Intelligence 58 (2016) 22–32

Contents lists available at ScienceDirect

Intelligence

# The influence of item sampling on sex differences in knowledge tests

Ulrich Schroeders [a,*], Oliver Wilhelm [b], Gabriel Olaru [b]

[a] Department of Educational Science, University of Bamberg, Germany
[b] Department of Psychology and Education, Ulm University, Germany

ARTICLE INFO

*Article history:*
Received 5 February 2016
Received in revised form 14 June 2016
Accepted 18 June 2016
Available online 1 July 2016

*Keywords:*
Crystallized intelligence
Declarative knowledge
Sex differences
Item sampling
Ant colony optimization

ABSTRACT

Few topics in psychology have generated as much controversy as sex differences in intelligence. For fluid intelligence, researchers emphasize the high overlap between the ability distributions of males and females, whereas research on sex differences in declarative knowledge often uncovers a male advantage. However, on the level of knowledge domains, a more nuanced picture emerged: while females perform better in health-related topics (e.g., aging, medicine), males outperform females in domains of natural sciences (e.g., engineering, physics). In this paper we show that sex differences vary substantially depending on item sampling. Analyses were based on a sample of $n = 3306$ German high-school students (Grades 9 and 10) who worked on the 64 declarative knowledge items of the *Berlin Test of Fluid and Crystallized Intelligence* (BEFKI) assessing knowledge within three broad content domains (science, humanities, social studies). Using two strategies of item sampling—stepwise confirmatory factor analysis and ant colony optimization algorithm—we deliberately manipulate sex differences in multi-group structural equation models. Results show that sex differences considerably vary depending on the indicators drawn from the item pool. Furthermore, ant colony optimization outperforms the simple stepwise selection strategy since it can optimize several criteria simultaneously (model fit, reliability, and preset sex differences). Taken together, studies reporting sex differences in declarative knowledge fail to acknowledge item sampling issues. On a more general stance, handling item sampling hinges on profound considerations of the content of measures.



## 1. Introduction

Few topics in ability research are regarded as controversial as sex or gender differences in cognitive abilities (e.g., Halpern, 2000; Neisser et al., 1996). According to the *Gender Similarity Hypothesis* (Hyde, 2005), sex differences in cognitive abilities are mainly small and unsystematic (Halpern et al., 2007; Halpern & LaMay, 2000). This general notion seems to hold true for fluid intelligence (*gf*), which reflects individual differences in the ability to "arrive at understanding relations among stimuli, comprehend implications, and draw inferences" (Horn & Noll, 1997, p. 69). The occasionally reported sex differences in fluid intelligence are either capitalizing on chance (due to a small sample size) or method artifacts. For example, Irwing and Lynn (2005) found for the *Progressive Matrices* a male advantage of nearly one third of a standard deviation among university students (see also Lynn et al., 2004). However, the meta-analysis seemed to be biased (e.g., study selection) and statistically flawed (e.g., using the median of the estimated differences; Blinkhorn, 2005). Therefore, the predominant perspective is that the similarities in the ability distributions across sex outweigh the small differences and that the negative consequences of overinflated claims regarding sex differences reinforce stereotypes (Hyde & Linn, 2006).

However, this consensus might not hold for declarative knowledge, because the majority of studies showed that males outperform females in general knowledge (Ackerman, Bowen, Beier, & Kanfer, 2001; Ackerman & Rolfhus, 1999; Camarata & Woodcock, 2006; Furnham, Christopher, Garwood, & Martin, 2007; Lynn, Irwing, & Cammock, 2001; Lynn, Wilberg & Margraf-Stiksrud, 2004; Steinmayr, Bergold, Margraf-Stiksrud, & Freund, 2015). In the intelligence literature, the terms "declarative knowledge" and "crystallized intelligence (*gc*)" are often used interchangeably (for different operationalizations of *gc* see Schipolowski, Wilhelm, & Schroeders, 2014). This dates back to Cattell's (1943, 1963, 1971) original broad definition of crystallized intelligence that included skills and knowledge in different content areas. More specifically, *gc* reflects the "breadth and depth of knowledge of the dominant culture" (Horn & Noll, 1997, p. 69), taking into account knowledge that is considered important, commonly rewarded in society, and a cultural good deemed worthy to impart to the next generation. Thus, *gc* has a high theoretical and empirical overlap with educational

* Corresponding author at: Bamberg Graduate School of Social Sciences, University of Bamberg, 96045 Bamberg, Germany.
*E-mail address:* ulrich.schroeders@uni-bamberg.de (U. Schroeders).

http://dx.doi.org/10.1016/j.intell.2016.06.003
0160-2896/

**Table 1**
Meta-analysis of sex differences in declarative knowledge.

| Study | *n* | Participants | Nation | Knowledge measure | Cohen's *d*[a] |
|---|---|---|---|---|---|
| Ackerman et al. (2001) | 320 | Freshmen<br>17–20 years | USA | 19 knowledge tests (e.g., physics, art, U.S. literature) | 0.39 [0.26; 0.53] |
| Beier and Ackerman (2003)[b] | 345 | 19–70 years | USA | 10 health-related knowledge tests (e.g., aging, bones, nutrition) | −0.56 [−0.67; −0.46] |
| Camarata and Woodcock (2006) | 10,465 | 5–79 years, heterogeneous sample | USA | Academic knowledge test (science, social studies, and humanities), based on the standardization samples of WJ-77, WJ-R, and WJ III | 0.21 [0.17; 0.24] |
| Engelberg (2015) | 247 | Grades 11 and 12, Gymnasium | Germany | 11 knowledge tests with an intended oversampling of domains with a female advantage (e.g., biology, nutrition, pedagogics, social work) | 0.08 [−0.12; 0.28] |
| Furnham, Christopher, Garwood, and Martin (2007) | 430 | 17–43 years, university students, female overrepresented | UK | General knowledge test, six broad domains (i.e., literature, general science, medicine, games, fashion and finance) | 0.24 [0.15; 0.33] |
| Hossiep and Schulte (2007) | 2,415 | 14–78 years, mainly graduates, male overrepresented (59.1%) | Germany | BOWIT, 11 knowledge tests (e.g., arts/architecture, biology/chemistry), assigned to two broader domains (humanities/social studies and natural/technical sciences) | 1.03 [1.12; 0.95] |
| Keith, Reynolds, Patel, and Ridley (2008) | 6,156 | 6–59 years | USA | Academic knowledge test (general information science, geography, cultural information), WJ III | 0.18 [0.13; 0.23] |
| Liepmann, Beauducel, Brocke, and Amthauer (2007) | 661 | 15–60 years, mostly graduates | Germany | Knowledge test with six different domains (i.e., geography/history, economics, science, mathematics, arts, and daily life), IST 2000-R | 0.36 [0.21; 0.52] |
| Lynn et al. (2001)[c] | 635 | Undergraduates<br>11–48 years | UK | General knowledge test with 19 knowledge domains assigned to six broad domains (i.e., current affairs, fashion, family, arts, science, and physical health/recreation) | 0.32 [0.15; 0.50] |
| Lynn, Wilberg and Margraf-Stiksrud (2004) | 302 | Grade 12 | Germany | German version of the general knowledge test, 17 knowledge domains (e.g., sport, politics, medicine, film) | 0.30 [0.08; 0.52] |
| Steinmayr et al. (2015)[c] | 977 | Grades 11 and 12, Gymnasium<br>16-18 years | Germany | Knowledge test with six different domains (i.e., geography/ history, economics, science, mathematics, arts, and daily life), IST 2000-R | 0.50 [0.26; 0.74] |
| Wilhelm et al. (2014)[c] | 4,213 | Grades 8–10<br>13–18 years | Germany | 16 knowledge domains (e.g., literature, chemistry, finance, politics), assigned to three broader domains (science, humanities, and social studies), BEFKI 8-10 | 0.04 [−0.05; 0.13] |
| Overall effect | | | | | 0.26 [0.06; 0.47] |

*Note.* The meta-analysis (*random-effect model*) was conducted with the R package *metafor* (Viechtbauer, 2010). Following the usual convention, positive values indicate an advantage for men.
[a] Values in brackets represent the lower and upper boundaries of the 95% confidence interval.
[b] The scales measuring technology and current-events were excluded because they were administered in a power format.
[c] Meta-analyzed on the level of subtests. BEFKI 8–10 = Berlin Test of Fluid and Crystallized Intelligence for Grades 8–10 (Wilhelm et al., 2014) BOWIT = Bochumer Wissenstest (Hossiep & Schulte, 2007) IST 2000-R = Intelligence-Structure-Test 2000 R (Liepmann et al., 2007) WJ-77 = Woodcock–Johnson Psycho-Educational Battery (Woodcock & Johnson, 1977) WJ-R = Woodcock–Johnson Psycho-Educational Battery—Revised (Woodcock & Johnson, 1989) WJ III = Woodcock–Johnson III (Woodcock, McGrew, & Mather, 2001).

achievement (Ackerman & Lohman, 2003). In the following, we reanalyze and measure—in accordance with Cattell's original definition—*gc* in terms of a factual or declarative knowledge test.

To get an overview of previous studies with different knowledge tests (e.g., IST 2000-R, General Knowledge Test) and samples (e.g., in terms of age, ability, country), we summarize the findings with a meta-analysis (see Table 1). A literature search using PsycInfo, Psyndex, and Google Scholar and the following search term combinations: (gender OR sex) AND differences AND (crystallized intelligence OR knowledge) yielded 12 relevant studies (total $N = 27{,}166$). In comparison to a descriptive and qualitative review or simply averaging effect sizes, we deem this meta-analytical approach superior because it allows to estimate the average effect size more precisely and to assess the heterogeneity of the results (see Cumming, 2010). Please note that we conducted a thorough, but limited literature research because, as we argue later, the variation of sex differences not only lies *between* the studies, but *within*. The random-effect meta-analysis was conducted with the R package *metafor* (Viechtbauer, 2010) and yielded an overall effect of $d = 0.26$ ($CI_{95\%}$: 0.06–0.47).[1] The homogeneity of the effect was tested with Cochran's *Q*-test (Cochran, 1954), which computes the sum of the squared deviations of each study's estimate from the overall estimate, and $I^2$, which relates the total heterogeneity to the total variability (Higgins, Thompson, Deeks, & Altman, 2003).[2] Both statistics pointed to large heterogeneity in the effect sizes: $Q(df = 11) = 450.69$, $p < .0001$, and $I^2 = 98.28\%$ (with values between 75 and 100% indicating "considerable heterogeneity", Higgins & Green, 2008, p. 278). An examination on the level of domains revealed a more complex pattern and demonstrated that women clearly outperform men in health-related domains such as aging and nutrition (Beier & Ackerman, 2003), whereas large differences in favor of males were found for technology and the natural sciences (Ackerman et al., 2001). This pattern could also be replicated for the domains of the *Berlin Test of Fluid and Crystallized Intelligence* (BEFKI; Wilhelm, Schroeders, & Schipolowski, 2014; see Fig. 1). There are small advantages for female students in the humanities, whereas the effects were reversed for the natural sciences and social studies. This level of aggregation may also represent a simplification of the actual conditions, since the health-related questions of medicine also showed a female advantage (in line with Beier & Ackerman, 2003). Nevertheless, there is preliminary evidence that both direction and size of sex differences vary substantially across domains, which indicates that the effects depend to a large extent on the sampling of domains and presumably also of items.

[1] Following the usual convention, positive values indicate an advantage for the male group.

[2] The ratio can be expressed as follows: $I^2 = 100\% \times (Q - df) / Q$, where *Q* is Cochran's heterogeneity statistic and *df* the degrees of freedom.

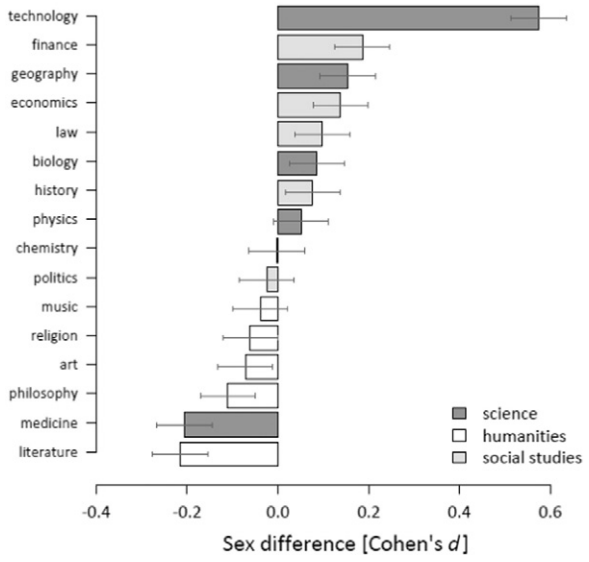


**Fig. 1.** Sex differences as a function of domain in the Berlin Test of Fluid and Crystallized Intelligence. *Note.* Mean sex differences were based on four items for each of the sixteen knowledge tests ($n = 4213$). Error bars indicate the 95% confidence intervals of Cohen's *d*. Knowledge domains are ordered from the highest values (indicating male advantage) to lowest values (indicating female advantage). Shading of the bars reflect theoretically derived broader knowledge domains.

### 1.1. The origin of sex differences in knowledge — trait difference or test bias?

If a psychological measure identifies a mean level difference between groups (e.g., sex or race), it could be either due to a true difference in the trait or due to a test bias (e.g., Reynolds & Suzuki, 2012). If sex differences reflect true trait differences between groups, what are the determinants of these effects? Different explanations are discussed (see also Engelberg, 2015), for example, that intellectual investment traits are essential for knowledge acquisition. In a recent meta-analysis, substantial relations between *need for cognition* (NFC) or *typical intellectual engagement* (TIE) and general knowledge were reported (von Stumm & Ackerman, 2012). On a domain level, it has been pointed out that the relation between TIE is stronger "to humanities-type knowledge than to the sciences" (Rolfhus & Ackerman, 1999, p. 513). The sex differences especially in the TIE facet *reading* that have been reported in the literature (Schroeders, Schipolowski, & Böhme, 2015; Wilhelm, Schulze, Schmiedek, & Süß, 2003) could partially account for the small female advantage in the humanities. Differences in interest have also been suggested to be causally related to differences in knowledge (e.g., Halpern et al., 2007). In the PPIK theory (*intelligence-as-process, personality, interests, and intelligence-as-knowledge*; Ackerman, 1996; Ackerman & Heggestad, 1997), Ackerman tied specific interests to certain knowledge structures; for example, the investigative interest is more closely related to physical and social science than to arts and literature. In a meta-analysis, Su, Rounds, and Armstrong (2009) reanalyzed the data of 47 interest inventories and found higher interest for males in STEM (e.g., in engineering, $d = 1.11$, and science, $d = 0.36$) and in the realistic domain of Holland's RIASEC model ($d = 0.84$). In contrast, females showed more interest in the social ($d = -.68$) and artistic ($d = -.35$) domains. These sex differences in interest match the differences in the knowledge domains described above. For the differences in interest, socialization processes in terms of sex role types and stereotypical expectations are most likely essential. Such a complex and mutual interplay between biological and social/environmental variables and its joint influence on cognitive abilities is proposed in the biopsychosocial model of Halpern (2000, 2004) and Halpern et al. (2007).

Besides true differences in the trait, a measurement bias could account for the sex differences. According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), bias is present when test scores are not solely dependent on the construct in question, but also dependent on another variable—in the present case, membership in a sex group. Different procedures for testing measurement bias across groups have been proposed. For observable variables the most commonly used methods are the Mantel–Haenszel procedure (Mantel & Haenszel, 1959), the non-parametrical test for simultaneous item bias (SIBTEST; Shealy & Stout, 1993), and the standardization method (Dorans & Kulick, 1986). These approaches are easily set up, but are also outdated (Osterlind & Everson, 2009). Contemporary approaches use latent variable modeling, that is, item response theory (IRT) models and confirmatory factor analysis (CFA) models (e.g., Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993). The IRT perspective of bias is often related to the concept of *differential item function* (DIF), which refers to a difference on a specific item after controlling for ability differences. In contrast, the CFA perspective is often associated with the concept of *measurement invariance* (MI; e.g., Wicherts, 2007). With respect to a specific item, differences between groups can originate either from differences in the intercepts (corresponds to uniform DIF in the IRT context) or the factor loadings (non-uniform DIF). Besides different tradition and focus on the item level (IRT) or scale level (CFA) there is a large conceptual overlap between both methods (Millsap, 2011). The logic behind both modeling approaches is to identify items that contribute to test bias and to level out potential sex differences by removing these items.

However, Steinmayr et al. (2015) showed that this attempt does not necessarily lead to a test version without sex differences: removing nearly half of the items of a German general knowledge test that showed a substantial amount of DIF resulted in a smaller but still substantial male advantage of $d = .32$ (instead of $d = .78$). Such studies indicate that the removal of biased items can change characteristics of a measure—including its validity—substantially. Therefore, we tackle the question of sex differences in declarative knowledge from a perspective of item sampling.

### 1.2. The item universe and generalization across persons, items, and occasions

Analogous to the selection of persons from a population participating in a study (*person sampling*), items can be thought of as being drawn from a population of items (*item sampling*). Furthermore, in the construction and validation of psychological measures, it is usually assumed that items are drawn from a theoretically infinite item universe (Markus & Borsboom, 2013; McDonald, 1999, 2003). According to Cronbach and Meehl (1955, p. 282), content validity is often established deductively "by defining a universe of items and sampling systematically within this universe". The item sampling process has to provide a representative sample, not necessarily a random one (Robitzsch, 2015, p. 163). Whereas the generalizability of results across different samples is often discussed in the psychological literature, the influence of specific item sets on test scores and trait levels—and *mutatis mutandis* to sex differences—is usually neglected and assumed to be small or zero.

The issue of item sampling is explicitly handled in the *Generalizability Theory* (Brennan, 2001; Cronbach, Linn, Brennan, & Haertel, 1997) by disentangling and quantifying different sources of variation (or facets). These facets usually include a) persons, b) items, and c) occasions or time. In this context, generalization means to replicate the results of an initial study across different samples, measures, and occasions and use the information to estimate the generalizability of the findings. Even though this approach can determine what role item sampling plays for the occurrence of sex differences, it requires a complex test design and comprehensive data. More precisely, analyses within the Generalizability Theory are often carried out in a within-subject design, that is, all subjects working on all levels of all facets (Shavelson & Webb, 2006). Perhaps due to these restrictions, the question of item specificity is not often raised in the assessment of intelligence. We want to present

a more parsimonious method to examine the effect of item sampling within a given measure.

For conventional psychological measures, items are considered to be drawn from a larger pool of potential items, but often neither the domain nor the item selection process is clearly defined. In the present paper, we apply the idea of item sampling to the final item set of a knowledge test and study its influence on sex differences. In contrast to previous research that was mainly concerned with asking if males outperform females in a certain knowledge domain, the aim of this study is to show that the answer to such questions crucially depends on the items sampled. More precisely, we want to demonstrate that depending on the item subsample sex differences can substantially change and favor either males *or* females. Thus, we intend to illustrate the importance of item sampling for the validity of a measure. In order to maximize sex differences, two different item strategies are implemented: on the one hand, we gradually remove items that increase sex differences on a latent level in confirmatory factor analysis. On the other hand, we apply *ant colony optimization* algorithms that are suitable to optimize several criteria (model fit, reliability of the scale, and maximized sex differences). Because it has been previously pointed out that shortening a measure with respect to a certain criterion can have detrimental effects on the reliability and validity of a test (Kruyen, Emons, & Sijtsma, 2013; Schipolowski, Schroeders, & Wilhelm, 2014; Ziegler, Kemper, & Kruyen, 2014), we compare both item selection strategies with respect to sex differences, factor saturation, and overall model fit to examine in what way the shortening of the measure affects its reliability.

## 2. Method

### 2.1. Design and participants

Data derived from the German standardization sample of the *Berlin Test of Fluid and Crystallized Intelligence* (Wilhelm et al., 2014). The sample included students from all German federal states (except for the smallest one) and all school types of the general educational system. In the German school system students are relatively early separated into different types of schools (usually after Grade 4), which are—depending on the federal state—differently cut and labeled. In general, the following types of schools can be distinguished: academic-track schools (*Gymnasium*), intermediate-track schools (*Realschule*), vocational-track schools (*Hauptschule*), mixed-track school types (e.g., *Integrierte Gesamtschule*), and schools with multiple educational qualifications. The standardization sample was part of a large-scale educational pilot study that intentionally left out more capable students in Grade 8; no students from *Realschule* and *Gymnasium* were sampled for this specific grade. Such an oversampling could lead to a male disadvantage in knowledge, due to the male disadvantage in the segregated school system. Therefore, we limited our analyses to Grades 9 and 10. Table 2 shows the distribution of students across school types in the present sample in comparison to the distribution in the population. Overall, the deviations are small and almost level out across grades. Analyses presented in this paper were based on a sample of $n = 3306$ students working on the *gc* section. Half of the sample was female (49.5%); sex was reported by school officials. Participation in the study was mandatory and students were not graded or rewarded in any way (see Table 2).

In terms of demographics, the sample was representative of German students at the end of secondary education. The socio-economic situation of the family was assessed by means of the *International Socio-Economic Index* (ISEI; Ganzeboom, De Graaf, & Treiman, 1992), which is based on the income and education level of the parents. The calculation of the ISEI is based on the ISCO-88 classification system of occupations by the International Labour Office (1990), which ranks professions hierarchically on a common scale ranging from 16 to 90 points. As a best proxy of the available socio-economic resources, we used the highest ISEI (HISEI) of the family. The mean was $M = 48.7$ ($SD = 16.1$), which, for example, corresponds to "physical and engineering science technicians" or "customer services clerks". The average and *SD* almost perfectly matched the values reported in a representative large-scale study in Germany in 2009 ($n = 2559$, $M = 48.7$, $SD = 15.8$; Köller, Knigge, & Tesch, 2010, p. 244). The sample was equally representative in regards to migration status: 17.2% of the students had a migration background—as defined by at least one parent been born abroad (Stanat & Christensen, 2006)—compared to 17.6% in the respective large-scale study (Köller et al., 2010, p. 214). The main countries of origins were a) Turkey, b) Russia, Kazakhstan or other former Soviet republic, and c) Poland.

### 2.2. Measurement instrument

Knowledge was assessed with 64 multiple-choice items of the *Berlin Test of Fluid and Crystallized Intelligence* (Wilhelm et al., 2014) that was composed of three subscales: *sciences*, *humanities*, and *social studies*. Each subscale included several content domains: physics, chemistry, biology, medicine, geography, technology (sciences), art, literature, music, religion, philosophy (humanities), and history, law, politics, economy, finance (social studies). Item development and the composition of knowledge domains aimed to cover the "breadth and depth of the knowledge of the dominant culture" (Horn & Noll, 1997, p. 69), taking into account both curriculum-related and out-of-school knowledge that is commonly deemed culturally valuable and considered socially important. Thus, the knowledge arguably differs from trivial, incidental, and short-lived knowledge, such as soccer results or bus schedules. The final item set was compiled based on several pilot-studies testing the psychometric property of several hundred items. All items had four response alternatives, one of which being correct. For the sample at hand, reliability of the original 64-item version in terms of McDonald's $\omega$ (1999), was .83 for *science*, .79 for *humanities*, and .80 for *social studies*.

**Table 2**
Representativeness of the data with respect to school types.

| | Population | | | | Sample | | | | Difference | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Absolute | | Relative [%] | | Absolute | | Relative [%] | | Relative [%] | |
| Grade | 9 | 10 | 9 | 10 | 9 | 10 | 9 | 10 | 9 | 10 |
| School type | | | | | | | | | | |
| HS | 9,422 | 4,349 | 28.11 | 15.25 | 734 | 69 | 34.56 | 5.84 | 6.45 | −9.41 |
| MBG | 2,857 | 2,155 | 8.52 | 7.56 | 214 | 164 | 10.08 | 13.87 | 1.55 | 6.32 |
| RS | 8,385 | 8,985 | 25.01 | 31.50 | 416 | 380 | 19.59 | 32.15 | −5.43 | 0.65 |
| Gym | 10,059 | 10,201 | 30.01 | 35.76 | 415 | 487 | 19.54 | 41.20 | −10.47 | 5.44 |
| IGS | 2,797 | 2,833 | 8.34 | 9.93 | 345 | 82 | 16.24 | 6.94 | 7.90 | −2.99 |

Note. HS = *Hauptschule* (vocational-track school); MBG = schools with multiple educational qualifications; RS = *Realschule* (intermediate track school); Gym = *Gymnasium* (academic-track school); IGS = *Integrierte Gesamtschule* (mixed-track school). Sample sizes of the population are taken from the official school statistics (Federal Office of Statistics, 2010, p. 40).

### 2.3. Statistical analyses

Sex differences are maximized based on two item selection strategies—*stepwise confirmatory factor analysis* (SCOFA) and *ant colony optimization* (ACO). Data preparation, recoding, and analyses were conducted with *R 3.2.0* (R Development Core Team, 2011); CFA models were estimated with the R package *lavaan 0.5–20* (Rosseel, 2012). The ACO script is a revised and adopted version of the script provided by Leite (2015) and is available from the authors' website (Schroeders, Wilhelm, & Olaru, 2016).

#### 2.3.1. Stepwise confirmatory factor analysis (SCOFA)

The first method of item sampling adopts a simple stepwise approach. More precisely, in the first iteration 64 multiple-group CFAs (MGCFAs) are estimated; each model excludes a different item and includes the other 63 items. The item causing the greatest mean difference is removed from the item set and the procedure is repeated for the reduced item set. The process is reiterated until the predetermined number of items for the short version is reached (e.g., 32 items); in this algorithm the relative weighting of the content domains is retained (i.e., 12 items for *natural science* and 10 items each for *humanities* and *social studies*). MGCFAs are estimated with equal factor loadings and thresholds across sex groups; these constraints of measurement parameters correspond to strong measurement invariance (for the procedure of MI testing with categorical data see Schroeders & Wilhelm, 2011), which is deemed a prerequisite for comparing the means of latent variables, that is, true group differences (Vandenberg & Lance, 2000). Furthermore, for estimation purpose we set the variance of the latent variables to 1 in both groups. The CFAs were estimated with the *Weighted Least Squares Mean and Variance adjusted* (WLSMV) estimator, which is superior to a maximum likelihood estimator for categorical data in terms of model rejection rates and appropriateness of the factor loadings (Beauducel & Herzberg, 2006). Values of the *Comparative Fit Index* (CFI) ≥ .95 and values of the *Root Mean Square Error of Approximation* (RMSEA) ≤ .08 were taken as indication of good model fit (Hu & Bentler, 1999).

#### 2.3.2. Ant colony optimization algorithm (ACO)

The second method for maximizing differences across sex groups while maintaining model fit and factor saturation is an ant colony optimization algorithm. Such agent-based models are used to solve various problems of combinatorial optimization (Dorigo & Stützle, 2010). Originally, these algorithms date back to the systematic and meticulous observations of the foraging behavior of Argentine ants (Deneubourg, Aron, Goss, & Pasteels, 1990; Goss, Aron, Deneubourg, & Pasteels, 1989). In psychological assessment, the universal mechanisms of ACO have been used to compile effective short test versions of existing long forms (Janssen et al., in press; Leite, Huang, & Marcoulides, 2008; Olaru, Witthöft, & Wilhelm, 2015).

These algorithms are not yet widely used in psychology. Therefore, we shortly outline the procedure for the current context. Different sets of items (= ants) are randomly drawn from the larger item pool of the long version. For each item set, a structural equation model is estimated and evaluated with respect to an optimization criterion (= shortest route), for example, model fit or the magnitude of sex differences. Items belonging to an item set that optimized best, obtain a higher probability to get selected in the next iteration. Fig. 2 illustrates the change in the probability of utilization across iterations. Initially, all items have the same probability which is represented by every item having the same gray value (row 1). With every iteration the pattern becomes more distinct and refined: Items with a higher probability of getting selected for the final solution are represented in darker gray values.

In comparison to other approaches of drawing item samples, ACO has several advantages: first, agent-based algorithms such as ACO were specifically developed to solve complex problems that are so computationally demanding that a full search for the proper solution is not feasible. For example, the complete computational solution for finding the best 32 item short version out of a unidimensional measure with 64 item demands the calculation of 1,832,624,140,942,592,256 models. As a metaheuristic approach ACO is more efficient than the complete computational solution, but it may not find the best solution (Dorigo & Stützle, 2010). Second, the optimization can be done with respect to several criteria simultaneously (Janssen et al., in press). For example, sex differences in favor of one group or the other can be maximized while simultaneously retaining model fit. Traditional methods focus on optimizing a single criterion. Third, ACO is a flexible method and relatively free of prerequisites. For instance, it can be used on a manifest or latent level. Fourth, the number of items per factor can be set in advance, maintaining the relative strength of the factors and the representativeness of the content of a measure. Removing items proportionally reduces the risk of a shift in the meaning of the construct due to biased selection across factors (Smith, McCarthy, & Anderson, 2000). Fifth, ACO's strength involves evaluating the effects of item sampling for the overall model. In contrast to DIF analyses, this also means that optimization and evaluation have the same unit of analysis (a specific item set rather than an item). Finally, the results generated by ACO are not affected by sequence effects, as item samples of fixed size are directly drawn from the item pool instead of removing indicators in a stepwise manner. Stepwise selection procedures are dependent on the sequence of items removed and are prone to local optima during the selection process, with no further improvements possible.

The optimization function in the present case consisted of three parts: a) overall model fit including measurement invariance, b) factor saturation, and c) latent mean differences between sex groups. ACO models were estimated for several abbreviated test forms (39, 36, 32, 26 and 23 items); sex differences were maximized either in favor of females or males. With respect to model fit, we used a combination of the incremental fit index *Comparative Fit Index* (CFI) and the absolute fit

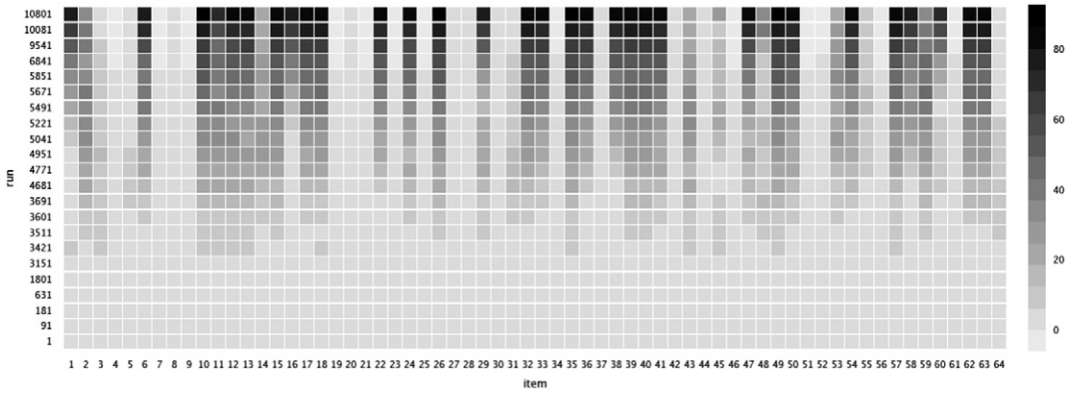


**Fig. 2.** Changes in drawing probabilities for each item across iterations. *Note.* Darker shades of gray indicate higher selection probabilities for an item.

index *Root Mean Square Error of Approximation* (RMSEA)—as proposed in the two-index strategy presentation (Hu & Bentler, 1999). Analogous to Janssen et al. (in press), model parameters were logit-transformed in order to scale the value range between 0 and 1 and to differentiate most around a given cutoff value. For example, values for the CFI above .95 correspond to a transformed pheromone level greater than .50. In contrast to values at the extremes, small model fit differences near the cutoff are weighted more heavily.

$$\varphi_{CFI} = \frac{1}{1 + e^{95 - 100CFI}} \tag{1}$$

In addition to the incremental fit index CFI, the RMSEA was used as absolute fit index. The RMSEA cutoff value indicating good model fit was .05.

$$\varphi_{RMSEA} = 1 - \frac{1}{1 + e^{5 - 100RMSEA}} \tag{2}$$

In order to interpret meaningful differences in the means of latent variables between sex groups, we estimated a strong measurement invariance model (i.e., equal factor loadings and thresholds across groups). Ideally, measurement invariance is tested in a series of models in which more and more measurement parameters are constrained. Violations to parameter constraints do not necessarily manifest in absolute model fit indices, rather they will lead to a deterioration in model fit in relation to a less constrained model. In comparison to invariance testing with continuous variables, the procedure to test for measurement invariance differs for categorical data (Muthén & Asparouhov, 2002; Schroeders & Wilhelm, 2011), because factor loadings and thresholds have to be varied in tandem. In the first model of measurement invariance testing, configural invariance, only the pattern of loadings has to be identical across sex groups. In the continuous case, the second step is metric invariance (i.e., equal factor loadings across groups), which has no direct equivalent in the categorical case. The next step, therefore, is strong measurement invariance with equal factor loadings and thresholds. Usually, a difference in CFI > .01 between two consecutive models is considered a serious deterioration in model fit (Cheung & Rensvold, 2002). Therefore, in the present case we added a term that favors differences in CFI between the configural and strong measurement invariance models below.01:

$$\varphi_{MI} = 1 - \frac{1}{1 + e^{5 - 500\ \Delta CFI}} \tag{3}$$

with

$$\Delta CFI = abs\left(CFI_{config} - CFI_{strong}\right) \tag{4}$$

All three parts, that is, model fit in terms of CFI and RMSEA, and the aspect of measurement invariance are averaged:

$$\varphi_{Fit} = \frac{\varphi_{CFI} + \varphi_{RMSEA} + \varphi_{MI}}{3} \tag{5}$$

The second criterion in the optimization function dealt with an index of measurement precision (Mellenbergh, 1996): McDonald's $\omega$ (1999). This coefficient represents factor saturation in a three-dimensional factor model and relates the squared sum of the factor loadings to the sum of the residuals (see formula 6). In contrast to Cronbach's $\alpha$ that requires an essentially $\tau$-equivalent measurement model (Zinbarg, Revelle, Yovel, & Li, 2005), McDonald's $\omega$ is suitable also in a $\tau$-congeneric case (i.e., varying instead of fixed factor loadings), which is often encountered in real data sets. Even though it has been correctly pointed out that any cutoff values for reliability estimates should be treated with caution (Lance, Butts, & Lawrence, 2006), we consider values greater than .70 satisfactory in the present case. Another reason for optimizing factor saturation is that cutoff values for model fit cannot be interpreted independently of the factor loadings. Therefore, allowing for low factor loadings "might draw an overoptimistic picture of model fit" (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011, p. 330). Reliability coefficients were calculated for all three factors in both groups and averaged in order to get a proxy for the average factor saturation for the female and male group:

$$\varphi_{Rel} = \frac{1}{1 + e^{7 - 10\omega}} \quad \text{with } \omega = \frac{\omega_{sci} + \omega_{hum} + \omega_{soc}}{3}$$
$$\text{and } \omega_{\text{factor}} = \frac{\left(\sum_{i=1}^{n} \lambda_i\right)^2}{\left(\sum_{i=1}^{n} \lambda_i\right)^2 + \sum_{i=1}^{n} 1 - \lambda_i^2} \tag{6}$$

The last term in the optimization functions was the average mean difference in the three-dimensional model across sex groups. The average standardized mean differences in the original 64-item model showed a minimal advantage for males: .080 = (0.223 (*science*) − 0.158 (*humanities*) + 0.174 (*social studies*)) / 3. For the abbreviated test forms, we estimated two series of models: One that maximized sex differences in favor of males:

$$\varphi_{Diff} = 1 + 3\left(\frac{\alpha_{sci} + \alpha_{hum} + \alpha_{soc}}{3} - .080\right) \tag{7}$$

and one that maximized sex differences in favor of females:

$$\varphi_{Diff} = 1 + 3\left(\frac{-\alpha_{sci} - \alpha_{hum} - \alpha_{soc}}{3} - .080\right) \tag{8}$$

The overall optimization function took into account all three pheromone trails, but not with equal weight, as $\varphi_{Fit}$ and $\varphi_{Rel}$ used logit-transformed values (see Eqs. (1), (2), and (4)), whereas $\varphi_{Diff}$ started with a value of 1 for no sex difference. Because we were mainly interested in maximizing group differences in factor means, the adjusted mean difference was multiplied by 3, thus weighting these pheromones more strongly (see Eqs. (7) and (8)):

$$\max f(x) = \varphi_{Fit} + \varphi_{Rel} + \varphi_{Diff} \tag{9}$$

## 3. Results

The original three-dimensional model included the three factors of *science* (24 items), *humanities* (20 items), and *social studies* (20 items). These 64 items provided a good fit to the data: $\chi^2 = 5690.45$ ($df = 3959$), CFI = .952, and RMSEA = .016. The standardized mean differences for the long version were 0.405 for *science*, −0.158 for *humanities*, and 0.174 for *social studies*. In the following, we present the results for both selection strategies (ACO and SCOFA) that optimized either in favor of males or females (see formulas (7) and (8)). The upper panel of Fig. 3 shows the sex difference in the three factor means for the different abbreviated short versions, where negative values represent a female advantage. The lines represent the item selection strategies (i.e., SCOFA and ACO) that were used once to select items with a maximum female advantage, and a second time to maximize the advantage for males. For example, the ACO approach in favor of females compiled a 26 item version (10 items for *science* and 8 items each for *humanities* and *social studies*) with a mean sex difference in science knowledge of −.26, in humanities of −.68, and in social studies of −.09. In comparison to the long version, there was a significant shift in the means (see solid line in Fig. 3, upper panel). The same shift in the opposite direction was found, if the ACO algorithm was used to maximize the male advantage: for the 26 item version, the group mean difference was .53 for science, .02 for humanities, and .40 for social studies. Keeping in mind that ACO finds an efficient approximation—not necessarily the best solution—these results might not represent the solution with the largest difference.

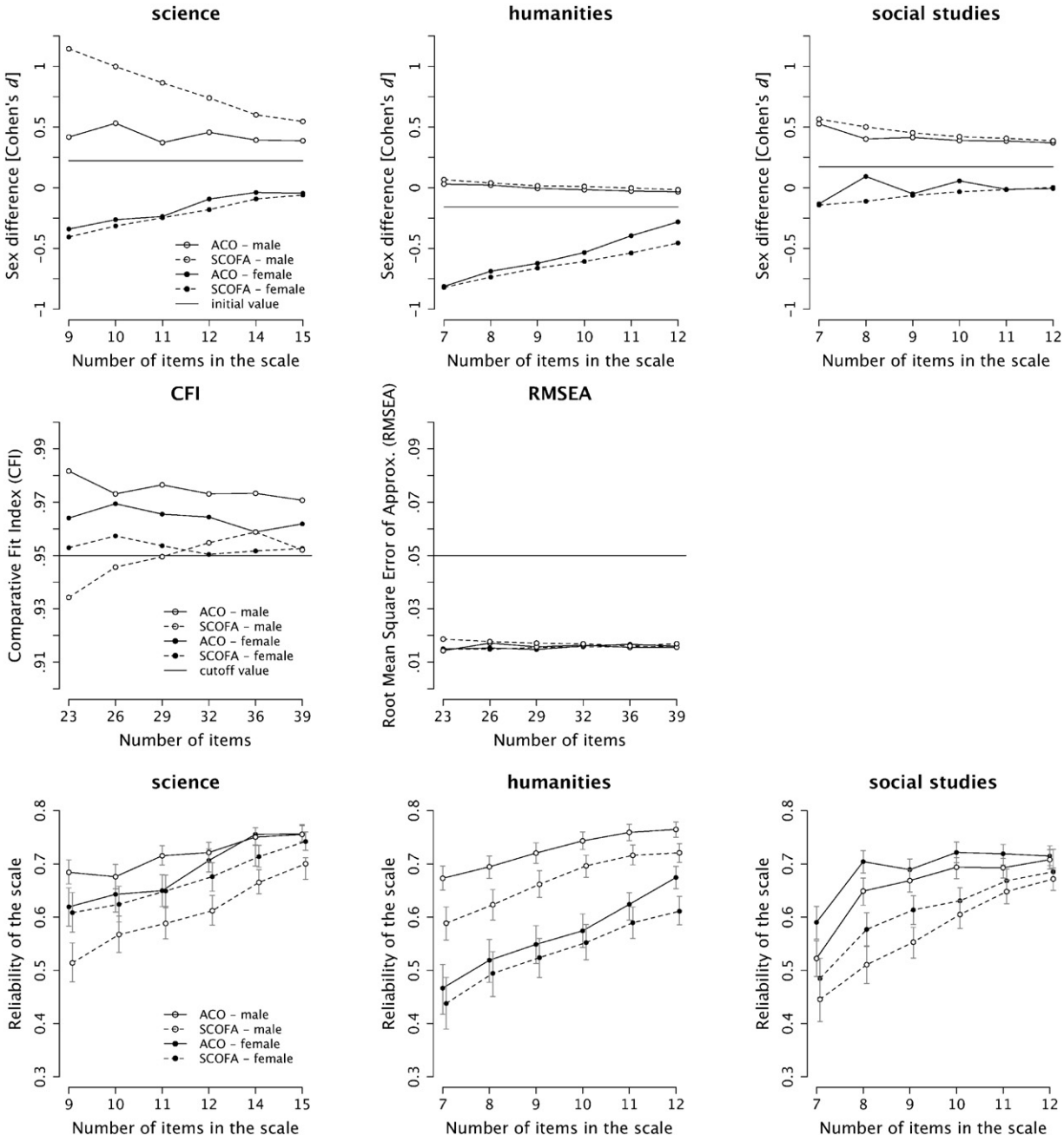


**Fig. 3.** Effects of item sampling on sex differences (upper panel), model fit (middle panel), and reliability of the scale (lower panel). *Note*. ACO = ant colony optimization algorithm. SCOFA = stepwise confirmatory factor analysis. Upper panel: Solid lines indicate the mean sex differences in the long version. Middle panel: Solid lines represent the cutoffs indication good model fit (Hu & Bentler, 1999). Lower panel: Error bars indicate the 95% confidence intervals computed with naive bootstrapping (Padilla & Divers, 2013; Zhang & Yuan, 2016).

Several conclusions can be derived from the comparison of mean sex differences (Fig. 3, upper panel): (1) Item sampling has a large impact on the size and direction of the sex differences; for almost all models, the item sets favor either males or females (in reference to the absolute zero point). (2) The sex differences get higher for stronger abbreviated test versions (with only few exceptions). (3) For the optimization in favor of females, both ACO and SCOFA provided similar results for all factors. (4) For males, ACO and SCOFA also provided almost identical results for the *humanities* and *social studies*, yet different results for science. Thus, as to maximize sex differences, SCOFA outperformed ACO, but this comes at the cost of unreliability (Fig. 3, lower panel) and is due to some measurement variant items (Table 3). Model fit was sufficient for both item selection strategies (Fig. 3, middle panel), more specifically, CFI ranged between .934 and .959 for SCOFA, and between .959 and .982 for ACO; the range of the RMSEA was .015–.019 for SCOFA and .014–.017 for ACO. Since the difference in CFI between the configural and strong measurement invariant models was part of the optimization function for the ACO models (see formulas (3) to (5)), there were (expectedly) no significant deviations (Table 3). However, the differences for the models that were derived from the SCOFA algorithm maximizing male advantages were quite large and beyond the recommended threshold of $\Delta CFI = .01$ (Cheung & Rensvold, 2002). That is, these measurement models were *not* strong invariant preventing a conclusive comparison of the latent variable means. Therefore, the reported male advantages for SCOFA were inflated by invariant items. With respect to the reliability of the scale that was identically estimated as McDonald's (1999) for both selection strategies (see $\omega_{factor}$ in formula (6)), item selection with ACO yielded superior results for all short forms and all factors (Fig. 3, lower panel). Error bars indicate the 95% confidence intervals computed with naive bootstrapping, which

**Table 3**
Measurement invariance testing for short forms SCOFA and ACO.

| | | SCOFA | | | ACO | | |
|---|---|---|---|---|---|---|---|
| | Items | Configural | Strong | Δ(CFI) | Configural | Strong | Δ(CFI) |
| Female | 23 | .958 | .953 | .005 | .970 | .964 | .006 |
| | 26 | .962 | .957 | .005 | .973 | .969 | .004 |
| | 29 | .958 | .954 | .004 | .971 | .966 | .005 |
| | 32 | .956 | .950 | .006 | .968 | .964 | .004 |
| | 36 | .957 | .952 | .005 | .963 | .959 | .004 |
| | 39 | .959 | .953 | .006 | .967 | .962 | .005 |
| Male | 23 | .959 | .934 | .025 | .983 | .982 | .001 |
| | 26 | .971 | .946 | .025 | .979 | .973 | .006 |
| | 29 | .974 | .950 | .024 | .978 | .977 | .001 |
| | 32 | .976 | .955 | .021 | .979 | .973 | .006 |
| | 36 | .976 | .959 | .017 | .980 | .973 | .007 |
| | 39 | .973 | .952 | .021 | .976 | .971 | .005 |

*Note*. ACO = ant colony optimization algorithm. SCOFA = stepwise confirmatory factor analysis. Configural = equal item-factor structure across sex groups. Strong = equal factor loadings and thresholds across sex groups. For more information on measurement invariance testing with categorical data see *Statistical analyses*. ΔCFI > .01 between consecutive models is considered a serious deterioration in model fit (Cheung & Rensvold, 2002).

repeatedly resamples from the original data set with replacement (Padilla & Divers, 2013; Zhang & Yuan, 2016). Naive bootstrapping has been recommended for obtaining confidence intervals for parameter estimates in SEM (Nevitt & Hancock, 2001). The advantage of ACO over SCOFA concerning knowledge in sciences and the humanities was statistically significant only within the male group. Although the present sample is quite large, there is some chance of random fluctuation, as indicated, for example, by the larger confidence intervals for *humanities* in the female group. Put differently, the specific item sets analyzed here might be responsible for the non-significance of the ACO–SCOFA difference in females. Similar efforts to investigate sex differences might use new item sets or a more fine-grained analysis across domains, thus, adding evidence to superiority of ACO broadly found here.

## 4. Discussion

In this study, we examined to what extent the magnitude of sex differences in declarative knowledge frequently reported in the research literature (e.g., Ackerman & Rolfhus, 1999; Beier & Ackerman, 2003; Lynn, Wilberg et al., 2004) is affected by item sampling. Previous research on sex differences was largely concerned with removing items that showed substantial bias across sex groups in order to derive or approach an unbiased test form (e.g., Steinmayr et al., 2015) without broaching the topic of item selection. Interestingly, the notion that sex differences in cognitive abilities, such as mathematical abilities, vary if samples do not represent the populations of males and females had been previously discussed (Hyde, Fennema, & Lamon, 1990). With respect to the item side, we demonstrated that sex differences also vary largely with different item sampling procedures—from effects in favor of females to the opposite direction—solely depending on the composition of the reduced item set. These findings advocate strong item specificity of sex differences, thereby questioning the soundness of effect sizes reported in the literature.

From a measurement perspective, research on sex differences in cognitive abilities is often concerned with the aggregation of findings from different studies and measures across several domains (Else-Quest, Hyde, & Linn, 2010; Hyde, 2014). Such systematic reviews and meta-analytical studies are indispensable, but they can only answer the question about a true sex difference if item sampling is explicitly considered. Please note that with the study at hand we did not attempt to find an answer to the true size of sex differences. Instead, we argue that there is no possible answer to this question until sampling issues are more profoundly addressed. In the literature of educational large-scale assessment, this fact is more thoroughly acknowledged (Rutkowski, Gonzales, Joncas, & von Davier, 2010; Wu, 2010). In large-scale educational studies, such as PISA (*Programme for International Student Assessment*), several hundred items are administered in a multiple matrix sampling design, facilitating a sufficient content coverage of the constructs. In contrast, in intelligence research, measurement issues of item (or test) sampling are exacerbated because only a small subset of the item universe is sampled. For declarative knowledge tests, this item sampling applies both to the selection and compilation of the knowledge domains (e.g., humanities vs. natural sciences) and the actual items within a given domain. In the present case, items for each broad content area (e.g., science) form a unidimensional scale. Thus, findings concerning the mean structure differences are not the result of mixing up items of different dimensionality. Put differently, even *within* one domain the range of sex differences is large. For example, items assessing knowledge in medicine might show a female advantage if topics such as aging and nutrition are covered (Beier & Ackerman, 2003) versus a male advantage when more technical topics are involved (e.g., X-rays, neurotransmitters).

On a more general stance, the consequences of the reported results also affect the construction of measures of maximal effort in general, because one common assumption is that the final item set is randomly drawn from an infinite item universe. Items are supposed to be used interchangeably, which is an idea that is particularly intriguing for declarative knowledge items for which it is hard to predict sex differences. However, the results of the present analyses question this assumption. Given that the item-universe-analogy is one of the fundamental principles in item construction, the degree of fluctuation is remarkable. The variation in group differences we found connects to earlier cautionary notes of psychometricians and theorists in the 1960s (see also Kane, 2002). For example, Loevinger (1965, p. 147) noted that the random sampling assumption of items (and tests) is unrealistic, because test development is "almost invariably expert selection rather than sampling". Even more disquieting, she pointed out that the "term population implies that in principle one can catalog or display, or index all possible members even though the population is infinite and the catalogue cannot be completed". In some closely circumscribed and narrow measures tapping, such as mental speed or working memory, the item generation process is sufficiently clear so that items can be generated automatically (for a good example of automated, ability tailored item development see Schmiedek, Lövdén, & Lindenberger, 2010). In the case of declarative fact knowledge measures, we are clearly not sampling from a population. From a practical point of view, Thorndike (1966, p. 288) added that test developers tend to design items in a special way (e.g., wording, distractor construction), which means that the item "universe is considerably restricted, is hard to define, and the sampling from it is hardly to be considered random." Against this background, theory-driven test construction efforts that take into account domain sampling when considering sex group differences should be highly appreciated (see Engelberg, 2015, for an example). For the construction of tests in other domains, it is best to rely on theoretical assumptions about gender differences and to sample accordingly. For example, a math competency test should include word and mental rotation tasks (Halpern et al., 2007). However, often test authors cannot rely on sound theory and the empirical findings are inconclusive. In this case, a viable approach could be to construct and administer several parallel test forms with a broad sampling in order to quantify the influence of item sampling. Finally, as presented, item sampling issues could also be addressed for a preselected set of items with ACO.

We demonstrated the influence of item sampling on mean differences in sex groups. It is likely that the examination of other sociodemographic variables, such as ethnicity or school tracks, would reveal a similarly strong dependence of the group differences on the items sampled. Importantly, such considerations are not restricted to mean differences of latent factors in a multiple-group CFA context, but also apply to other analyses and other parameters. Item sampling strategies can be used to maximize (or minimize) effects in both reflective

and formative models. The influence of item sampling on group differences or persons' estimates may be very prominent in formative models in which items are viewed as causes of constructs (Edwards & Bagozzi, 2000). Formative modeling is also discussed in the context of intelligence assessment, currently most prominent in the mutualism model (van der Maas et al., 2006), as an alternative conceptualization for the classical *g*-factor model (van der Maas, Kan, & Borsboom, 2014). In this view, cognitive ability tests correlate substantially (*positive manifold*) not because a hypothesized underlying entity *g* is working (*reflective model*); rather, this is due to the mere result of reciprocal positive interactions between abilities and developmental processes. Accordingly, *g* is understood as index variable without causal meaning (*formative model*). A distinctive feature between both conceptualizations is the role of the indicators. In reflective models, indicators are in principle interchangeable: within acceptable boundaries adding, removing, or replacing indicators should not change the nature of the construct. In this sense, we included a term in the optimization function that favors models with high factor saturation. In formative models, however, the optimization problem is different. The mix of indicators influences or constitutes the construct in question and there is no optimal way to compute such a composite score. Thus, one strategy could be to maximize the prediction of relevant outcomes (van der Maas et al., 2014). The optimization function of ACOs can be easily adapted to find an index that maximizes the correlation to some external variables (e.g., job performance). Taken together, even though item sampling is essential in both formative and reflective models, the procedure by which good indicators are sampled can differ substantially.

In the present application, we showed that ACO is a powerful tool to create an abridged test version that fulfills several conditions (model fit, factor saturation, and sex differences). Often, reducing the item pool with respect to a single criterion (i.e., maximizing sex differences) comes at a cost. Among others, this shortcoming could refer to the content coverage of each factor in the abridged test, the reliability of the scale, or the factor structure (Smith et al., 2000). Accordingly, the stepwise confirmatory factor analysis with only one criterion showed reasonable model fit, but decreasing reliability estimates, whereas ACO took into account model fit and factor saturation, which resulted in good fit and reliability. Therefore, our findings add evidence to the body of research that shows that ACO outperformed simple item selection strategies (Janssen et al., in press; Leite et al., 2008; Olaru et al., 2015). Optimizing several criteria is crucial to accomplish a comprehensive evaluation of model fit (Heene et al., 2011), and ACO is a useful tool for this purpose.

For some test developers, the idea may sound discomforting that a “dumb” optimization algorithm can find better and unbiased solutions than a human. This reservation seems to be fueled by the fact that experts' judgment of the item content is often regarded as a valid and indispensable source of information in item selection in addition to test or item statistics (Kruyen et al., 2013). However, in principle every item characteristic (including linguistic features, complexity, and content) can be explicitly coded and incorporated in the optimization process. With a sufficiently large item pool, ACO can then be used to optimize a specific criterion set in order to develop psychometrically sound measurement instruments. For instance, one could argue that it is best to remove any sex difference while maintaining the depth and breadth of the knowledge test, which at least renders separate sex norms obsolete. Of course, the presented item sampling methods can be used to achieve this goal of a sex-fair measurement instrument. However, this would only allegedly solve fairness issues often encountered in psychological assessment: if true sex differences exist, this procedure would provide biased person parameters in the sense that items selected for the final test form do not represent representative sample of the item universe (see also AERA, APA, & NCME, 2014 on fairness in testing).

## References[3]


Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, *22*, 227–257. http://dx.doi.org/10.1016/S0160-2896(96)90016-1.

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, *121*, 219–245. http://dx.doi.org/10.1177/1069072703011002006.

Ackerman, P. L., & Lohman, D. F. (2003). Education and g. In H. Nyborg (Ed.), *The Scientific Study of General Intelligence* (pp. 275–292). Oxford: Pergamon.

Ackerman, P. L., & Rolfhus, E. L. (1999). The locus of adult intelligence: Knowledge, abilities, and nonability traits. *Psychology and Aging*, *14*, 314–330. http://dx.doi.org/10.1037/0882-7974.14.2.314.

Ackerman, P. L., Bowen, K. R., Beier, M., & Kanfer, R. (2001*). Determinants of individual differences and gender differences in knowledge. *Journal of Educational Psychology*, *93*, 797–825. http://dx.doi.org/10.1037/0022-0663.93.4.797.

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

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 186–203. http://dx.doi.org/10.1207/s15328007sem1302_2.

Beier, M. E., & Ackerman, P. L. (2003*). Determinants of health knowledge: An investigation of age, gender, abilities, personality, and interests. *Journal of Personality and Social Psychology*, *84*, 439–448. http://dx.doi.org/10.1037/0022-3514.84.2.439.

Blinkhorn, S. (2005). Intelligence: a gender bender. *Nature*, *438*, 31–32. http://dx.doi.org/10.1038/438031a.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer New York.

Camarata, S., & Woodcock, R. (2006*). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, *34*, 231–252. http://dx.doi.org/10.1016/j.intell.2005.12.001.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129. http://dx.doi.org/10.2307/3001666.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. http://dx.doi.org/10.1037/h0040957.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399. http://dx.doi.org/10.1177/0013164497057003001.

Cumming, G. (2010). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York, London: Routledge.

Deneubourg, J. -L., Aron, S., Goss, S., & Pasteels, J. M. (1990). The self-organizing exploratory pattern of the argentine ant. *Journal of Insect Behavior*, *3*, 159–168.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Dorigo, M., & Stützle, T. (2010). Ant colony optimization: overview and recent advances. In M. Gendreau, & J. -Y. Potvin (Eds.), *Handbook of Metaheuristics* (pp. 227–263) Springer US.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*, 155–174. http://dx.doi.org/10.1037/1082-989X.5.2.155.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127. http://dx.doi.org/10.1037/a0018053.

Engelberg, P. M. (2015*). *Ursachen für Geschlechterdifferenzen in Tests des Allgemeinen Wissens [Causes of sex differences in general knowledge tests].* Doctoral dissertation Wuppertal, Germany: University of Wuppertal.

Federal Office of Statistics (2010). *Fachserie 11, Schuljahr 2007/08 [Technical series 11, school year 2007/08].* Wiesbaden: Statistisches Bundesamt.

Furnham, A., Christopher, A. N., Garwood, J., & Martin, G. N. (2007*). Approaches to learning and the acquisition of general knowledge. *Personality and Individual Differences*, *43*, 1563–1571. http://dx.doi.org/10.1016/j.paid.2007.04.013.

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard internationalsocio-economic index of occupational status. *Social Science Research*, *21*, 1–56. http://dx.doi.org/10.1016/0049-089X(92)90017-B.

Goss, S., Aron, S., Deneubourg, J. -L., & Pasteels, J. M. (1989). Self-organized shortcuts in the Argentine ant. *Naturwissenschaften*, *76*, 579–581.

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science*, *13*, 135–139. http://dx.doi.org/10.1111/j.0963-7214.2004.00292.x.

Halpern, D. F., & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review*, *12*, 229–246.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest*, *8*, 1–51. http://dx.doi.org/10.1111/j.1529-1006.2007.00031.x.


[3] All references marked with * were included in the meta-analysis.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*, 319–336. http://dx.doi.org/10.1037/a0024917.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 53–91). New York: Guilford Press.

Hossiep, R., & Schulte, M. (2007*). *BOWIT - Bochumer Wissenstest [BOWIT — Bochum Knowledge Test]*. Göttingen: Hogrefe.

Higgins, J. P. T., & Green, S. (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, England: John Wiley & Sons.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*, 557–560. http://dx.doi.org/10.1136/bmj.327.7414.557.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. http://dx.doi.org/10.1080/10705519909540118.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592. http://dx.doi.org/10.1037/0003-066x.60.6.581.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*, 373–398. http://dx.doi.org/10.1146/annurev-psych-010213-115057.

Hyde, J. S., & Linn, M. C. (2006). Gender similarities in mathematics and science. *Science*, *314*, 599–600. http://dx.doi.org/10.1126/science.1132154.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139–155. http://dx.doi.org/10.1037/0033-2909.107.2.139.

International Labour Office (1990). *International standard classification of occupation(ISCO-88*. Genf: International Labour Office.

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*, 505–524. http://dx.doi.org/10.1348/000712605X53542.

Janssen, A. B., Schultze, M., & Grötsch, A. (2015). Following the ants: Development of short scales for proactive personality and supervisor support by ant colony optimization. *European Journal of Psychological Assessment*. http://dx.doi.org/10.1027/1015-5759/a000299 (in press).

Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement*, *39*, 165–181.

Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008*). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, *36*, 502–525.

Köller, O., Knigge, M., & Tesch, B. (Eds.). (2010). *Sprachliche Kompetenzen im Ländervergleich. [Language skills across German federal states]*. Münster: Waxmann.

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*, 223–248. http://dx.doi.org/10.1080/15305058.2012.703734.

Lance, C. E., Butts, M. M., & Lawrence, C. M. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*, 202–220. http://dx.doi.org/10.1177/1094428105284919.

Leite, W. L. (2015). Ant colony optmization (ACO) algorithm [computer software]. Retrieved January 1, 2016, from http://education.ufl.edu/leite/code/

Leite, W. L., Huang, I. -C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*, 411–431. http://dx.doi.org/10.1080/00273170802285743.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007*). *Intelligenz-Struktur-Test 2000R [Intelligence-Structure-Test 2000R]*. Göttingen: Hogrefe.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*, 143–155. http://dx.doi.org/10.1037/h0021704.

Lynn, R., Allik, J., Pullmann, H., & Laidra, K. (2004a). Sex differences on the progressive matrices among adolescents: Some data from Estonia. *Personality and Individual Differences*, *36*, 1249–1255. http://dx.doi.org/10.1016/S0191-8869(02)00240-4.

Lynn, R., Irwing, P., & Cammock, T. (2001*). Sex differences in general knowledge. *Intelligence*, *30*, 27–39. http://dx.doi.org/10.1016/S0160-2896(01)00064-2.

Lynn, R., Wilberg, S., & Margraf-Stiksrud, J. (2004*). Sex differences in general knowledge in German high school students. *Personality and Individual Differences*, *37*, 1643–1650. http://dx.doi.org/10.1016/j.paid.2004.02.018.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748. http://dx.doi.org/10.1093/jnci/22.4.719.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

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

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*, 212–230.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299. http://dx.doi.org/10.1037/1082-989X.1.3.293.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

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

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., ... Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101. http://dx.doi.org/10.1037/0003-066X.51.2.77.

Nevitt, J., & Hancock, G. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 353–377. http://dx.doi.org/10.1207/S15328007SEM0803_2.

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale Big-Five assessments. *Journal of Research in Personality*, *59*, 56–68. http://dx.doi.org/10.1016/j.jrp.2015.09.001.

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning*. Thound Oaks, CA: Sage Publications.

Padilla, M. A., & Divers, J. (2013). Bootstrap interval estimation of reliability via coefficient omega. *Journal of Modern Applied Statistical Methods*, *12*, 78–89. http://dx.doi.org/10.1177/0013164413492765.

R Development Core Team (2011). *R: A language and environment for statistical computing (Version 3.2.0)*. Vienna: R Foundation for Statistical Computing Retrieved from http://www.R-project.org/.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529. http://dx.doi.org/10.1037/0021-9010.87.3.517.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566. http://dx.doi.org/10.1037/0033-2909.114.3.552.

Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment. An empirical review and recommendations. In I. B. Weiner, J. A. Naglieri, & J. R. Graham (Eds.), *Handbook of psychology. Volume 10. Assessment psychology* (pp. 82–113). Hoboken, New Jersey: John Wiley & Sons Ltd.

Robitzsch, A. (2015). *Essays zu methodischen Herausforderungen im Large-Scale Assessment [Essays on methodological challenges in large-scale assessment] (Doctoral dissertation)*. Berlin, Germany: Humboldt-Universität zu Berlin.

Rolfhus, E. L., & Ackerman, P. L. (1999). Assessing individual differences in knowledge: Knowledge, intelligence, and related traits. *Journal of Educational Psychology*, *91*, 511–526. http://dx.doi.org/10.1037/0022-0663.91.3.511.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. http://dx.doi.org/10.18637/jss.v048.i02.

Rutkowski, L., Gonzales, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*, 142–151. http://dx.doi.org/10.3102/0013189X10363170.

Schipolowski, S., Schroeders, U., & Wilhelm, O. (2014). Pitfalls and challenges in constructing short forms of cognitive ability measures. *Journal of Individual Differences*, *35*, 190–200. http://dx.doi.org/10.1027/1614-0001/a000134.

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: the relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156–168. http://dx.doi.org/10.1016/j.intell.2014.05.014.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, *2*, 1–10. http://dx.doi.org/10.3389/fnagi.2010.00027.

Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, *71*, 849–869. http://dx.doi.org/10.1177/0013164410391468.

Schroeders, U., Schipolowski, S., & Böhme, K. (2015). Typical intellectual engagement and achievement in math and the sciences in secondary education. *Learning and Individual Differences*, *43*, 31–38. http://dx.doi.org/10.1016/j.lindif.2015.08.030.

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Ant Colony Optmization (ACO) Algorithm [Computer software]. Retrieved from http://ulrich-schroeders.de/en/publikationen/.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In J. L. Green, G. Camilli, P. B. Elmore, A. Skukauskaiti, & E. Grace (Eds.), *Handbook of complementary methods in education research* (pp. 309–322). Washington D.C.: Lawrence Erlbaum Associates.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*, 102–111. http://dx.doi.org/10.1037/1040-3590.12.1.102.

Stanat, P., & Christensen, G. S. (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003*. Paris: OECD.

Steinmayr, R., Bergold, S., Margraf-Stiksrud, J., & Freund, P. A. (2015). Gender differences on general knowledge tests: Are they due to differential item functioning? *Intelligence*, *50*, 164–174. http://dx.doi.org/10.1016/j.intell.2015.04.001.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*, 859–884. http://dx.doi.org/10.1037/a0017364.

Thorndike, R. L. (1966). Reliability. In A. Anastasi (Ed.), *Testing problems in perspective* (pp. 284–291). Washington, D.C.: American Counsil of Education.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*, 842–861. http://dx.doi.org/10.1037/0033-295X.113.4.842.

van der Maas, H. L. J., Kan, K. -J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence*, *2*, 12–15. http://dx.doi.org/10.3390/jintelligence2010012.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70. http://dx.doi.org/10.1177/109442810031002.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. http://dx.doi.org/10.18637/jss.v036.i03.

von Stumm, S., & Ackerman, P. L. (2012). Investment and intellect: A review and meta-analysis. *Psychological Bulletin*, *139*, 841–869. http://dx.doi.org/10.1037/a0030746.

Wicherts, J. M. (2007). *Group differences in intelligence test performance.* Doctoral dissertation Amsterdam, The Netherlands: University of Amsterdam.

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014*). Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. *Jahrgangsstufe [Berlin test of fluid and crystallized intelligence for grades 8–10].* Göttingen: Hogrefe.

Wilhelm, O., Schulze, R., Schmiedek, F., & Süß, H. -M. (2003). Interindividuelle Unterschiede im typischen intellektuellen Engagement [Individual differences in typical intellectual engagement]. *Diagnostica*, *49*, 49–60. http://dx.doi.org/10.1026//0012-1924.49.2.49.

Woodcock, R. W., & Johnson, K. S. (1977). *Woodcock–Johnson psycho-educational battery.* Itasca, IL: Riverside Publishing.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson psycho-educational battery—Revised.* Allen, TX: DLM.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III.* Itasca, IL: Riverside Publishing.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*, 15–27. http://dx.doi.org/10.1111/j.1745-3992.2010.00190.x.

Zhang, Z., & Yuan, K. -H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, *76*, 387–411. http://dx.doi.org/10.1177/0013164415594658.

Ziegler, M., Kemper, C., & Kruyen, P. (2014). Short scales — Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*, 185–189. http://dx.doi.org/10.1027/1614-0001/a000148.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's $\alpha$, Revelle's $\beta$, and McDonald's $\omega_H$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133. http://dx.doi.org/10.1007/s11336-003-0974-7.