Measurement and structural invariance of cognitive ability tests after computer-based training

Hermes, Michael; Albers, Frank; Böhnke, Jan; Huelmann, Gerrit; Maier, Julia; Stelling, Dirk

Published in:
Computers in Human Behavior

DOI:
10.1016/j.chb.2018.11.040

Publication date:
2019

Measurement and structural invariance of cognitive ability tests after computer-based training

Michael Hermes[a], Frank Albers[a], Jan R. Böhnke[b], Gerrit Huelmann[a], Julia Maier[a], and Dirk Stelling[a]

[a] German Aerospace Center DLR, Department of Aviation and Space Psychology, Hamburg, Germany

[b] Dundee Centre for Health and Related Research, School of Nursing and Health Sciences, University of Dundee, Dundee, UK

Paper accepted for publication at Computers in Human Behavior

Author Note

Correspondence concerning this article should be addressed to Michael Hermes.

Michael Hermes, German Aerospace Center DLR, Aviation and Space Psychology, Sportallee 54a, 22335 Hamburg, Germany.

E-mail: michael.hermes@dlr.de, Phone: +49 40 513096 44, Fax: +49 40 513096 60

Declarations of interest: none.

Highlights

- The link between computer-based training and high-stakes assessments was investigated.
- Training and selection test data of 15,752 pilot trainee applicants was analyzed.
- The amount of training predicted test performance in curvilinear fashion as expected.
- The ability test scores' structure was invariant across different amounts of training.
- Free training was not linked to the psychometric structure of the high-stakes tests.

Measurement and structural invariance of cognitive ability tests after computer-based training

Abstract

Ability tests are core elements in performance research as well as in applied contexts and are increasingly carried out using computer-based versions. In the last few decades a whole training and coaching industry has developed to prepare individuals for computer-based assessments. Evidence suggests that such commercial training programs can result in score gains in ability tests, thereby creating an advantage for those who can afford it and challenging the fairness of ability assessment. As a consequence, several authors recommended freely offering training software to all participants to increase measurement fairness. However, it is still an open question whether the unsupervised use of training software could have an impact on the measurement properties of ability tests. The goal of the present study is to fill this gap by examining the subjects' ability scores for measurement and structural invariance across different amounts of computer-based training. Structural equation modeling was employed in a sample of 15,752 applicants who participated in high-stakes assessments with computer-based ability tests. Across different training amounts, our analyses supported measurement and structural invariance of ability scores. In conclusion, free training software is a means that provides fair preparation opportunities without changing the measurement properties of the tests.

Keywords: computer-based training, computer-based testing, cognitive ability, measurement invariance, test fairness

Measurement and structural invariance of cognitive ability tests after computer-based training

1. Introduction

In the last decades the use of computers in psychological assessment has grown enormously. In the domain of cognitive abilities, computer-based assessments have increasingly substituted paper-pencil tests and are now core elements in performance research and applied contexts such as personnel selection, admission decisions in educational contexts, or neuropsychological assessments. In these applications computer-based tests allow the assessment of cognitive abilities with high levels of standardization and offer the possibility to realize designs that could not be appropriately implemented with paper-pencil tests (Greiff, Scherer, & Kirschner, 2017; Tippins, 2015).

However, particularly in high-stakes settings, there are still challenges with cognitive ability tests. One is the requirement of test fairness in the sense of ensuring that all test takers have comparable opportunities to demonstrate the abilities measured by the test. With cognitive ability tests, it is of particular importance to provide examinees with equal opportunities to prepare for the test (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). In the interest of fairness, the materials provided should closely resemble the actual test with regards to appearance and format (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Yet, the requirement of equal opportunities for test preparation is increasingly challenged by a growing market of service companies offering commercially distributed training software. Although the quality of such software may vary substantially between providers, there is

evidence that training[1] may result in substantial score gains in cognitive ability tests, not only in educational but also in selection settings (Chung-Herrera, Ehrhart, Ehrhart, Solamon, & Kilian, 2008; Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007). It can be expected that only individuals with sufficient financial resources will be able to afford such training programs, calling into question the fairness of ability assessments (Sackett, Burris, & Ryan, 1989; Stemig, Sackett, & Lievens, 2015). Such training gains can be partly attributed to retest effects, i.e. score gains resulting from the mere repetition of a test (Freund & Holling, 2011; Hausknecht et al., 2007). This implies that not only individuals with training course experience but also individuals who had the opportunity to repeat an examination may have an advantage over individuals who conduct an ability test for the first time. Taken together, commercially available training and the retest policies of institutions give rise to considerable concern about the fairness of these ability assessments. To deal with this problem, several authors suggested to freely offer training and practice materials to all participants (Arendasy et al., 2016; Freund & Holling, 2011; Sackett, Borneman, & Connelly, 2008; Zwick, 2002). For example, Arendasy et al. (2016) proposed that "making more informal student-centered practice opportunities accessible to all test-takers could resolve issues of fairness associated with differential access to test preparation opportunities without compromising measurement fairness" (p. 54).

Today most organizations involved in high-stakes computer-based testing actually provide opportunities for test familiarization by distributing free practice items or tutorials as recommended by the International Test Commission (The International Test Commission,

[1] It is important to note that the terms training, practice, retest, and coaching are not exclusive categories but belong a continuum of preparation activities, with different emphasis on test familiarization and the development of test-specific skills (Arendasy, Sommer, Gutiérrez-Lobos, & Punter, 2016; Messick & Jungeblut, 1981).

2006). The range of complexity of practice materials currently offered by test-administering institutions is, however, mostly restricted to downloadable information brochures and example items in paper and pencil format. There is some slightly more sophisticated online material available. It must be expected though, that in case of cognitive ability tests, the practice gains increase as a function of the equivalence of the practice items with test items (Hausknecht et al., 2007). Therefore, to ensure test fairness, it is crucial to offer practice items that resemble the test items as closely as possible. Recently, more efforts have been made to increase fairness. In the United States for example, the College Board cooperated with the Khan Academy to offer freely available test practice programs for the SAT (formerly Scholastic Aptitude Test), a common test for college admissions. Overall, however, still only little consideration is given to the unequal opportunities to practice and training activities, for example by making effective preparation material freely available to all test takers and informing them openly about available training options.

One reason for this deficiency may be that in the context of personnel selection there is virtually no published research on the effects and consequences of using sophisticated computer-based training tools, especially on the validity of the tests. This is surprising given the great potential of computer-based training tools in this field. With computer-based systems it is possible to offer training tools with high equivalence to selection tests. Such systems allow a very standardized presentation of training items (including timing and navigation issues), complex interfaces and item-selection algorithms, a wide range of multimedia features, reliable response recording, or feedback mechanisms. Finally, computer-based training systems can be economically distributed, making them especially attractive for larger user pools.

2. Possible effects of training

For the current study possible consequences of practice and training are inferred from two related research areas. The first area comprises research regarding practice and retest effects (Sackett et al., 1989; Scharfen, Peters, & Holling, 2018; Van Iddekinge & Arnold, 2017), the second includes research regarding the acquisition of cognitive skills (Ackerman, 1987, 1988; Fitts, 1964).

Since Ebbinghaus' learning experiments it is well-established that practice of cognitive tasks normally results in asymptotic learning curves (Donner & Hardy, 2015; Heathcote, Brown, & Mewhort, 2000; Newell & Rosenbloom, 1981). Practice effects seem to occur in virtually all types of mental ability tests, with especially large sizes of effect for psychomotor coordination and spatial orientation tests (Sackett et al., 1989). Practice effects are generally larger when identical compared to alternate test forms are employed, i.e. practice gains increase the more equivalent both test forms are (Hausknecht et al., 2007; Scharfen et al., 2018). Since training with computer-based training modules can be considered as practice with parallel test forms, the possible consequences of freely offering training software on measurement properties may be derived from studies that analyze the impact of practice and retesting on test validity.

Although research has shown that criterion-oriented validity is either unaffected or positively affected by practice and retest effects (Van Iddekinge & Arnold, 2017), the empirical literature for construct validity is rather mixed. An early study with psychomotor tests showed that the factor structure of a psychomotor task changed over practice to a simpler pattern (Fleishman & Hempel, 1954). Current research has demonstrated evidence of measurement invariance over retest scores of several cognitive ability tests (Reeve & Lam, 2005) but also of non-invariance (Lievens, Reeve, & Heggestad, 2007). Here, variance in retest scores reflected test-specific abilities to a larger degree than variance of initial scores. This could be attributed to a reduction of construct irrelevant variance, like reduced test

anxiety or greater experience with the test and the test setting (Hausknecht et al., 2007; Lievens, Buyse, & Sackett, 2005; Van Iddekinge & Arnold, 2017). However, to our knowledge, there is no research on the question whether different training quantities could have an impact on the invariance of cognitive ability measurements. Different amounts of training could influence the measured variables without changing the basic ability of interest, thereby creating measurement bias. Likewise, different amounts of training could change the relations between different constructs.

For the invariance of cognitive ability measurement theories of skill acquisition play an important role as well. Most of these theories suggest that practice of cognitive or psychomotor tasks may lead to the development of a (rather narrow) skill set (Fitts, 1964; Rosenbaum, Carlson, & Gilmore, 2001). Therefore, there is a certain risk that the intense use of computer-based training tools for cognitive tasks results in a change of the construct from a rather broad ability to a narrow skill. In this case, the test conducted after extensive practice would not be useable to measure the ability. Ackerman (1987, 1988) presented a comprehensive theory which states that by practice, a skill is acquired in three phases in which different determinants are important for individual performance: content abilities in the first phase (which correspond to different aspects of intelligence), perceptual speed in the second, and psychomotor skills in the third phase. For the construct-oriented validity of a cognitive test this implies that for those who have the opportunity to practice, in every phase of skill acquisition (depending on the amount of practice) the test would assess a different underlying ability. Importantly, Ackerman's theory states that skill acquisition is only pertinent if the practice material is rather consistent and not complex (see also Ackerman & Schneider, 1985). For the construction of cognitive ability tests this implies that the test material should be rather complex and have an inconsistent structure so that even after intense practice, test-takers should never reach phases two or three of skill acquisition.

3. Goals of the current study

Ability diagnostic in high-stakes settings is in the predicament that on the one hand providing free training opportunities is a promising way to (re-) establish fair initial conditions for every test-taker. On the other hand, this opportunity to practice increases the risk that test-takers may develop a cognitive skill in the tasks in question, thereby compromising the construct validity of the cognitive tests. For the context of personnel selection Ackerman's theory (1987, 1988) shows a way of how the development of skills as a consequence of training may be prevented and the construct validity may be preserved: The nature of the tasks in the cognitive ability tests has to be more complex and inconsistent, so that more sophisticated and changing mental operations are required for the test items. The ability tests used in the present study were developed with a special focus on more complex and inconsistent material and items. Therefore we expected that training did not alter the measurement structure and the relations between constructs.

Hypothesis 1: Increasing levels of training do not alter the factor structure of the measurements.

Likewise, with reference to Ackerman's framework there should be a consistent relationship between the ability constructs and perceptual speed as well as psychomotor coordination for different practice levels.

Hypothesis 2a: Increasing levels of training do not alter the importance of perceptual speed within the factor structure.

Hypothesis 2b: Increasing levels of training do not alter the importance of psychomotor coordination within the factor structure.

In both cases we expect that the size of the respective factor loadings is not moderated by the amount of training. The goal of this study was to test these hypotheses in a sample of N = 15,752 applicants whose data were gathered in high-stakes application contexts. Since

the applicants were encouraged to use our freely distributed computer-based training tools, this sample provided a unique opportunity to test these hypotheses in a sample with high external validity. In contrast to previous studies, we used computer-based training tools that were specifically designed to resemble the different cognitive ability tests as closely as possible. In addition, the applicants provided data on their amount of individual training so that a quantitative measure of these training activities was available.

4. Method

4.1. Participants and procedure

The present sample consisted of candidates applying for a pilot trainee position at major European airlines. Applicants were included in the present study if they had participated in an assessment between 2010 and 2017. During this period, all applicants were offered identical computer-based training programs and conducted the same ability tests. The present study only considered applicants who had participated for the first time in an assessment in our institution. Data sets with missing values were excluded from the analyses (there were 100 applicants, i.e. 0.63 % of the sample with incomplete training data) so that only complete data sets were used. The remaining 15,752 candidates were between 17 and 50 years old (M = 20.30, SD = 2.68), 86% were male and 14% were female. All applicants had completed a high school education adequate for university entrance.

In the following data from a multi-stage selection procedure were analyzed. The procedure started with a set of computer-based tests measuring cognitive abilities, knowledge and psychomotor abilities, as required for aviation training and the pilot's profession. Further stages comprised a flight in a fixed-base flight simulator, an assessment center, and an interview. Examinations with computer-based tests were conducted in groups of up to 44 subjects. The measurement procedure was highly standardized; all tests were administered at the same time of the day and in an identical order. Applicants completed all tests within one

day, in an air-conditioned and well-lit testing room. Each ability test started with an introductory text on the screen; the cognitive ability tests also included examples and a few practice items. Before the main test began, comprehension questions were answered by the test administrator if necessary.

4.2. Measures

For the present study, we analyzed nine computer-based tests. Six tests measured basic cognitive abilities: visual perception speed, selective attention, auditory and visual memory capacity, mental rotation and spatial visualization (both aspects of spatial abilities). In order to examine possible effects of the training on the relation between the measured construct of cognitive ability and other constructs, we also included two tests measuring technical knowledge and comprehension and one test measuring psychomotor coordination and the capacity for multiple task coordination. All tests have been developed in our institution and are employed in pilot selection for several years. The following cognitive ability tests were employed:

The Optical Perception Test (test-retest reliability r_{tt} = .90; Zierke, 2014) as a measure of visual perception speed required the subject to read four specific indicator values from a complex display consisting of nine dials. The dials differed in color (black/white) and shape (round/angular) and were displayed for two seconds. Before each task was presented, information was given about which dials were critical (black/white/round/angular) and thus which had to be read.

The Symbol Concentration Test (r_{tt} = .93; Zierke, 2014) as measure of selective attention required the subject to apply changing rules to long sequences of displayed triangles. The triangles differed in color, orientation, and number of dots displayed in each triangle. Before each sequence, a rule indicated which of these features were critical. The

subjects had to decide whether or not two consecutive triangles were identical with respect to the critical features.

The Running Memory Span Test (r_{tt} = .76; Zierke, 2014) is a measure of auditory memory. It required the subject to memorize acoustically presented sequences of digits and enter them in reverse order. The sequences differed in length of up to 35 digits.

The Visual Memory Capacity Test (r_{tt} = .74; Hermes & Stelling, 2016) contained an n-back task, requiring the subjects to compare a sequence of symbols and to react when the present symbol matched the one from n steps earlier in the sequence. The symbols differed in shape and color. Two-back, 3-back, 4-back, and 5-back sequences were administered.

In the Mental Rotation Test (r_{tt} = .91; Zierke, 2014) the subject had to visualize a cube with one face marked by a cross. Then sequences of acoustic orders were presented—differing in length and speed—which specified how to mentally rotate the cube. At the end of each sequence the subjects had to indicate the position of the mark on the cube.

The material in the Spatial Visualization Test (Cronbach's α = .91; Zierke, 2014) consisted of dice with different dot markings. The subjects saw one unfolded die of which all faces could be seen and five different dice of which three faces could be seen. The task was then to decide which of the five dice was unfolded.

The Test of Knowledge in Physics (Cronbach's α = .78; Zierke, 2014) evaluated the subjects' physics and technical knowledge, which mainly covered scholastic knowledge. The areas included technical systems, mechanics, electronics, thermodynamics, hydraulics, and aerodynamics.

The Mechanical Comprehension Test (Cronbach's α = .78; Zierke, 2014) contained questions covering technical problems which, in most cases, were illustrated by pictures. The test did not assess textbook knowledge but rather measured the understanding of mechanical and physical principles and devices.

The “Monitoring and Instrument Coordination” test provided two measures, one for psychomotor coordination and one for multiple task coordination (r_{tt} = .75 and .76 respectively; Hoermann, 2016). The test was a complex, flight simulator like task mainly operated with a joystick. Different parameters (heading, speed, altitude) had to be coordinated via tracking tasks and button presses. In addition, an auditory secondary task had to be accomplished. The two measures were obtained from different parts of the test: the first represented psychomotor measures of the tracking performance and the second represented the coordination of a complex task consisting of several tracking tasks and the auditory task.

4.3. Training and test preparation

At least three weeks ahead of the computer-based assessment, all applicants were offered online access to the test preparation materials. Here, the computer-based training modules —which are executable software files—could be downloaded for offline use. The download package also included two documents: one contained training recommendations and the other was a template which was be used to record how often each training module was used. With these software modules and documents each candidate had the opportunity and was advised to practice for the ability tests prior to the examination. The training records were submitted on the testing day.

The preparation concept differed between test domains. For basic cognitive abilities like visual perception speed, selective attention, auditory and visual memory capacity, mental rotation, and spatial visualization, the design of each training module resembled the actual test as closely as possible, yet without disclosing any test items. After instructions concerning the test principles, the user chose between up to three levels of difficulty, whereby the highest level equaled the item difficulty in the actual test. To ensure variation of training contents between repeated executions, training items were either randomly drawn at runtime from a larger pool of items or dynamically generated, thereby following specific rules to control item

difficulty. In the training program for the Running Memory Span Test for instance, the series of numbers that make up an item are randomly generated according to predefined sequence length specifications and restrictions on the frequency of occurrence and sequence of individual digits (e.g., to prevent the direct succession of identical digits). At the end of each training run the candidates received a summarized feedback about the percentage of correctly solved items. Recommendations were to start the training on the lowest level of difficulty, then work through all other levels. The same module should not be repeated more than three times in a row and regular breaks should be introduced. To keep the training varied, training modules for different abilities should be used alternately. In total each of them should be worked on 20 times. If candidates had the impression that their performance was still increasing, they were advised that they should practice even more.

In contrast, the software module for tests of technical knowledge and comprehension was not developed as a training tool. It merely offered example items for familiarization with the test principle and the use as a self-diagnostic tool to identify personal deficiencies in the subject matter. A repetition of the module was recommended only to check the individual learning progress. Hence, the number of executions of the module could not be interpreted as an indicator of training quantity. Finally, as the measurement of psychomotor coordination and the capacity for multiple task coordination requires specific input devices, a different preparation strategy was chosen. Instead of a software module, the candidates received a comprehensive information brochure and the actual training took place on the testing day during a standardized, yet unscored familiarization phase directly before the test.

4.4. Statistical analyses

Test reporting and interpretation for the selection procedure is done on a 9-point standard scale (stanine). Applicants' raw scores of each ability test are converted to stanine

values based on an archival data set of pilot trainee applicants. These individual stanine scores were used in the statistical analyses.

In a first step, we analyzed the relationship between the scores of the ability measurements and the reported number of training runs. In a second step, we assessed whether increasing numbers of training runs had an impact on the psychometric properties of ability tests and the relationship between constructs. Our hypotheses 1, 2a, and 2b implied that with an increasing number of training runs, the relationship between the measured scores and the underlying (latent) constructs as well as the relationship between constructs remain unchanged. Such hypotheses can be tested by examining the measurement and structural invariance of ability scores across different levels of training (Putnick & Bornstein, 2016; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). As a first step the sample was divided in six groups, based on the mean number of training runs across the six cognitive abilities: visual perception speed, selective attention, auditory and visual memory capacity, mental rotation and spatial visualization (see Table 1). Two of these groups were defined based on conceptual reasons: the first group contained applicants who had not used any of the training modules. The last group contained applicants who had used the training modules on average more often than recommended, i.e. more than 20 runs. The other four groups included applicants who were situated between these two ends of the training distribution (with equal ranges in training quantity across groups and reasonable group sizes).

Table 1

Mean number of training runs and sample sizes of training groups

Group No.	Training quantity (X)	*N*
1	$X = 0$	60
2	$0 < X \leq 5$	2061
3	$5 < X \leq 10$	3140
4	$10 < X \leq 15$	3578
5	$15 < X \leq 20$	4393
6	$X > 20$	2520

These six groups were simultaneously analyzed in a multigroup structural equation model (Putnick & Bornstein, 2016). Cognitive ability, technical comprehension, and complex psychomotor coordination were conceptualized as latent variables and the test scores (stanine scores) of the computer-based tests were conceptualized as manifest variables (see Figure 1).

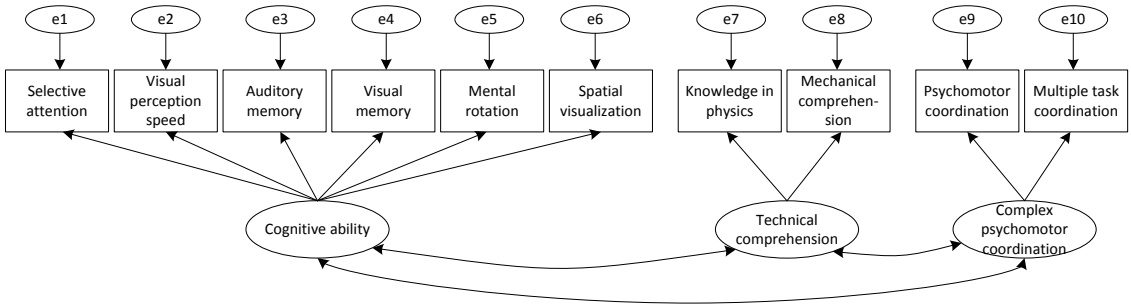

Figure 1. Confirmatory factor analysis model.
[2-column fitting image]

The assessment of measurement and structural invariance involved a sequence of model comparisons with increasingly stringent models (Putnick & Bornstein, 2016; see Table 2). The first model is the configural invariance model. This is the least stringent model since it imposes no parameter constraints across groups. With such a model the hypothesis is tested whether the same tests load on the same latent variables across groups. Within each group, the factor loading of one manifest variable was set to 1 to achieve model identification. The second model was the metric invariance model. Here, each factor loading was constrained to be equal across groups. This model allowed to test whether the amount of computer-based training has an effect on the relative weights of the constructs on the measured test scores. The third model was the scalar invariance model. In this model the loadings and the intercepts of the measured variables were constrained to be equal across groups, which means that an individual with average ability would have the same expected manifest score on a test, irrespective of how much they trained. The fourth model was the residual invariance model. Here, the loadings, intercepts, and residual variances were constrained to be equal across

groups, which states that the amount of variance in manifest indicators not explained by the latent variables remains the same, irrespective of how much an individual has trained. The fifth model tests for invariant factor variances; here the variances of the latent factors are additionally constrained to be equal across groups implying that irrespective of the amount of training the variability of the latent ability remains the same. The sixth model was the structural invariance model. In this model, the loadings, intercepts, residual variances, factor variances, and the covariances between factors were constrained to be equal across groups. If the structural invariance model can be accepted, it implies that in addition the relations between latent constructs remain constant across training groups. In this case, the whole factorial structure of the measurements is not affected by training (hypothesis 1).

Table 2

Overview of equality constraints imposed across training groups

No.	Model	Factor loadings	Intercepts	Residual variances	Factor variances	Factor covariances
1	Configural invariance	free	free	free	free	free
2	Metric invariance	<u>invariant</u>	free	free	free	free
3	Scalar invariance	invariant	<u>invariant</u>	free	free	free
4	Residual invariance	invariant	invariant	<u>invariant</u>	free	free
5	Invariant factor variances	invariant	invariant	invariant	<u>invariant</u>	free
6	Structural invariance	invariant	invariant	invariant	invariant	<u>invariant</u>

Note. All models are nested; the parameter constraint that is tested in a model is underlined.

With reference to the Ackerman model, it was hypothesized that across different training levels, there should be a constant relationship between the ability constructs and perceptual speed (hypothesis 2a) and between the ability constructs and psychomotor coordination (hypothesis 2b). The measurement invariance analyses already serve as omnibus tests for both hypotheses. In particular, if metric invariance is supported, there is evidence of a common pattern of loadings across groups. This also comprises the loading of perceptual speed on the latent cognitive ability factor, i.e. irrespective of the amount of training, the

relationship between the latent variable and its indicators is not changed. To more specifically test hypothesis 2a, we compared the metric invariance model to a modified model where only the loading of perceptual speed on the latent cognitive ability factor was released. If there is no significant difference between both models, this indicates that the importance of perceptual speed was not changed due to different amounts of training: the size of the loading remained the same independent of training (Byrne, 2016). A similar analysis can be performed for psychomotor coordination (hypothesis 2b). Again, the metric invariance model can be compared to a modified model, in this case, psychomotor coordination was allowed to load on the latent cognitive ability factor. If there is no significant difference between the metric invariance and this modified model, it more specifically indicates that the importance of psychomotor coordination was not changed as a result of different amounts of training.

The nested structural equation models were analyzed using the statistics environment R (version 3.4.3; R Core Team, 2017) together with the "lavaan" package (version 0.5-23; Rosseel, 2012). Maximum likelihood was employed to estimate the model parameters. Model fit of the *baseline/configural model* was assessed by a set of approximate fit indices: comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). The configural model was considered acceptable with CFI $\geq .95$, RMSEA $< .08$, and SRMR $< .08$ (Byrne & Stewart, 2006). Chi-square was not used as an acceptance criterion since chi-square is overly sensitive to sample size (Bentler & Bonett, 1980; Cheung & Rensvold, 2002) and chi-square is not considered as an appropriate criterion in very large samples (Rutkowski & Svetina, 2013).

When chi-square is used as a criterion in large samples such as the present study, models tend to be rejected even due to differences of trivial size. However, it is not enough to claim that the chi-square value is inflated by sample size (Ropovik, 2015), when it is possible to empirically analyze the impact of sample size on the chi-square statistic. We therefore

estimated the impact of sample size on chi-square in the present study: we repeated the analysis with the baseline model in a subset of our sample. The subsample was generated by randomly drawing 10% of applicants out of each group (with replacement) resulting in sample sizes around $N = 300$ for each stratum, which is the usual recommendation to perform such analyses. Only for the first group, the complete sample was employed because of the small sample size. To minimize the impact of sampling error, we conducted the sampling procedure 1,000 times, analyzed the baseline model each time and computed the mean of the 1,000 chi-square estimates. A comparison of this mean chi-square value with the chi-square value in the complete sample allows an estimation of the impact of sample size on chi-square.

All models were nested models with increasing constraints. Therefore, these *constrained models* were assessed using a direct model comparison to the less restricted model. A model was determined to be invariant when the difference in the comparative fit index (ΔCFI) was smaller or equal to .01 (Cheung & Rensvold, 2002; Kim, Cao, Wang, & Nguyen, 2017).

5. Results

5.1. Descriptive analyses

Figure 2 shows how often the training modules for the cognitive ability tests were actually used. Aggregated across the six cognitive domains, 10% of all candidates reported a mean training level of 20 runs, thereby following the recommendation exactly. In contrast, 75% reported a lower level of training and 15% a higher one. As Table 3 illustrates, most modules were practiced 14-15 times on average. The module for selective attention was used less often and the module for visual perception speed was used more often.

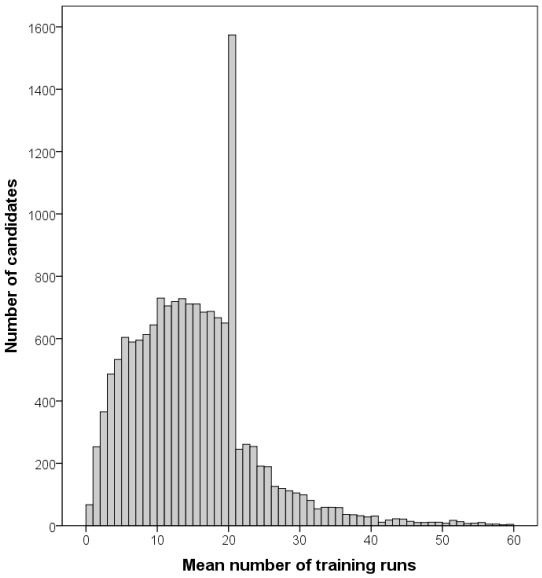

Figure 2. Frequency distribution of the mean number of computer-based training runs across 6 ability measures, for display purposes rounded to integer values. *N* = 15,752.
[single column fitting image]

Table 3

Descriptive statistics for the number of computer-based training runs and performance in ability tests

	Training runs	Test performance	
	M (*SD*)	*M* (*SD*)	$r_{\text{Training,Test}}$
Selective attention	12.68 (8.49)	5.18 (1.79)	.32
Visual perception speed	17.16 (11.64)	5.22 (1.87)	.50
Auditory memory	14.73 (10.16)	5.33 (1.96)	.37
Visual memory	14.04 (9.96)	5.18 (2.07)	.36
Mental rotation	14.34 (9.93)	5.65 (2.22)	.37
Spatial visualization	14.30 (9.51)	4.91 (1.85)	.39
Knowledge in physics	-	4.90 (2.13)	-
Mechanical comprehension	-	4.63 (1.97)	-
Psychomotor coordination	-	4.72 (2.00)	-
Multiple task coordination	-	4.96 (1.91)	-

Note. For each measure the mean number of training runs and the mean stanine score of test performance is shown, together with the Spearman correlation of both variables. For the latter 4 measures no data for training quantity were available due to a different preparation concept. Due to the large sample size, indications of statistical significance for correlation coefficients are omitted. *N* = 15,752.

Table 3 also shows the performance data for all computer-based tests in stanines with *M* = 5 and *SD* = 2. The mean test scores and the respective standard deviations measured in

our sample were comparable to the normative group, with slightly better performance than the norm average of $M = 5$ in measures where training modules were offered in advance, and a slightly lower performance than the norm in the others.

For the six measures of cognitive abilities, the relationship between the number of training runs and test performance was analyzed. As a visual inspection indicated that it was a non-linear, monotonic relationship, we performed the analyses using Spearman's rank order correlation. In contrast to the Pearson product-moment correlation, Spearman's coefficient evaluates the monotonic relation between two variables and not the linear. As Table 3 shows, there were substantial correlations between training runs and test performance. The correlation coefficients ranged from r = .32 (selective attention) to r = .50 (visual perception speed), with a mean correlation of $M = .39$ (the mean correlation was computed by transforming the correlation coefficients to Fisher's Z values and back-transforming the mean Z value). The more often a training module had been practiced on average, the higher was the achieved test result. It must be noted, however, that higher average training scores were also associated with higher standard deviations, so that the pattern of correlations between training and test performance may be mediated by the variability in training runs. Figure 3 further illustrates the relationship between the number of training runs and test performance. When aggregated across the six cognitive ability measures, there was a non-linear relationship between the number of training runs and test performance. Generally, more training predicted on average better performance. However, the training gains became smaller with additional training amounts. This pattern was also present when data were analyzed for each measure separately.

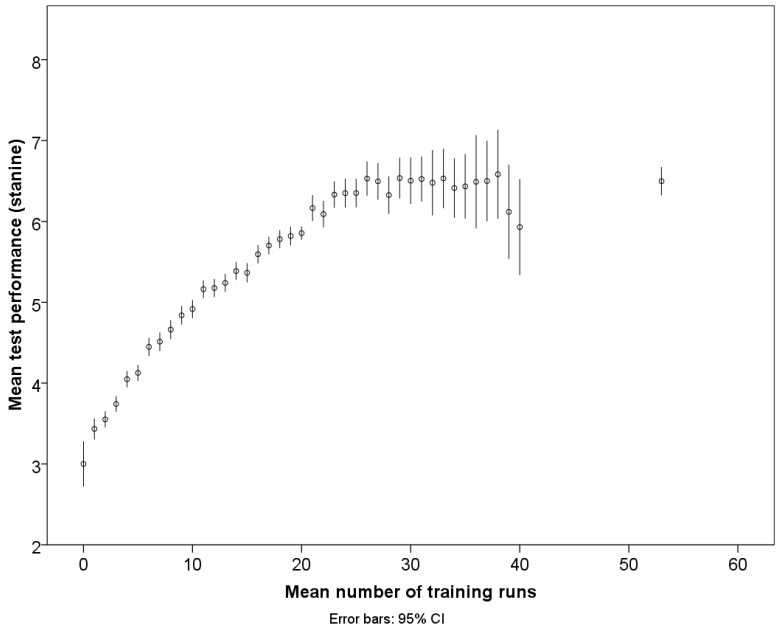

Figure 3. Relation between the mean number of computer-based training runs and mean test performance. X axis: mean number of training runs across 6 ability measures, for display purposes rounded to integer values; y axis: mean stanine scores aggregated across 6 ability tests, corresponding to the 6 training modules. Error bars indicate 95% confidence intervals. Due to small sample sizes with more than 40 training runs and increasingly larger confidence intervals, applicants with on average more than 40 training runs have been aggregated (only for display purposes; n = 262, M = 53 training runs). N = 15,752; the data point with the smallest sample size is at 39 training runs with n = 28.
[1-column fitting image]

5.2. Confirmatory factor analyses

The ability scores were tested across different levels of training for measurement and structural invariance. The sequence of increasingly stringent models started with the configural invariance model which imposes no parameter constraints across groups. As Table 4 shows, the fit of the configural invariance model was acceptable, $\chi^2(df = 192) = 2{,}396.23$, $p < .001$, CFI = .97, RMSEA = .07, SRMR = .03. Imposing restrictions between groups also resulted in acceptable model fit statistics and also the CFI differences between models were acceptable: the metric invariance model (ΔCFI = .001), scalar invariance model (ΔCFI = .003), residual invariance model (ΔCFI = .003), invariant factor variances model (ΔCFI = .004), and the structural invariance model (ΔCFI = .001). Thus, the factorial structure of the

ability measurements remained constant across different levels of computer-based training, supporting hypothesis 1.

Table 4

Goodness of fit for models of invariance across different levels of computer-based training

Model	χ^2 (*df*)	CFI	RMSEA (90% CI)	SRMR	Model compa-rison	ΔCFI	Decision
Configural invariance	2,396.23 (192)	.970	.066 (.064-.069)	.030	---	---	Accept
Metric invariance	2,552.27 (227)	.969	.062 (.060-.065)	.035	configural	.001	Accept
Scalar invariance	2,811.60 (262)	.966	.061 (.059-.063)	.037	metric	.003	Accept
Residual invariance	3,019.40 (312)	.963	.057 (.056-.059)	.038	scalar	.003	Accept
Invariant factor variances	3,394.23 (327)	.959	.060 (.058-.062)	.087	residual	.004	Accept
Structural invariance	3,438.54 (342)	.958	.059 (.057-.061)	.087	factor	.001	Accept

Note. Acceptance criterion for nested models: $\Delta CFI \leq .01$ (Cheung & Rensvold, 2002). CFI = comparative fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual. N = 15,752 (see Table 1 for individual group sizes).

As outlined in the Methods section, a significant chi-square value is not a reliable indicator of model fit in very large samples. To estimate the impact of sample size on chi-square in the present study, we repeated the analysis with the finally accepted model (i.e., the structural invariance model) in subsets of our sample. Mean chi-square in the subsamples was $\chi^2(df = 342, N = 1{,}629) = 691.29$. While still statistically significant, there was a large drop in the value of chi-square, indicating that it was considerably influenced by sample size.

Since measurement invariance across the different amounts of training was supported, this also supported the hypotheses of a constant relationship between the ability constructs and perceptual speed (hypothesis 2a) and between the ability constructs and psychomotor coordination (hypothesis 2b). To specifically assess the importance of perceptual speed, a modified version of the metric invariance model was analyzed where only the loadings of

perceptual speed on the latent cognitive ability factor was released, $\chi^2(df = 222) = 2{,}542.31$, $p < .001$, CFI = .97, RMSEA = .06, SRMR = .04. However, a direct model comparison with the metric invariance model showed no significant difference between the two models, $\Delta\chi^2(df = 5, N = 15{,}752) = 9.96$, $p = .076$. This implies that the liberalized, less parsimonious model, allowing different loadings between groups, did not result in a significantly better fit. It can be concluded that the importance of perceptual speed remained constant across different amounts of training. Hypothesis 2a was supported. To specifically evaluate the importance of psychomotor coordination, again a modified version of the metric invariance model was analyzed. In this case psychomotor coordination was allowed to load on the latent cognitive ability factor, $\chi^2(df = 221) = 2{,}549.09$, $p < .001$, CFI = .97, RMSEA = .06, SRMR = .04. Again, a direct model comparison showed no significant difference between the two models, $\Delta\chi^2(df = 6, N = 15{,}752) = 3.18$, $p = .786$, suggesting that the importance of psychomotor coordination remained constant across different training quantities, supporting hypothesis 2b.

6. Discussion

With the growing use of computer-based diagnostics, especially in applied contexts, the question of good and fair preparation for applicants has become increasingly crucial. Preparation for ability tests has become an industry in many selection contexts and the financial resources of the applicant and differences in quantity and quality of test preparation constitute growing sources of variance and unfairness. Our approach to cope with this challenge was to offer freely available computer-based training modules to all applicants. The present study showed that there were training effects consistent with previously published research, but most importantly, the structure of the tests was not altered by offering free training to all applicants. The analyses for measurement and structural invariance indicated that there was no change in the factorial structure of the ability measures. With regard to our specific hypotheses, the invariance analysis showed that increasing levels of training did not

alter the importance of perceptual speed and psychomotor coordination within the factorial structure. With respect to Ackermann's theory (1988), this finding is interpreted as evidence for the fact that training did not alter the measured constructs from rather broad abilities to task-specific skills. This can be attributed to the way the ability tests of the present study were designed: they were characterized by an inconsistent structure, i.e. the rules for item processing changed constantly during the test. This inconsistency impeded the process of automation and required continuous active information processing. Therefore, even after intense practice, the content abilities (Ackerman, 1988) remained the main factor for solving the test items and hence, the structural validity of the tests was not affected.

Previous studies have already analyzed the measurement invariance of cognitive ability tests in the context of practice and retest effects. However, as a result of different criteria for defining the groups, these are not directly comparable to the present study. For example, previous studies analyzed invariance across classes of training prior to admission testing (Arendasy et al., 2016), practice with different cognitive tasks between repeated intelligence measurements (Estrada, Ferrer, Abad, Román, & Colom, 2015), different methods used to construct alternate test forms (Arendasy & Sommer, 2013), or authors analyzed invariance across different measurement occasions (Arendasy & Sommer, 2017; Freund & Holling, 2011; Lievens et al., 2007; Reeve & Lam, 2005; Sommer, Arendasy, & Schützhofer, 2017). To our knowledge, the present study is the first that analyzes measurement and structural invariance across different amounts of test-specific training. As most of the previous invariance studies, our invariance analyses were conducted within a multi-group framework. Recently, moderated nonlinear factor analysis was proposed as a more general approach (Bauer, 2017): it allows investigating measurement invariance also for continuous variables which could be used in future research.

Additionally, we found that although there were instructions given for computer-based training in the present study, applicants did not always follow them, resulting in a large variability in the use of the training modules. The reasons for the large differences in training quantity may be diverse: individual differences in motivation, time resources, beliefs about the effectiveness of training or correctness of self-evaluations (for relevant factors in commercial coaching programmes, see Ryan, Ployhart, Greguras, & Schmit, 1998). Therefore, when psychologists decide to offer computer-based training to all applicants, they must decide how to cope with these possible influences. One strategy could be to allow applicants to participate in the examination only if they confirm having trained up to a specified level. This could enhance construct validity since the measurements are less biased by differences in training and hence, potential differences in motivation, time resources, or correctness of self-evaluations. However, it should also be considered that such differences may provide diagnostically relevant information. Moreover, to the extent that these differences meet requirements for the later job, the criterion-related validity of measurements will be enhanced if applicants are allowed to participate in the examination irrespective of their actual training level. In this case, it is mandatory to offer the training software to all applicants, with sufficient time for an adequate test preparation. In addition, institutions have to underline the importance of training to all applicants, otherwise the risk that applicants only fail the examination due to insufficient training will be unacceptably large. Such false negative decisions are probable since in most cases the introduction of training software necessitates the adjustment of test norms—which become stricter as a result of the training effects. In particular, if the number of potentially suitable applicants is rather small, the costs of false negative decisions are high. Taken together, there is some investment required with the introduction of computer-based training but it ensures that training and the resulting score gains are not restricted to applicants with high financial resources.

Consistent with studies on practice effects and learning curves, we found a positive, non-linear relationship between the quantity of computer-based training and cognitive performance (Bartels, Wegrzyn, Wiedl, Ackermann, & Ehrenreich, 2010; Donner & Hardy, 2015). Such a relation was also reported by studies that used the time spent on training and not the number of training runs (cf. Hausknecht et al., 2007), which is plausible since the time spent on training and the number of training runs should be highly correlated. As discussed in several lines of research, the relationship between training and cognitive performance is mainly based on familiarization with the item format and an increase in test-specific skills (for an overview of the causes of retest and practice effects see Randall & Villado, 2017; Van Iddekinge & Arnold, 2017). Recent evidence supported the hypothesis that score gains after repeated testing are due to refinements in the solution strategies of subjects (Arendasy & Sommer, 2017; Hayes, Petrov, & Sederberg, 2015). As Arendasy and Sommer (2017) point out, when subjects repeatedly conduct an ability test, information on speed and accuracy of the solution strategy and salient item design characteristics are stored in working memory. With further retests, the cognitive schema becomes more automatized and hence more working memory resources become available. Depending on the nature of the cognitive ability analyzed, such an automation process may proceed at different speeds, with for example simple speed tasks reaching the plateau faster than memory or reasoning tasks (Scharfen, Jansen, & Holling, in press; Scharfen et al., 2018).

A limitation of the present study is that the number of training runs was not experimentally manipulated and not directly monitored but only reported by each applicant. Therefore, it cannot be excluded that the training data were biased by record inaccuracies or socially desirable behavior. For example, we do not know whether the participants actually adhered to the suggestion or whether the spike after run 20 in figure 3 could be due to social desirability. Furthermore, it is possible that some applicants had used commercial preparation

methods in addition to our training software. However, the relationship between training data and test performance, which is in line with theoretical expectations based on Ebbinghaus' learning curves (cf. Figure 1) suggest that such influences are—at least on average—of only minor importance. Moreover, the training guidance with recommendations and the record template, together with the upcoming high-stakes setting, are likely to have increased the reliability of the training records. It would be desirable to more closely monitor the training activities of the applicants, for example by tracking the usage of the software. However, in this case the monitoring process should be made transparent to the user to ensure data protection and to prevent negative reactions such as reactance or enhanced socially desirable behavior (cf. Ketelaar & van Balen, 2018).

Particularly in times of internet testing and the growing (online) availability of information about tests, to be reliable and fair, psychological diagnostics has to react. Our results suggest that free computer-based training is a possible way to react to today’s challenges without affecting the measurement quality in diagnostic decision making and tests for measurement invariance are an efficient way to screen for such potential biases. However, offering training software to all applicants is certainly not the optimal strategy in every applied context. For example, if sufficient training cannot be implemented due to insufficient individual prerequisites (such as in neuropsychology or psychiatry) or contextual conditions (e.g., short time lags between registration and assessment), this training concept is not appropriate. Similarly, if the ability tests have a rather simple structure which makes them vulnerable to skill development after practice, other strategies may be more appropriate. In these cases, the application of alternative ability tests and test versions should be preferred. Alternatively, the test-retest interval may be extended (Scharfen et al., in press). For personnel selection contexts, however, offering computer-based training to all subjects offers a powerful tool to ensure sufficient fairness of the diagnostic process.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3-27. doi:10.1037/0033-2909.102.1.3

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117,* 288-318. doi:10.1037/0096-3445.117.3.288

Ackerman, P. L., & Schneider, W. (1985). Individual differences in automatic and controlled information processing. In R. F. Dillon & R. F. Schmeck (Eds.), *Individual differences in cognition* (pp. 35-66). Orlando: Academic Press.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arendasy, M. E., & Sommer, M. (2013). Quantitative differences in retest effects across different methods used to construct alternate test forms. *Intelligence, 41,* 181-192. doi:10.1016/j.intell.2013.02.004

Arendasy, M. E., & Sommer, M. (2017). Reducing the effect size of the retest effect: Examining different approaches. *Intelligence, 62,* 89-98. doi:10.1016/j.intell.2017.03.003

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence, 55,* 44-56. doi:10.1016/j.intell.2016.01.004

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience, 11,* 118. doi:10.1186/1471-2202-11-118

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychol Methods, 22,* 507-526. doi:10.1037/met0000077

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606. doi:10.1037/0033-2909.88.3.588

Byrne, B. M. (2016). *Structural equation modeling with amos : Basic concepts, applications, and programming* (Third edition. ed.). New York: Routledge, Taylor & Francis Group.

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The macs approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling: A Multidisciplinary Journal, 13*, 287-321. doi:10.1207/s15328007sem1302_7

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. doi:10.1207/s15328007sem0902_5

Chung-Herrera, B. G., Ehrhart, K. H., Ehrhart, M. G., Solamon, J., & Kilian, B. (2008). Can test preparation help to reduce the black—white test performance gap? *Journal of Management, 35*, 1207-1227. doi:10.1177/0149206308328506

Donner, Y., & Hardy, J. L. (2015). Piecewise power laws in individual learning curves. *Psychonomic Bulletin & Review, 22*, 1308-1319. doi:10.3758/s13423-015-0811-x

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence, 50*, 93-99. doi:10.1016/j.intell.2015.02.004

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243-285). New York: Academic Press.

Fleishman, E. A., & Hempel, W. E. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. *Psychometrika, 19*, 239-252. doi:10.1007/bf02289188

Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence, 39*, 233-243. doi:10.1016/j.intell.2011.02.009

Greiff, S., Scherer, R., & Kirschner, P. A. (2017). Some critical reflections on the special issue: Current innovations in computer-based assessments. *Computers in Human Behavior, 76*, 715-718. doi:10.1016/j.chb.2017.08.019

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385. doi:10.1037/0021-9010.92.2.373

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence, 48*, 1-14. doi:10.1016/j.intell.2014.10.005

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185-207. doi:10.3758/bf03212979

Hermes, M., & Stelling, D. (2016). Context matters, but how much? Latent state-trait analysis of cognitive ability assessments. *International Journal of Selection and Assessment, 24*, 285-295. doi:10.1111/ijsa.12147

Hoermann, H.-J. (2016). *MIC: Monitoring & instrument coordination – documentation*. Unpublished report. German Aerospace Center (DLR). Hamburg, Germany.

Ketelaar, P. E., & van Balen, M. (2018). The smartphone as your follower: The role of smartphone literacy in the relation between privacy concerns, attitude and behaviour towards phone-embedded tracking. *Computers in Human Behavior, 78*, 174-182. doi:10.1016/j.chb.2017.09.034

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 524-544. doi:10.1080/10705511.2017.1304822

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981-1007. doi:10.1111/j.1744-6570.2005.00713.x

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92*, 1672-1682. doi:10.1037/0021-9010.92.6.1672

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the sat. *Psychological Bulletin, 89*, 191-216. doi:10.1037/0033-2909.89.2.191

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale: Lawrence Erlbaum.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71-90. doi:10.1016/j.dr.2016.06.004

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Randall, J. G., & Villado, A. J. (2017). Take two: Sources and deterrents of score change in employment retesting. *Human Resource Management Review, 27*, 536-553. doi:10.1016/j.hrmr.2016.10.002

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence, 33*, 535-549. doi:10.1016/j.intell.2005.05.003

Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology, 6*, 1715. doi:10.3389/fpsyg.2015.01715

Rosenbaum, D. A., Carlson, R. A., & Gilmore, R. O. (2001). Acquisition of intellectual and perceptual-motor skills. *Annual Review of Psychology, 52*, 453-470. doi:10.1146/annurev.psych.52.1.453

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*. doi:10.18637/jss.v048.i02

Rutkowski, L., & Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*, 31-57. doi:10.1177/0013164413498257

Ryan, A. M., Ployhart, R. E., Greguras, G. J., & Schmit, M. J. (1998). Test preparation programs in selection contexts: Self-selection and program effectiveness. *Personnel Psychology, 51*, 599-621. doi:10.1111/j.1744-6570.1998.tb00253.x

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215-227. doi:10.1037/0003-066X.63.4.215

Sackett, P. R., Burris, L. R., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145–183). Oxford, UK: Wiley.

Scharfen, J., Jansen, K., & Holling, H. (in press). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin & Review*.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence, 67*, 44-66. doi:10.1016/j.intell.2018.01.003

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222. doi:10.1016/j.hrmr.2008.03.003

Sommer, M., Arendasy, M. E., & Schützhofer, B. (2017). Psychometric costs of retaking driving-related cognitive ability tests. *Transportation Research Part F: Traffic Psychology and Behaviour, 44*, 105-119. doi:10.1016/j.trf.2016.10.014

Stemig, M. S., Sackett, P. R., & Lievens, F. (2015). Effects of organizationally endorsed coaching on performance and validity of situational judgment tests. *International Journal of Selection and Assessment, 23*, 174-181. doi:10.1111/ijsa.12105

The International Test Commission. (2006). International guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*, 143-171. doi:10.1207/s15327574ijt0602_4

Tippins, N. T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 551-582. doi:10.1146/annurev-orgpsych-031413-091317

Van Iddekinge, C. H., & Arnold, J. D. (2017). Retaking employment tests: What we know and what we still need to know. *Annual Review of Organizational Psychology and Organizational Behavior, 4*, 445-471. doi:10.1146/annurev-orgpsych-032516-113349

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. doi:10.1177/109442810031002

Zierke, O. (2014). Predictive validity of knowledge tests for pilot training outcome. *Aviation Psychology and Applied Human Factors, 4*, 98-105. doi:10.1027/2192-0923/a000061

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York: Routledge Falmer.