students, the interaction between motivation and item format for 15-year-old students had an additional and significant positive effect, $\omega = 0.07$. Model 2_(mot × CR) had a better fit than a 1PL model, including only main effects of gender and intrinsic reading motivation, 1PL mot: $p < .001$ (see Table 3). Highly motivated 15-year-old readers performed better when responding to CR items compared with less motivated readers with the same level of reading competence. Thus, the data supported hypothesis 2.2 because there was a significant interaction between intrinsic reading motivation and item format for students at the age of 15. Nevertheless, hypothesis 2.1 had to be rejected because no specific advantage of highly motivated readers on CR reading items could be observed in students at the age of 10 years.

The Effect of Intrinsic Reading Motivation on the Interaction Between Gender and Item Format

To investigate the effect of intrinsic reading motivation on the interaction between gender and item format, we specified a third DIF model (model 3_(gender × CR; mot)) following the GLMM framework proposed by De Boeck et al. (2011), and once again studied item subset DIF. For both data sets, model 3_(gender × CR; mot) included the main effects of gender and intrinsic reading motivation as well as an interaction effect between gender and item format. Looking at 10-year-old students (PIRLS), the estimated coefficient of the main effect of gender was not significant ($\omega_{\text{female}} = -0.01$, $p = .59$; see Table 2), but there was a significant interaction between gender and item format ($\omega = 0.04$, $p < .001$). Model 3_(gender × CR; mot) showed a better fit than a 1PL model, including only main effects of gender and intrinsic reading motivation without interaction, 1PL mot: $p < .001$ (see Table 3). Even after controlling for intrinsic reading motivation, 10-year-old girls performed better on CR reading items compared with equally skilled 10-year-old boys, and the effect did not decrease.

Results for students at the age of 15 (PISA) did not differ from the PIRLS findings: Whereas the estimated coefficient of gender failed to attain significance ($\omega_{\text{female}} = 0.03$, $p = .39$; see Table 2), the main effect of intrinsic reading motivation differed significantly from zero ($\omega_{\text{mot}} = 0.51$, $p < .05$; see Table 2), and model 3_(gender × CR; mot) showed a better fit than a 1PL model, including only main effects of gender and intrinsic reading motivation, 1PL mot: $p < .001$ (see Table 3). Moreover, there was a significant positive interaction between gender and item format, $\omega = 0.06$, $p < .05$ (see Table 2). As in the 10-year-old students, the specific advantage of females on CR reading items compared with males remained significant in the 15-year-old group, even after controlling for intrinsic reading motivation. Also, the interaction effect did not decrease. Contrary to hypotheses 3.1 and 3.2, the effect of the interaction between item format and gender on reading achievement did not decrease in both age groups after controlling for intrinsic reading motivation. Therefore, hypotheses 3.1 and 3.2 had to be rejected. After controlling for the girls' higher level of intrinsic reading motivation, 10- and 15-year-old girls performed better in responding to CR items compared with equally skilled boys.

Discussion

Summary of Findings

This study used data from two large-scale reading tests to systematically investigate the complex relationships among gender, item format, and intrinsic reading motivation in two core school-age groups. As an extension of previous studies, the findings indicate that there is a specific advantage of girls in responding to CR reading items compared with equally skilled boys, both for younger and older students (ages 10 and 15). This explains, at least in part, the overall differences in reading achievement.

The interaction between intrinsic reading motivation and item format reveals a differentiated pattern of results across age groups: Whereas there was a significant advantage of highly motivated 15-year-old readers in responding to CR items, this was not the case for 10-year-old students. Nevertheless, contrary to our theoretically derived expectations, 10- and 15-year-old girls' better performance in responding to CR items did not decrease after controlling for intrinsic reading motivation. Although girls showed higher levels of intrinsic reading motivation, this did not explain their advantage in responding to CR items.

Discussion

The specific advantage on CR items for female students at the age of 15 corresponds to the findings reported

in several studies (e.g., Lafontaine & Monseur, 2009; Routitsky & Turner, 2003). However, other studies did not corroborate these results, especially in regard to 10-year-old students (Geske & Ozola, 2010; Schwippert et al., 2004). In this context, our findings strengthen the assumption of an advantage of girls on CR reading items across all ages. The discrepancies found in comparison with previous studies might result from the method chosen. Whereas previous analyses investigated DIF on an item level, the current study focused on sets of items. Therefore, the relatively small effect in the group of 10-year-old readers found in our analyses might not be covered by influences of other item characteristics such as content or difficulty.

The result that boys are impeded from fully demonstrating their (reading) competence in responding to text-related CR items is a crucial issue from both an assessment and an educational perspective. First, the validity of the assessment might be limited due to the finding of DIF because DIF can indicate validity problems in the test items (e.g., Kane, 2012). In this context, the construct relevance of the item format needs to be discussed. Second, the range of application of text-related CR tasks is not limited to the special setting of assessments. Such requirements are part of daily lessons and class tests in different subjects. This means that boys might be disadvantaged in these contexts whenever productive language skills are involved and that these disadvantages might accumulate.

There are various plausible explanations for the observed specific advantage of girls on CR items for both age groups. These explanations might provide insights into possible starting points for interventions to close the gap between boys and girls in responding to CR items. A first starting point is the girls' dominance in verbal domains, especially their advantage in the area of language production (Priess & Hyde, 2010; Simkin & Kuechler, 2005); a second is the different test-taking behavior of girls and boys (Ben-Shakhar & Sinai, 1991; Hannon, 2014; Willingham & Cole, 1997). Furthermore, other aspects such as item content can at least partly account for this interaction (Lafontaine & Monseur, 2009). This investigation covers the core aspect of the higher level of intrinsic reading motivation of girls as a potential explanation for the gender gap in reading test items. Intrinsic reading motivation might also serve as another starting point for interventions.

Alongside the specific advantage of girls in CR items—illustrated by significant interaction effects in the DIF framework applied—our study points to considerable main effects of intrinsic reading motivation for both ages and of gender for 15-year-old students on reading achievement. This finding, especially the size of the estimators, underlines the central role played by intrinsic motivation in the context of reading and reading assessment.

Results on the interaction between intrinsic reading motivation and item format differ for 10- and 15-year-old students. Findings suggest that high motivation explains specific advantages in responding to CR reading items only in the group of 15-year-old readers, whereas 10-year-old readers do not profit from higher levels of intrinsic reading motivation in responding to those items. One possible explanation for this result might be that, on average, intrinsic reading motivation depends less on proficiency in elementary compared with secondary school students. The latter students have gained or lost motivation in line with their individual progression in acquiring reading proficiency (Archambault et al., 2010). Another reason for the divergent results for 10- and 15-year-old students might be characteristics of our assessment of intrinsic reading motivation, stemming from self-report (see Schiefele et al., 2012, for a critical discussion on questionnaire-based assessment of motivational variables).

The finding that the interaction between gender and item format does not decrease after controlling for intrinsic motivation in both age groups suggests that other aspects that were not included in this study might influence this interaction more strongly. These might be, for example, the type of text or the required cognitive component of the reading process (see Lafontaine & Monseur, 2009). The comparability of our observations in regard to different age groups is in line with prior research on intrinsic reading motivation because this has suggested that effects of intrinsic reading motivation are consistent across ages (Archambault et al., 2010; McGeown et al., 2012), whereas absolute levels of intrinsic reading motivation decrease with increasing age (Becker et al., 2010).

Limitations

Nevertheless, a number of factors are likely to limit the generalizability of the results reported. First, the investigation of reading achievement is based on a specific context: We studied a Western European country in which, in international comparative studies (e.g., PIRLS, PISA) on average, a good reading competence is revealed at the age of 10, a medium reading competence is found at the age of 15, and gender differences in reading achievement and intrinsic reading motivation exist at all school ages. Moreover, we analyzed a definite set of items administered in these studies (PIRLS and PISA). To expand our knowledge about the gender gap in general, we could consider different countries with other conditions regarding the level of competence and the appearance of a gender gap, as well as more items and item types.

Second, as mentioned earlier, we treated the multi-dimensional construct of intrinsic reading motivation as unidimensional for measurement reasons. It is quite

possible that the results of this study would be more revealing if we had been able to incorporate measures that (a) depict the multidimensionality of intrinsic reading motivation and (b) assess affective predictors of intrinsic reading motivation. Moreover, we focused on long-time motivation as a trait, but effects of state motivation (e.g., test motivation) might be of influence, especially in the assessment context studied.

Third, the categorization based on biological gender could be differentiated further by focusing more on the social and cultural gender role (see McGeown et al., 2012). Disentangling cultural, socioeconomic, and linguistic influences may provide deeper insights into this field.

Fourth, we employed just one method to identify DIF, and it is not known whether the identified DIF effects would vary if some other method were to be used for DIF detection. Finally, the matching criterion of students with the same abilities inherent to the test under investigation may have already been contaminated by DIF items, so it may not be an accurate indicator of the overall proficiency.

Implications for Future Research and Educational Practice

Despite these restrictions, the present results have important implications for both future research and educational practice. Subsequent studies should further investigate (a) the explanatory power of the aspects assumed to be determinants of the specific advantage of girls and (b) the differential effect of intrinsic reading motivation for younger and older readers. Subsequent research should clarify whether one aspect is the key influence behind the advantage of girls on CR items or whether a complex interplay of aspects, such as productive language skills and intrinsic reading motivation, must be considered. Moreover, future studies might include other variables such as test motivation and test-taking behavior (e.g., Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011) or executive functioning such as attention shifting and inhibitory control (e.g., Kieffer, Vukovic, & Berry, 2013) because differential test results might depend on these variables as well.

Another question raised by the current study that could also be addressed in future research is whether there is a threshold stage at which intrinsic reading motivation gains a crucial influence in testing situations. Schaffner et al. (2014) have recently demonstrated that the relationship between reading achievement and intrinsic reading motivation differs with respect to characteristics of the students studied, such as their track affiliation within the tracked German school system in which achievement-based tracks are associated with specific degrees. Therefore, it seems possible for the relationship

between intrinsic reading motivation and specific advantages of subgroups of test takers to be moderated by factors such as the age of the students—as suggested by the present results—or even other aspects, such as the readers' level of achievement, family background, and social status (Becker et al., 2010).

Considering educational practice, the findings of this study have implications for (a) standardized testing, especially in the case of large-scale assessments of reading competence, and (b) educational practice in schools to improve both the productive language skills and intrinsic reading motivation of boys and the intrinsic reading motivation of older students. A high percentage of standardized tests currently combine MC and CR item formats, and this is particularly true in large-scale assessments. Crucial educational decisions are often made on the basis of such test results because the monitoring of education systems from an output perspective has become increasingly important in recent years (von Davier, Gonzalez, Kirsch, & Yamamoto, 2013). The scientific knowledge gained from studying large data sets is made available to educational researchers, to educational administrations, and to principals and teachers in schools. Therefore, it is very important to better understand the nature of the competencies measured by each of the item formats used and the degree to which the formats are relevant to and valid for what is being assessed. Our investigation becomes even more relevant in light of the fact that CR items will play a large role in the new assessments being developed by PARCC and Smarter Balance in response to the Common Core State Standards, because we demonstrated a core restriction of CR items in the domain of reading.

To improve the reading achievement of boys within the educational context, our findings clearly support the notion that instruction should provide boys with the necessary skills required by CR items, particularly focusing on reading motivation and productive language skills to make sure that boys will catch up with their female classmates. The study underlines the relevance of specific support for boys. This should begin at an early age and continue up to higher grades because our results are equivocal across different school ages. Moreover, it seems advisable that the choice of texts in reading lessons should consider the needs of boys and meet their interests, which differ greatly from the interests of girls (e.g., Farris, Werderich, Nelson, & Fuhle, 2009; Senn, 2012). An interesting result in this context is that boys' reading comprehension depends largely on the text content, whereas girls' achievement is hardly influenced by this factor (Oakhill & Petrides, 2007). Taking into account that actual reading achievement is affected by interest-related factors such as text content, meeting the interests of boys by, for example, providing suitable reading materials becomes even more important. The NAEP 2011 finding that fourth-grade boys

prefer informational text over fiction (National Center for Education Statistics, 2011) might serve as an orientation for an appropriate selection of materials.

Conclusion

The present investigation was conducted to extend previous research on the gender gap in reading comprehension by focusing on (a) a core aspect in assessment (the item format) and (b) an important characteristic of the test takers studied (their intrinsic reading motivation). The results reported here suggest a promising way to achieve a fuller explanation of the gender gap in reading achievement, and they have implications for subsequent research and the further development of educational practice and school administration.

NOTES

This article is based on considerations and findings resulting from a project funded by the German Federal Ministry of Education and Research (BMBF).

¹ The PIRLS sample is referred to as 10-year-olds because of its mean age. Nonetheless, we are aware that the students were selected on the basis of grade 4 attendance.

² The estimator of the main effect illustrates the mean differences in ability between the studied groups. All models include this main effect. Therefore, the interaction effects are controlled for mean differences in ability between the groups.

³ *Equally skilled* means that within the DIF framework, students were compared while controlling for mean differences in reading competence measured by the whole item set. In simple terms, subgroups of students, who reached equal reading test scores but differed by gender, were compared.

REFERENCES

- Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Erlbaum.
- Archambault, I., Eccles, J., & Vida, M. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology, 102*(4), 804–816. doi:10.1037/a0021075
- Asparouhov, T., & Muthén, B. (2008). *Chi-square statistics with multiple imputation: Technical appendix*. Los Angeles, CA: Muthén & Muthén.
- Bacon, D.R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education, 25*(1), 31–36. doi:10.1177/0273475302250570
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.
- Becker, M., McElvany, N., & Kortenbruck, M. (2010). Intrinsic and extrinsic reading motivation as predictors of reading comprehension: A longitudinal study. *Journal of Educational Psychology, 102*(4), 773–785. doi:10.1037/a0020084
- 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.
- Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*(1), 23–35. doi:10.1111/j.1745-3984.1991.tb00341.x
- Clinton, V., Seipel, B., van den Broek, P., McMaster, K., Kendeou, P., Carlson, S., & Rapp, D.N. (2012). Gender differences in inference generation by fourth-grade students. *Journal of Research in Reading, 00*, 1–18. doi:10.1111/j.1467-9817.2012.01531.x
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. doi:10.1177/001316446002000104
- Connor, C.M., Morrison, F.J., Fishman, B., Giuliani, S., Luck, M., Underwood, P.S., & Bayraktar, A. (2011). Testing the impact of child characteristics × instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46*(3), 189–221. doi:10.1598/RRQ.46.3.1
- Conradi, K., Jang, B.G., & McKenna, M.C. (2014). Motivation terminology in reading research: A conceptual review. *Educational Psychology Review, 26*(1), 127–164. doi:10.1007/s10648-013-9245-z
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–28.
- DESI-Konsortium. (Ed.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Teaching and competence acquisition in German and English: Results of the DESI study]. Weinheim, Germany: Beltz.
- Duckworth, A.L., Quinn, P.D., Lynam, D.R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences of the United States of America, 108*(19), 7716–7720. doi:10.1073/pnas.1018601108
- Farris, P.J., Werderich, D.E., Nelson, P.A., & Fuhle, C.J. (2009). Male call: Fifth-grade boys' reading preferences. *The Reading Teacher, 63*(3), 180–188. doi:10.1598/RT.63.3.1
- Geske, A., & Ozola, A. (2010, July). *Differential item functioning in the aspect of gender differences in reading literacy*. Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.
- Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69–86. doi:10.1016/j.stueduc.2007.01.006
- Guthrie, J.T., & Wigfield, A. (2005). Roles of motivation and engagement in reading comprehension assessment. In S.G. Paris & S.A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 187–214). Mahwah, NJ: Erlbaum.
- Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- Halpern, D.F. (2000). *Sex differences in cognitive abilities*. Mahwah, NJ: Erlbaum.
- Hannon, B. (2014). Are there gender differences in the cognitive components of adult reading comprehension? *Learning and Individual Differences, 32*, 69–79. doi:10.1016/j.lindif.2014.03.017
- Hyde, J.S., & Linn, M.C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104*(1), 53–69. doi:10.1037/0033-2909.104.1.53
- Kane, M. (2012). All validity is construct validity: Or is it? *Measurement: Interdisciplinary Research and Perspectives, 10*(1/2), 66–70. doi:10.1080/15366367.2012.681977
- Kiefer, T., Robitzsch, A., & Wu, M. (2013). *TAM: Test analysis modules* (R package version 0.7–35). Retrieved from CRAN.R-project.org/package=TAM
- Kieffer, M.J., Vukovic, R.K., & Berry, D. (2013). Roles of attention shifting and inhibitory control in fourth-grade reading comprehension. *Reading Research Quarterly, 48*(4), 333–348. doi:10.1002/rrq.54

- Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal*, 8(1), 69–79. doi:10.2304/eerj.2009.8.1.69
- Lietz, P. (2006). Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: A meta-analytic view. *The International Education Journal*, 7(2), 127–149.
- Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where these differences lie. *Journal of Research in Reading*, 32(2), 199–214. doi:10.1111/j.1467-9817.2008.01389.x
- Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *Trames*, 13(1), 3–13. doi:10.3176/tr.2009.1.01
- Martin, M.O., Mullis, I.V.S., & Kennedy, A.M. (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: International Study Center, Boston College.
- McGeown, S., Goodwin, H., Henderson, N., & Wright, P. (2012). Gender differences in reading motivation: Does sex or gender identity provide a better account? *Journal of Research in Reading*, 35(3), 328–336. doi:10.1111/j.1467-9817.2010.01481.x
- McKenna, M.C., Conradi, K., Lawrence, C., Jang, B.G., & Meyer, J.P. (2012). Reading attitudes of middle school students: Results of a U.S. survey. *Reading Research Quarterly*, 47(3), 283–306. doi:10.1002/RRQ.021
- Mucherah, W., & Yoder, A. (2008). Motivation for reading and middle school students' performance on standardized testing in reading. *Reading Psychology*, 29(3), 214–235.
- Mücke, S. (2009). Schulleistungen von Jungen und Mädchen in der Grundschule: Eine metaanalytische Bilanz [School achievement of boys and girls in elementary school: Results from a meta-analysis]. *Empirische Pädagogik*, 23(3), 290–337.
- Mullis, I.V.S., Martin, M.O., Fierros, E.G., Goldberg, A.L., & Stemler, S.E. (2000). *Gender differences in achievement: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.
- Mullis, I.V.S., Martin, M.O., Foy, P., & Drucker, K.T. (2012). *The PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Mullis, I.V.S., Martin, M.O., Kennedy, A.M., Trong, K.L., & Sainsbury, M. (2009). *The PIRLS 2011 assessment framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- National Center for Education Statistics. (2011). *The Nation's Report Card: Reading 2011* (NCES 2012-457). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Oakhill, J.V., & Petrides, A. (2007). Sex differences in the effects of interest on boys' and girls' reading comprehension. *British Journal of Psychology*, 98(2), 223–235. doi:10.1348/000712606X117649
- OECD. (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Paris, France: Author. doi:10.1787/9789264062658-en
- OECD. (2010). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science* (Vol. 1). Paris, France: Author. doi:10.1787/9789264091450-en
- OECD. (2012). *PISA 2009 technical report*. Paris, France: Author. doi:10.1787/9789264167872-en
- Petscher, Y. (2010). A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. *Journal of Research in Reading*, 33(4), 335–355. doi:10.1111/j.1467-9817.2009.01418.x
- Pfost, M., Dörfler, T., & Artelt, C. (2012). Reading competence development of poor readers in a German elementary school sample: An empirical examination of the Matthew effect model. *Journal of Research in Reading*, 35(4), 411–426. doi:10.1111/j.1467-9817.2010.01478.x
- Pintrich, P.R., & Schunk, D.H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.
- Priess, H.A., & Hyde, J.S. (2010). Gender and academic abilities and preferences. In J.C. Chrisler & D.R. McCreary (Eds.), *Handbook of gender research in psychology* (Vol. 1, pp. 297–316). New York, NY: Springer. doi:10.1007/978-1-4419-1465-1_15
- R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.r-project.org
- Rodriguez, M.C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40(2), 163–184. doi:10.1111/j.1745-3984.2003.tb01102.x
- Routitsky, A., & Turner, R. (2003, April). *Item format types and their influences on cross-national comparisons of student performance*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley. doi:10.1002/9780470316696
- Schaffner, E., Phillip, M., & Schiefele, U. (2014). Reciprocal effects between intrinsic reading motivation and reading competence? A cross-lagged panel model for academic track and nonacademic track students. *Journal of Research in Reading*. Advance online publication. doi:10.1111/1467-9817.12027
- Schiefele, U., Schaffner, E., Möller, J., & Wigfield, A. (2012). Dimensions of reading motivation and their relation to reading behavior and competence. *Reading Research Quarterly*, 47(4), 427–463. doi:10.1002/RRQ.030
- Schwippert, K., Bos, W., & Lankes, E.M. (2004). Lesen Mädchen anders? Vertiefende Analysen zu Geschlechtsdifferenzen auf Basis der Internationalen Grundschul-Lese-Untersuchung IGLU [Do girls read differently? In-depth analyses of gender differences based on the Progress in Reading Literacy Study PIRLS]. *Zeitschrift für Erziehungswissenschaft*, 7(2), 219–234. doi:10.1007/s11618-004-0023-z
- Seipel, B., Clinton, V., & Carlson, S.E. (2012, May). *Gender differences in episodic memory reflected in text-connecting inferences*. Poster presented at the 24th annual meeting of the Association for Psychological Science, Chicago, IL.
- Senn, N. (2012). Effective approaches to motivate and engage reluctant boys in literacy. *The Reading Teacher*, 66(3), 211–220. doi:10.1002/TRTR.01107
- Simkin, M.G., & Kuechler, W.L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3(1), 73–98. doi:10.1111/j.1540-4609.2005.00053.x
- Snow, C.E. (2002). *Reading for understanding: Toward a R&D program in reading comprehension*. Santa Monica, CA: RAND.
- Snow, C.E., Burns, M.S., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Solheim, O.J. (2011). Impact of reading self-efficacy and task value on reading comprehension scores in different item formats. *Reading Psychology*, 32(1), 1–27. doi:10.1080/02702710903256601
- Swalander, L., & Taube, K. (2007). Influences of family-based prerequisites, reading attitude, and self-regulation on reading ability. *Contemporary Educational Psychology*, 32(2), 206–230. doi:10.1016/j.cedpsych.2006.01.002
- Taylor, C.S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, 25(3), 246–280. doi:10.1080/08957347.2012.687650
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (2013). *The role of international large-scale assessments: Perspectives from*

technology, economy, and educational research. New York, NY: Springer.

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(2), 103–118. doi:10.1207/s15324818ame0602_1

Wigfield, A., Eccles, J.S., Schiefele, U., Roeser, R.W., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon & R.M. Lerner (Series Eds.), & N. Eisenberg (Vol. Ed.), *Social, emotional, and personality development: Vol. 3. Handbook of child psychology* (6th ed., pp. 933–1002). Hoboken, NJ: Wiley.

Wigfield, A., & Guthrie, J.T. (1997). Relations of children's motivation for reading to the amount and breadth of their reading. *Journal of Educational Psychology*, 89(3), 420–432. doi:10.1037/0022-0663.89.3.420

Willingham, W.W., & Cole, N.S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *The Journal of Educational Research*, 80(6), 352–358.

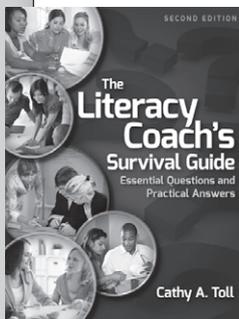
Zenisky, A.L., Hambleton, R.K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment*, 9(1/2), 61–78. doi:10.1080/10627197.2004.9652959

Submitted April 23, 2014
 Final revision received October 6, 2014
 Accepted October 12, 2014

FRANZISKA SCHWABE (corresponding author) is a research scientist at the Federal Institute for Education Research, Innovation and Development of the Austrian School System, Salzburg, and has formerly been employed at the Institute for School Development Research (IFS) at TU Dortmund University, Germany; e-mail f.schwabe@bifie.at. She conducts research on assessment of reading literacy in heterogeneous student populations.

NELE MCELVANY is a professor of empirical educational research and the director of the Institute for School Development Research (IFS) at TU Dortmund University, Germany; e-mail mcelvany@ifs.tu-dortmund.de. Her research interests include teaching and learning processes and outcomes in educational institutions as well as at home, and incorporate individual, social, and institutional factors for the development and promotion of reading literacy.

MATTHIAS TRENDTEL is a research scientist at the Federal Institute for Education Research, Innovation and Development of the Austrian School System, Salzburg; e-mail m.trendtel@bifie.at. He is principally involved in scaling, linking, and test equating, as well as automated test assembly in educational assessments.



© 2014
 ISBN 978-0-87207-156-8
 Nonmembers: \$28.95
 Members: \$23.15

The Literacy Coach's Survival Guide: Essential Questions and Practical Answers

Second Edition

CATHY A. TOLL

The Go-To Handbook for New and Experienced
 Literacy Coaches!

Heavily revised and updated to reflect changes in education, this go-to handbook guides new and experienced literacy coaches through important topics such as communicating well, effecting change, dealing with difficult situations, and coaching around special initiatives such as CCSS and RTI.



INTERNATIONAL
LITERACY
 ASSOCIATION

ORDER NOW! reading.org/literacycoach2

Enter this code for priority processing: LCSG

800.336.7323 (U.S. and Canada) | 302.731.1600 (all other countries)