A Validation Study of the German Complex-Span Tasks and Some General Considerations on Task Translation Procedures in Cognitive Psychology

Jan Rummel,[1] Lena Steindorf,[1] Ivan Marevic,[1] and Daniel Danner[2]

[1]Department of Psychology, Heidelberg University, Germany
[2]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

Abstract: Automated complex-span tasks are widely used to assess working-memory capacity and the English versions show good psychometric properties (Unsworth, Heitz, Schrock, & Engle, 2005). However, it is generally an open question whether translated task versions have the same properties as the original versions and whether results obtained with translated tasks can be interpreted equivalently to those obtained with the original tasks. We translated the complex-span tasks and had a sample of German participants perform these tasks as well as a running-memory-span task and a reasoning test. We assessed the reliabilities of the German complex-span tasks and their construct and criterion-related validities. Extrapolating from cross-cultural literature, we also employed a test of measurement invariance to compare the correlational patterns as well as the construct structure between the German sample and a similar North-American sample. Results show that the German complex-span tasks are reliable and valid indicators of working-memory capacity and that they are metrically and functionally equivalent to the original versions. As measurement equivalence is an important but often neglected topic in basic cognitive psychology, we also highlight the general benefits of using equivalence tests when translating cognitive tasks.

Keywords: working-memory capacity, German complex-span tasks, task equivalence, Bayesian measurement invariance tests

In cognitive psychology – and also in several other psychological disciplines, such as clinical, developmental, or educational psychology – working memory is a construct of central interest, as individual differences in working-memory capacity (WMC) relate to performance differences in various other areas (cf. Barrett, Tugade, & Engle, 2004). Standard instruments for the assessment of WMC are the automated complex-span tasks developed by Engle, Unsworth, and colleagues (Redick et al., 2012; Unsworth, Heitz, Schrock, & Engle, 2005; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). In this article, we report a validation study of German versions of the complex-span tasks, in which we tested the tasks' reliabilities, validities, and their equivalence to the original English versions. Additionally, we discuss the general merits of employing measurement invariance tests when translating cognitive tasks – a method regularly used by cross-cultural psychologists when testing for equivalence between different language versions of psychological assessment tools but rarely used by cognitive psychologists, who are less interested in cross-language comparisons. We will further elaborate this issue after quickly introducing the WMC tests whose German versions we validated in the present study.

Assessing Working-Memory Capacity

Working-memory capacity (WMC) tests are among the most frequently used tools for the objective assessment of cognitive abilities in psychology (Conway et al., 2005). A number of different tasks have been suggested for WMC assessment, such as *n*-back, recall *n*-back, pattern transformation, memory updating, simple-span, and complex-span tasks (see Oberauer, Süss, Schulze, Wilhelm, & Wittmann, 2000). Because working memory is considered a complex system comprising numerous processes that allow for maintaining, accessing, manipulating, updating, and coordinating information in active memory (Miyake & Friedman, 2012), it seems unlikely that there is a single "true" indicator of WMC. Nevertheless, in recent years, complex-span tasks have been most frequently used to assess WMC. Complex-span tasks require participants to store information while performing an intervening processing task and to recall the to-be-stored information in the correct order in a later test. Storage performance in complex-span tasks has been shown to be highly correlated and to account for similar variance in cognitive ability tests – suggesting that they measure a unitary construct (i.e., WMC; Kane et al., 2004). These tasks are frequently utilized for WMC assessment in English-speaking countries (Redick et al., 2012; Unsworth et al., 2005, 2009), which is probably due to their excellent psychometric properties, that is, their high reliability (Redick et al., 2012) and construct validity, indicated by their high correlation with other WMC measures (e.g., the running-memory-span task; Harrison, Shipstead, & Engle, 2015) as well as with measures of fluid intelligence (e.g., the Raven matrices test; Conway, Kane, & Engle, 2003; Unsworth, 2010). Complex-span tasks have been translated into several other languages. But do translated versions show similarly satisfying psychometric properties and can they be used in a similar manner as the original versions? We are only aware of three validation studies – one for a Dutch version of the operation-span task (De Neys, D'Ydewalle, Schaeken, & Vos, 2002), another for a French version of the reading-span task (Delaloye, Ludwig, Borella, Chicherio, & de Ribaupierre, 2008), and the most recent for French shortened versions of the reading-span, symmetry-span, and operation-span tasks (Gonthier, Thomassin, & Roulin, 2016). These studies indicate that the respective translated tasks were reliable and valid. To our best knowledge, however, our study is the first to test the reliability and validity of German translations of the complex-span tasks. Additionally, unlike previous validation studies, we tested for measurement invariance with the original English versions. Measurement invariance tests are regularly applied to secure equivalence between different to-be-compared language versions of psychological assessment tools (Van de Vijver & Leung, 2011); but they are rarely applied when cognitive tasks are translated for usage in a language area which differs from the one of the original task version. Before we turn to the validation study, we will outline why we consider measurement equivalence testing important for cognitive task translation procedures.

Testing for Equivalence of Psychological Assessment Tools

Psychological instruments are often translated to be used analogously to the original instruments in language areas other than the one they were originally developed for (Gudmundsson, 2009). It has been pointed out in psychological assessment literature that the reliability and validity of translated assessment tools have to be tested anew and cannot be inferred from the psychometric properties of the original versions (Peña, 2007). Some researchers – and especially those involved in the International Test Commission (ITC) – further developed guidelines for proper language adaptation (Hambleton, 2001; Muniz, Elosua, & Hambleton, 2013; Van de Vijver & Hambleton, 1996) which comprise a set of a priori measures ensuring a good quality of the translated versions, such as recommendations regarding the translation process (see Harkness, 2003), as well as measures ensuring a posteriori that two language versions of an instrument produce comparable outputs.

Achieving equivalent language versions is especially relevant when comparing national or cultural groups, because group-dependent biases at the item, method, and construct level can render test scores obtained in different groups incomparable resulting in incorrect conclusions about national or cultural group differences (Van de Vijver & Leung, 2011). Consequently, several psychometric tools, including some cognitive tasks, have been developed or adapted for cross-national or cross-cultural study purposes, for example, for comparing students' cognitive abilities between different nations as it is done in the Program for International Student Assessment (e.g., Kankaras & Moors, 2014) or for the comparison of cognitive development between Western and African cultures (e.g., Grigorenko et al., 2007; Zuilkowski, McCoy, Serpell, Matafwali, & Fink, 2016).

However, whereas researchers who are directly concerned with cross-national or cross-cultural research questions tend to consider translation-immanent threats to task equivalence, researchers who simply intend to use translated task versions in their own language area are less aware of and/or concerned about these threats. When translated cognitive tasks are used to derive performance indicators (e.g., mean accuracy rates or response times) for an underlying construct that are then compared between (quasi-) experimental groups or correlated with other constructs within one language area, one might typically not think of (and often also not expect) nonequivalence between the original task and its translation. That is, at first glance, equivalence between different language versions of cognitive tasks seems to be of minor importance for this kind of research. However, because cognition researchers from different language areas intend to investigate the same latent psychological constructs with their respective language versions of a cognitive task, we argue that task equivalence is of critical importance.

Nowadays, more and more cognitive psychologists from different language areas publish their research preferably in English, often using North-American psychology journals as outlets (Piocuda, Smyers, Knyshev, Harris, & Rai, 2015). Thus, all researchers contribute to a shared knowledge base and, consequently, global cognitive psychology science (implicitly) relies on the assumption that basic findings of scientific value – for instance regarding basic cognitive principles – generalize across language areas. We argue that, for this reason, certain levels of equivalence between different language versions of cognitive tasks are warranted and should thus be tested for, namely, *metric equivalence* of the test scores that are derived from these tasks and that serve as indicators for a construct of interest as well as *functional equivalence* of the constructs themselves. According to Van de Vijver and Leung (2011), metric equivalence requires the relations of test scores within each to-be-compared (language) group to be comparable across groups. Functional equivalence requires the construct assessed with the tasks to be embedded within its nomological network across groups in a similar fashion. A lack of metric and/or functional equivalence will render findings (e.g., group comparisons or correlational patterns) obtained with one task version incompatible with the findings obtained with another version. Such incompatibility will likely cause confusion in the respective literature as well as the (wrong) impression that established findings are not reliable, and thus finally harm scientific progress.

At first sight, nonequivalence across different language versions of cognitive tasks may seem unlikely because many of these tasks comprise nonverbal test items that can thus be used in different language versions without modification. However, these tasks regularly rely on verbal instructions. Although it may often be sufficient for participants to understand the gist of what to do when receiving instructions for a cognitive task, translation-immanent variations in instructions can cause biases (He & Van de Vijver, 2012). For instance, if difficulties with understanding the instructions were more likely to occur in the translated than in the original task version, the average task-score reliability of the translated version would be reduced. Along these lines, subtle wording differences between original and translated task versions may alter the strategy with which participants approach the tasks or the level of motivation participants devote to the tasks. Of course, task equivalence is even more likely to be jeopardized when test items themselves are verbal (cf. Geisinger, 1994; Gudmundsson, 2009).

These considerations become especially relevant in the light of recent initiatives to conduct multi-lab replication studies (see Simons, Holcombe, & Spellman, 2014). In these initiatives, several researchers, often from different countries, run exact replications of one experiment to test the reliability and robustness of its outcome. Nonequivalence between measurement tools used in the original study and by the participating researchers is a serious threat to such replication attempts. We will further discuss implications of nonequivalence for replication science in the Discussion section.

In sum, we argue that a lack of metric and/or functional equivalence between different language versions of cognitive tasks – and especially between those that are standard instruments for the assessment of central cognitive constructs, like the WMC tasks validated here – is a potential caveat for basic cognitive science, especially since it has become more global. We therefore suggest, in line with recommendations for cross-cultural studies (Van de Vijver & Leung, 2011), to test not only for reliability and validity but also for measurement equivalence when translating cognitive tasks, even when tasks are used within one language area only. Such equivalence tests allow the required level of metric and functional equivalence between original and translated task versions to be secured a posteriori and without much additional effort. In the present study, we illustratively employed such tests, to ensure that the German complex-span tasks can be used and interpreted analogously to the original English versions.

Testing for Measurement Equivalence

Measurement invariance tests ensure that members from different groups who have the same standing on a construct will achieve the same test scores (Schmitt & Kuljanin, 2008). Measurement invariance tests as well as comparisons of reliabilities, in terms of the true-score to total-variance ratios, and convergent validities, in terms of factor variances and covariances, can be conducted as multigroup comparisons within the confirmatory factor analysis (CFA) framework (Byrne & Stewart, 2006). An illustrative CFA model is displayed in Figure 1. As the CFA framework can be used to establish both metric and functional equivalence it is especially suitable for the validation of translated tasks.

Vandenberg and Lance (2000) distinguish three levels of measurement invariance, namely configural, metric, and scalar invariance. To assess measurement invariance levels, the fits of the three measurement invariance models are often tested against each other (Cheung & Rensvold,

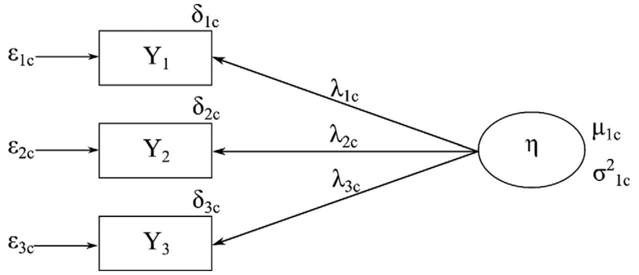

Figure 1. General structural equation model for the test of measurement invariance. Y_i represent items that are manifest indicators of any psychological construct; η represents the construct of interest; λ_i represent the factor loadings from the manifest on the latent variables; δ_i represent item intercepts; ε_i represent unique variances of the indicators; μ represents the construct mean; σ^2 represents the construct variance.

2002). However, these procedures test for *exact* invariance, that is, for any difference between groups – even those that are negligibly small and thus probably meaningless. Because of this very conservative assumption, exact measurement variance is hardly ever achieved in practice.

Alternatively, Vandenberg and Lance (2000) suggest partial measurement invariance tests that require only some parameters to be invariant across groups (or measures to be invariant across some groups). As partial invariance tests do not provide a solution for situations where only few groups and few measures are compared (as it is usually the case with validation studies), we will not discuss this approach in broader detail here.

Another alternative are *approximate* measurement invariance tests (e.g., van de Schoot et al., 2013). They use a Bayesian structural equation modeling (BSEM) approach to test for statistically reliable differences between groups, thereby tolerating minor parameter variations but requiring the mean of differences to be approximately zero across groups. To this end, the relevant parameter differences need not be exactly zero but only close to zero (i.e., follow a prior distribution centered around zero). This less conservative approach is especially useful when there are several small deviations from strict invariance (De Boeck, 2008; Muthén & Asparouhov, 2013). Additionally, the BSEM approach generally requires smaller sample sizes (Hox, Moerbeek, Kluytmans, & van de Schoot, 2014; Hox, van de Schoot, & Matthijsse, 2012; van de Schoot et al., 2013), and a parameters-to-observations ratio of 1:3 is often sufficient to achieve meaningful results with this approach (Lee & Song, 2004). For these reasons, the BSEM approach can be fruitfully applied to smaller-scale studies (e.g., van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015) as done in the present case.

The Current Study

For the present study, the complex operation-span, reading-span, and symmetry-span tasks were obtained from the authors of the original versions (Unsworth et al., 2005). Then, the first and the second author translated task instructions and all other verbal task materials into German. One author always translated the entire task with the other author critically reviewing the translation. A native German student assistant finally checked for incomprehensibility issues of the translated material. Our translations of the complex-span tasks can be downloaded from Dr. Engle’s website (Attention & Working Memory Laboratory, 2016).

To validate the German versions of the complex-span tasks we had a sample of German participants perform our translated task versions and assessed the tasks’ reliabilities and convergent validities by correlating WMC with performance in a reasoning test, college admission grades, and another WMC test (i.e., running-memory-span task). We further tested whether the German task versions were metrically and functionally equivalent to the English task versions, so that one was able to use them analogously. To this end, we identified a similar North-American sample from a previous study (Unsworth et al., 2009) and compared the correlational structure between samples.

Methods

Participants

One hundred students (18–31 years, $M_{\text{age}} = 22$; 76% female, all native German speakers) from a German university and from other schools of higher education in the area participated in the two-session study.[1] The students majored in various subjects and received monetary compensation for participation; psychology students could opt for course credit.

Materials and Procedure

All tasks used in this study were programmed with the software Eprime (Psychology Software Tools, 2012); data

[1] Different cognitive tasks were administered in Sessions 1 and 2 and data from the two sessions were combined via an anonymous individual code. Because we were not able to link four participant codes from the second session to the codes from the first session, only 96 data points from the tasks from the first session were analyzed.

collection was organized with the software *hroot* (Bock, Baetge, & Nicklisch, 2014).

Complex-Span Tasks

Participants of the *Operation-Span (OSpan) task* are presented with math problems followed by a number. Participants must decide whether this number is the solution to the preceding math problem or not (processing component). After each math problem, a letter is presented for 800 ms, which participants must remember for a final memory test (storage-component). Each OSpan trial comprises a set of three to seven processing-storage units with a final memory test. For the memory tests, participants must identify the previously presented letters from the set in a 4 × 3 matrix that contains all possible letters (i.e., F, H, J, K, L, N, P, Q, R, S, T, and Y) by clicking on them in correct serial order. Participants are asked to hit a "blank" button whenever they have forgotten a letter in a series. There is no time limit for the final memory test. The OSpan task comprises a total of 15 trials (i.e., three of each set size). Imperceptible for participants, the task is further divided into three subblocks. Each set size occurs once in each subblock but the set size is randomly determined within each subblock.

The *Reading-Span (RSpan) task* has a similar structure as the OSpan task, except that the processing component requires judging whether a given sentence is semantically meaningful or not. The storage component is identical to that of the OSpan task. RSpan trials comprise three to seven processing-storage presentations and 15 trials (divided into three subblocks) in total.

The processing component of the *Symmetry-Span (SymSpan) task* requires judging whether a given geometric figure is symmetrical or not. For its storage component, a 4 × 4 matrix is presented after each geometric figure, one field of which is highlighted. At the end of each trial, which can comprise two to five processing-storage presentations, participants must recall the highlighted screen positions by clicking on an empty matrix in the correct order. The SymSpan task contains 12 trials divided into three subblocks.

All complex-span tasks include three practice blocks, that is, a storage-only (4 trials), a processing-only (15 trials), and a processing-and-storage block (15 trials). Based on the performance in the processing-only block, individual time limits are calculated for the combined practice block and for the actual task (i.e., mean response time plus 2.5 *SDs*). To prevent participants from delaying their processing in order to rehearse the to-be-stored items, processing trials are terminated whenever participants' response times exceed the time limit. The terminated trials are counted as processing errors in these tasks. Each complex-span task lasted approximately 12 min (M_{OSpan} = 12.24 min; M_{RSpan} = 12.12 min; M_{SymSpan} = 11.43 min).

Running-Memory-Span (RunSpan) Task

In each trial of this task, participants are presented with a series of letters and have to remember a certain set size, that is, the last *n* letters of a series ($3 \leq n \leq 7$). The same letters as for the OSpan and RSpan tasks are used in this task. Participants are informed about the set size at the beginning of each trial and there are three trials of each set size ($n + 0$; $n + 1$; $n + 2$). Letters are presented one after the other in the middle of the screen for 300 ms with an inter-stimulus-interval of 200 ms. At the end of each trial, participants have to select the previously presented *n* letters in the correct order in a 4 × 3 matrix that contains all possible letters. On average, it took participants 13.15 min to perform this task.

Reasoning Task

Participants are presented with 18 spatial reasoning problems one after another with the level of difficulty increasing from one to the next. The problems were drawn from the advanced progressive Raven matrices set II (Raven, 1962). Participants are instructed to solve as many problems as possible within 8 min. Each problem consists of a 3 × 3 matrix with each cell, except for the lower right one, containing geometric figures. The arrangement of figures within the matrix follows a certain pattern. Participants must select the figure they think will complete the pattern out of an array of six optional figures. This task lasted approximately 12 min.

Participants of the present study attended two sessions. In the first session, all participants performed the OSpan task, an unrelated long-term memory experiment, the RunSpan task, and an unrelated prospective-memory experiment. In the second session, that took place two to four weeks later, participants performed the RSpan, SymSpan, and the reasoning tasks and indicated their college admission grade (CAG). These grades are calculated from both individual high-school grades and standardized final exams and have been shown to be strongly correlated with intelligence test scores (Roth et al., 2015).

Task Scores

The complex-span-task programs provide different scores to the experimenter. The *partial span scores* (in some older task versions also called *total span scores*) reflect the sum of items recalled in the correct position, independent of whether the whole item set or only parts of it were correctly recalled. Because of their higher internal consistency and stronger correlations with external criteria, the authors of the

Table 1. Descriptive statistics and correlations for the German sample

Measure	*M*	*SD*	α	*N*	1	2	3	4	5	6
1. OSpan	60.40	10.60	.79	96	–	[.34, .63]	[.09, .45]	[.45, .71]	[.01, .39]	[−.49, −.12]
2. RSpan	54.41	13.53	.85	100	.50*	–	[.21, .54]	[.32, .63]	[.10, .46]	[−.54, −.17]
3. SymSpan	27.10	6.43	.68	100	.28*	.39*	–	[.06, .44]	[.10, .46]	[−.36, .05]
4. RunSpan	41.67	10.15	.65	94	.60*	.49*	.26*	–	[−.01, .37]	[−.52, −.15]
5. Reasoning	10.23	3.07	.71	100	.21*	.29*	.29*	.19	–	[−.53, −.16]
6. CAG	1.62	0.62	–	88	−.32*	−.37*	−.16	−.35*	−.36*	–

Notes. OSpan = operation-span task; RSpan = reading-span task; SymSpan = symmetry-span task; RunSpan = running-memory-span task; Reasoning = Raven-test-like matrices test; CAG = college admission grade. CAG varies between 1 and 4.5, with 1 indicating the best performance. Due to a programming error, we did not collect grades from all participants in the first sessions. Correlations are displayed below and 95% confidence intervals of correlations are displayed above the diagonal. *$p < .050$.

automated complex-span tasks recommend using the partial scores as an indicator of WMC (Conway et al., 2005) and we followed this recommendation.[2] The RunSpan task program also provides partial scores to the experimenter, which are recommended to be used (cf. Broadway & Engle, 2010).

For the reasoning scores, the sum of correctly solved figural reasoning problems was divided by the total amount of figural reasoning problems (i.e., 18).

The assessed CAGs could theoretically range from 1.00 (best to-be-achieved grade) to 4.49 (worst to-be-achieved grade). CAGs of 4.50 or lower do not allow for college admission.

Results

We set the Type I error probability to α = .05 for all analyses. The software *Mplus 8.0* (Muthén & Muthén, 1998–2016) was used to conduct the structural equation model analyses. Software inputs and outputs for all analyses are provided in the Electronic Supplementary Materials (see ESM 1–9).

Descriptive Statistics and Preliminary Analysis

Descriptive statistics for all tasks are displayed in Table 1. As each set size occurred three times in all working-memory tasks, we combined the first occurrences of all set sizes into one sub-score, the second occurrences into another sub-score, and the third occurrences into a third sub-score. We then calculated Cronbach's α across these sub-scores (cf. Unsworth et al., 2005). For the reasoning test, we conducted Cronbach's α across individual items. We conducted a CFA to test whether the three complex-span scores would form a general WMC factor and whether reasoning and CAG would form a general cognitive ability (gCA) factor. Additionally, we examined whether WMC and gCA were correlated. The model is displayed in Figure 2a. The χ^2-value, the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) were used to assess model fit.

The CFA model achieved an acceptable fit, $\chi^2(4) = 5.29$, $p = .259$; CFI = .982; RMSEA = .057; and SRMR = .037. In line with previous findings, as evident from Figure 2, there were substantial correlations between WMC and gCA, $r = .69$ (cf. Shipstead, Harrison, & Engle, 2016) as well as between WMC and RunSpan performance, $\beta = .73$, $p < .001$ (cf. Harrison et al., 2013).

The preliminary analyses already showed that all span tasks loaded on a latent WMC factor and that WMC was correlated with external criteria, such as cognitive abilities and another WMC measure. Next, we formally tested for measurement invariance between the WMC measure derived from our German sample with the translated task versions and the same WMC measure derived from a North-American sample with the original tasks.

Equivalence Tests Between the German and English Complex-Span Task Versions

We employed CFA and BSEM analyses to evaluate measurement invariance between our German translations of the complex-span tasks and the original English versions.

[2] The alternative *absolute span scores* (in some older task versions also just called *span scores*) reflect the sum of items recalled in the correct position, given that the complete item set had been correctly recalled on that trial. Furthermore, the programs provide measures of processing errors (called *math error total*, *reading error total*, or *symm error total,* depending on the task), that is, the total amount of errors made in the processing task component. These errors can be further distinguished into *accuracy errors* (number of trials in which false responses were given) and *speed errors* (number of trials in which the time limit was exceeded). The authors of the complex-span tasks point out that these error scores should also be inspected carefully to ensure that participants did not prioritize the storage task component at a cost to the processing task component (Conway et al., 2005; Unsworth et al., 2005).

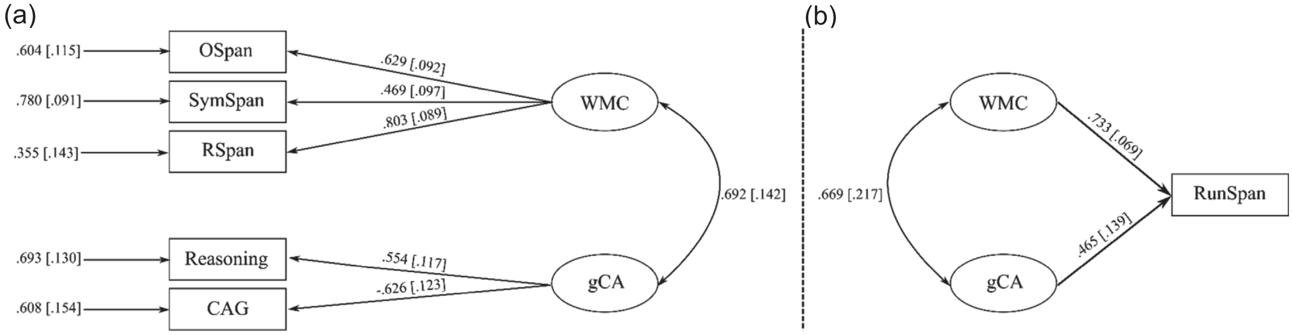

Figure 2. Confirmatory factor analysis for working-memory capacity (WMC) and its relation to general cognitive abilities (gCA). The path between the two latent variables (circles) represents their correlation, paths from the latent to the manifest variables (rectangles) represent loadings of the tasks on the latent factors. Numbers on the left side of the manifest variables represent error variances associated with the tasks. Standard errors are displayed in brackets. OSpan = operation span; SymSpan = symmetry span; RSpan = reading span; Reasoning = Raven-like matrices test; CAG = college admission grade; RunSpan = running-memory-span task.

Table 2. Exact and approximate measurement invariance tested with multi-group structural equation models

	Exact measurement invariance				Approximate measurement invariance	
	Δχ^2 (*df* = 2)	CFI	RMSEA	SRMR	*ppp*	BCI (95%)
Configural invariance	–	1.000	.000	.000	.480	(−16.980)–(16.632)
Metric invariance	1.164	1.000	.000	.038	.550	(−17.533)–(16.144)
Scalar invariance	7.924*	0.973	.111	.058	.257	(−9.647)–(21.273)

Notes. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; *ppp* = posterior predictive probability; BCI = Bayesian credibility interval for the difference between the observed and the replicated χ^2-value, *N* = 205, *$p < .050$.

For this analysis, we used the dataset from Unsworth et al. (2009) as the reference sample.[3] Because our sample was an all-student sample, we excluded all nonstudents from the Unsworth-et-al. sample. This resulted in $N_{\text{North-American sample}}$ = 105 (18–32 years, M_{age} = 21; 53% female). The OSpan, RSpan, and SymSpan scores were used as indicators for a latent WMC factor. In each sample, two subscores were obtained from the reasoning task by employing an odd/even split. The sub-scores were then used as indicators for a latent reasoning factor.

Measurement Invariance of the WMC Tasks

For the configural invariance model, all model parameters were estimated freely for both samples. For the metric invariance model, factor loadings (λ_i) were restricted to be equal across samples. For the scalar invariance model, the item intercepts (δ_i) were additionally restricted to be equal across samples. Changes in χ^2-values as well as in CFI, RMSEA, and SRMR were used to evaluate the exact measurement invariance between the configural and the metric invariance model as well as between the metric and the scalar invariance model. In general, a nonsignificant χ^2-value suggests adopting the more parsimonious model (scalar over metric, metric over configural). According to Chen (2007) ΔCFI $\leq -.005$, ΔRMSEA $\leq .010$, and ΔSRMR $\leq$ .025 between the configural and the metric model suggests metric invariance and ΔCFI $\leq -.005$, ΔRMSEA $\leq .010$, and ΔSRMR $\leq .005$ between the metric and the scalar model suggests scalar invariance in smaller samples. These values were developed based on simulation studies not only with samples of $N = 300$ but also with models with considerably more indicators (i.e., 8 or 12) than the present study. We thus decided to preferably rely on changes in χ^2-values and CFI, which are less affected by sample size and model complexity, in case of inconclusive evidence from the indices.

Model fit results are displayed in Table 2. Whereas χ^2-values, CFI, and RMSEA suggest metric invariance, SRMR indicated configural invariance. In light of the conflicting evidence, we chose to prioritize χ^2-values and CFI evidence, which both suggested metric invariance. Because the estimated model parameters based on the three models revealed only minor differences in factor loadings and intercepts, we additionally investigated approximate measurement invariance, which seemed to be most appropriate in the present case (De Boeck, 2008). Whereas the present

[3] We thank Nash Unsworth for sharing his dataset with us.

sample size was rather small for exact measurement invariance tests in the CFA framework, it was supposedly more adequate for Bayesian approximate invariance tests, as the BSEM approach requires smaller samples (Hox et al., 2014; Lee & Song, 2004; van de Schoot et al., 2013). Approximate measurement invariance was evaluated based on the posterior predicted probability (*ppp*) and the Bayesian credibility interval (BCI) for the difference between the observed and the replicated χ^2-value. A *ppp* > 0 and a 95%-BCI including zero were considered as indicative for invariance between two models (e.g., van de Schoot et al., 2013). As evident from Table 1, both *ppp* and BCI suggest accepting approximate scalar invariance.

Comparability of Reliabilities and Convergent Validities

Composite reliability was evaluated as the ratio of true-score variance of the performance score (σ_{Tp}) to the total performance variance $(\sigma_{\mathrm{p}}) : \omega = \sigma_{\mathrm{TP}}/\sigma_{\mathrm{P}} = [\Sigma\,(\lambda_i)]^2 \times \eta/[\Sigma\,(\lambda_i)]^2 \times \eta + \Sigma(\varepsilon_i)$. This composite reliability estimator, which is suggested by Raykov (1997; see also McDonald, 1999), for example, has the advantage – compared to traditional approaches like Cronbach's α – of the measurement model and the reliability being estimated simultaneously, without requiring any additional assumptions. Furthermore, the composite reliability provides an unbiased estimate whereas Cronbach's α tends to underestimate the reliability of tau-congeneric measurement models with unequal factor loadings across indicators (cf. Cortina, 1993). Finally, reliabilities obtained within different groups can be directly compared in the CFA framework (Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000).

We estimated the composite reliability within the CFA model described in Figure 1. The estimated reliability was .87 in the German and .75 in the North-American sample. We compared model fits of a model without restrictions and a model assuming equal loadings, true-score and error variances across the two samples. The difference between models was significant $\Delta\chi^2(6) = 13.67$, $p = .034$. At first glance, this result suggests that the WMC factor obtained with the German tasks is more reliable than that obtained with the English tasks. However, as reliability is defined as the ratio of true-score variance to total variance, it reflects both the magnitude of measurement error as well as the magnitude of the true-score variance (e.g., Lord & Novick, 1968). In the present case, the lower reliability in the North-American sample was due to a lower true-score variance rather than a higher measurement error and thus reflects sample characteristics rather than lower measurement precision.

The convergent validity of the WMC task was evaluated based on the correlation between the latent WMC and the latent reasoning factors. The correlation was $r = .69$ in the German sample and $r = .42$ in the North-American sample. We compared the model fit of a model without restrictions and a model assuming equal correlations. The difference between models was not significant, $\Delta\chi^2(1) = 2.42$, $p = .120$, suggesting similar convergent validities within the two samples.

Discussion

The German task versions showed satisfying reliabilities and decent levels of convergent validity with indicators of cognitive abilities and another WMC task. Additionally, we found strong evidence for exact metric and approximate scalar equivalence between the translated German and the original English versions. Finally, the correlations between the latent WMC factor and reasoning abilities did not differ significantly across the two language groups. Numerically, however, the translated task versions correlated more strongly with the reasoning test compared to the original versions. In light of the present sample size, which was sufficient for the (approximate) measurement invariance test, but rather small for the comparison of correlation coefficients, one could argue that the statistical power simply was not high enough to detect a difference between correlations. Thus we cannot make a strong argument that the correlation between WMC and reasoning in the German and the North-American sample was truly equivalent. However, the correlation in our German sample was quite in the range of correlations between WMC and higher-order cognitive abilities usually observed in North-American samples (Shipstead et al., 2016), whereas the correlation in the Unsworth-et-al. sample was a bit lower than one would usually expect. We therefore interpret the present finding as preliminary evidence for the convergent validity of the German complex-span tasks that lies within the to-be-expected range. Taken together, the present findings provide satisfactory support for metric measurement equivalence between the English and German tasks as well as preliminary evidence for functional equivalence as indexed by the convergent validities with a reasoning test.

There are other German WMC-assessment tools available, which partly rely on similar tasks but somewhat different WMC conceptualizations (e.g., Lewandowsky, Oberauer, Yang, & Ecker, 2010). We believe, however, that the German versions of the complex-span tasks that are comparable with the widely used English versions by Unsworth, Engle, and colleagues will be useful for researchers from various fields independent of which theoretical working-memory approach they favor (see also Conway et al., 2005, for an overview of different application domains).

The present results suggest that our translated tasks measured the construct of interest similarly well as the original tasks. However, when translating cognitive tasks, one cannot take such equivalence for granted. Due to the verbal aspects of task materials or instructions, translation-immanent biases may reduce the consistency of results obtained with supposedly parallel language versions, indicating methodological rather than theoretical shortcomings. Many cognitive psychology researchers who are concerned with drawing cross-national or cross-cultural comparisons usually take these issues into account (e.g., Zuilkowski et al., 2016; but see also Ellefson, Ng, Wang, & Hughes, 2017, for a recent example where this is not the case). Researchers who simply translate tasks to use them in their own language area, however, often largely ignore these challenges. For cognitive psychology as a global science, we think these considerations are especially important as a lack of metric and/or functional equivalence between tasks that should measure the same construct can be responsible for bad replicability of findings.

In the light of recent global initiatives investigating the replicability of psychological findings, one should consider the possibility of translation-immanent biases when evaluating "reproducibility failures." Stanley and Spence (2014) recently pointed out that artifacts, such as differences in sampling error, measurement error, and range restrictions, render exact replication very difficult. Whereas it is not possible to test the influence of unsystematic artifacts, one can test for potential systematic artifacts, such as translation-immanent biases. For example, in a recent unsuccessful multi-labs' replication attempt of the "ego-depletion" effect (Hagger & Chatzisarantis, 2016b), there was a trending effect across the English studies but not across other-language studies (Sripada, Kessler, & Jonides, 2016). This observation does not fully explain the replication failure because even across the English studies the effect was very small (Hagger & Chatzisarantis, 2016a) and the inability to produce the effect with task settings in other languages may well just speak against the effect's generalizability. However, there is a chance that that original and translated tasks were not equivalent, rendering the cross-language replications especially troublesome.

Also for the widely acknowledged Reproducibility Project (Open-Science-Collaboration, 2012, 2015), some of the tasks from the original studies were translated for replication in a different country – but without testing whether the translated versions were equivalent to the original versions. Others have raised somewhat similar concerns by arguing that cultural differences were responsible for the poor replicability rates reported in the Reproducibility Project (Gilbert, King, Pettigrew, & Wilson, 2016). To be clear, we do not believe that translation-immanent biases or actual cultural differences can explain the low replication success rate in cognitive psychology and we generally agree with the authors of the Reproducibility Project that the low replicability rate of psychological findings is alarming (Open-Science-Collaboration, 2016). Nevertheless, translation-immanent biases may have contributed to some of the replication failures in this project and future global replication projects should control for them. We therefore suggest ensuring metric and functional equivalence by employing Bayesian tests of approximate measurement invariance whenever different versions of a cognitive task are used in different language areas and results obtained with these task versions are then combined in one way or another.

In conclusion, because measurement invariance tests can be employed to test for translation-immanent biases without much additional effort, we believe they are a powerful tool that should not be limited to only being applied to tasks that are used in cross-national and cross-cultural research. Instead, we strongly suggest that they should also be used for ensuring the general appropriateness of task translations. In the present study, we showed that the test scores obtained with the German versions of the automated complex-span tasks are reliable, valid, and equivalent to the original English versions. These findings indicate that the German versions of the complex-span tasks can be applied to German populations in a similar manner as the original English tasks are applied to English populations. Furthermore, they illustrate the general usefulness of measurement invariance tests for the validation of translated cognitive tasks.

Electronic Supplementary Materials

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1015-5759/a000444

ESM 1. Text (pdf).
List of supplements with descriptions and abbreviations.
ESM 2. Data (out).
Factor loadings for WMC factor and correlation with gCA using CFA (input and output files).
ESM 3. Data (out).
Measurement invariance tests using multigroup CFA (input and output files).
ESM 4. Data (out).
Tests for measurement invariance using BSEM (input and output files).
ESM 5. Data (out).
Tests for measurement invariance using BSEM (input and output files).
ESM 6. Data (out).

Tests for measurement invariance using BSEM (input and output files).
ESM 7. Data (pdf).
Parameter estimates of all the invariance tests.
ESM 8. Data (out).
Comparison of reliabilities between language groups (input and output files).
ESM 9. Data (out).
Comparison of WMC-reasoning correlation between language groups (input and output groups).

References

Attention & Working Memory Laboratory. (2016). *Task downloads*. Retrieved from http://englelab.gatech.edu/tasks.html

Barrett, L. F., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin, 130*, 553–573. https://doi.org/10.1037/0033-2909.130.4.553

Bock, O., Baetge, I., & Nicklisch, A. (2014). Hroot: Hamburg registration and organization online tool. *European Economic Review, 71*, 117–120. https://doi.org/10.1016/j.euroecorev.2014.07.003

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods, 42*, 563–570. https://doi.org/10.3758/Brm.42.2.563

Byrne, B. M., & Stewart, S. A. (2006). 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. https://doi.org/10.1207/s15328007sem1302_7

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233–255. https://doi.org/10.1207/S15328007

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*, 769–786. https://doi.org/10.3758/Bf03196772

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences, 7*, 547–552. https://doi.org/10.1016/j.tics.2003.10.005

Cortina, J. M. (1993). What is coefficient alpha: An examination of theory and applications. *The Journal of Applied Psychology, 78*, 98–104. https://doi.org/10.1037/0021-9010.78.1.98

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559. https://doi.org/10.1007/s11336-008-9092-x

De Neys, W., D'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica, 42*, 177–190.

Delaloye, C., Ludwig, C., Borella, E., Chicherio, C., & de Ribaupierre, A. (2008). L'Empan de lecture comme épreuve mesurant la capacité de mémoire de travail: Normes basées sur une population francophone de 775 adultes jeunes et âgés [The reading span as a measure of working memory capacity: Norms based on a French speaking population of 775 younger and older adults]. *Revue Europeenne De Psychologie Appliquee – European Review of Applied Psychology, 58*, 89–103. https://doi.org/10.1016/j.erap.2006.12.004

Ellefson, M. R., Ng, F. F., Wang, Q., & Hughes, C. (2017). Efficiency of executive function: A two-generation cross-cultural comparison of samples from Hong Kong and the United Kingdom. *Psychological Science, 28*, 555–566. https://doi.org/10.1177/0956797616687812

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304–312. https://doi.org/10.1037/1040-3590.6.4.304

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "estimating the reproducibility of psychological science". *Science, 351*, 1037. https://doi.org/10.1126/science.aad7243

Gonthier, C., Thomassin, N., & Roulin, J. L. (2016). The composite complex span: French validation of a short working memory task. *Behavior Research Methods, 48*, 233–242. https://doi.org/10.3758/s13428-015-0566-3

Grigorenko, E. L., Jarvin, L., Kaani, B., Kapungulya, P. P., Kwiatkowski, J., & Sternberg, R. J. (2007). Risk factors and resilience in the developing world: One of many lessons to learn. *Development and Psychopathology, 19*, 747–765. https://doi.org/10.1017/S0954579407000375

Gudmundsson, E. (2009). Guidelines for translating and adapting psychological instruments. *Nordic Psychology, 61*, 29–45. https://doi.org/10.1027/1901-2276.61.2.29

Hagger, M. S., & Chatzisarantis, N. L. D. (2016a). Commentary: Misguided effort with elusive implications, and shifting signal from noise with replication science. *Frontiers in Psychology, 7*, 621. https://doi.org/10.3389/fpsyg.2016.00621

Hagger, M. S., & Chatzisarantis, N. L. D. (2016b). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573. https://doi.org/10.1177/1745691616652873

Hambleton, P. F. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*, 164–172. https://doi.org/10.1027/1015-5759.17.3.164

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (Vol. 1, pp. 35–56). Hoboken, NJ: Wiley.

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Memory & Cognition, 43*, 389–396. https://doi.org/10.3758/s13421-014-0473-3

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24*, 2409–2419. https://doi.org/10.1177/0956797613492984

He, J., & Van de Vijver, F. J. R. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*, 1–19. https://doi.org/10.9707/2307-0919.1111

Hox, J. J., Moerbeek, M., Kluytmans, A., & van de Schoot, R. (2014). Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in Psychology, 5*, 78. https://doi.org/10.3389/fpsyg.2014.00078

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*, 87–93. https://doi.org/10.18148/srm/2012.v6i2.5033

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*, 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kankaras, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology, 45*, 381–399. https://doi.org/10.1177/0022022113511297

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*, 653–686. https://doi.org/10.1207/s15327906mbr3904_4

Lewandowsky, S., Oberauer, K., Yang, L. X., & Ecker, U. K. H. (2010). A working memory test battery for MATLAB. *Behavior Research Methods, 42*, 571–585. https://doi.org/10.3758/Brm.42.2.571

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*, 8–14. https://doi.org/10.1177/0963721411429458

Muniz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traduccion y adaptacion de los tests: Segunda edicion [Guidelines for test translation and adaptation: Second edition]. *Psicothema, 25*, 151–157. https://doi.org/10.7334/psicothema2013.24

Muthén, B. O., & Asparouhov, T. (2013). *BSEM measurement invariance analysis: Mplus web note 17*. Retrieved from http://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, B. O., & Muthén, L. K. (1998–2016). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Oberauer, K., Süss, H. M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity – facets of a cognitive ability construct. *Personality and Individual Differences, 29*, 1017–1045. https://doi.org/10.1016/S0191-8869(99)00251-2

Open-Science-Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. https://doi.org/10.1177/1745691612462588

Open-Science-Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aa4716/4711–aa4716/4718. https://doi.org/10.1126/science.aac4716

Open-Science-Collaboration. (2016). Response to comment on "Estimating the reproducibility of psychological science". *Science, 351*, 1037. https://doi.org/10.1126/science.aad9163

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development, 78*, 1255–1264. https://doi.org/10.1111/j.1467-8624.2007.01064.x

Piocuda, J. E., Smyers, J. O., Knyshev, E., Harris, R. J., & Rai, M. (2015). Trends of internationalization and collaboration in US psychology journals 1950–2010. *Archives of Scientific Psychology, 3*, 82–92. https://doi.org/10.1037/arc0000020

Psychology Software Tools. (2012). *[E-Prime 2.0]*. Retrieved from http://www.pstnet.com

Raven, J. C. (1962). *Advanced progressive matrices: Set II*. London, UK: Lewis.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329–353. https://doi.org/10.1207/s15327906mbr3204_2

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*, 164–171. https://doi.org/10.1027/15-5759/a000123

Roth, B., Becker, N., Romeyke, S., Schafer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

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

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11*, 771–799. https://doi.org/10.1177/1745691616650647

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science, 9*, 552–555. https://doi.org/10.1177/1745691614543974

Sripada, C., Kessler, D., & Jonides, J. (2016). Sifting signal from noise with replication science. *Perspectives on Psychological Science, 11*, 576–578. https://doi.org/10.1177/1745691616652875

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science, 9*, 452. https://doi.org/10.1177/1745691614542787

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica, 134*, 16–28. https://doi.org/10.1016/j.actpsy.2009.11.010

Unsworth, N., Heitz, R. R., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*, 498–505. https://doi.org/10.3758/Bf03192720

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory, 17*, 635–654. https://doi.org/10.1080/09658210902998047

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology, 6*, 25216. https://doi.org/10.3402/ejpt.v6.25216

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89–99. https://doi.org/10.1027/1016-9040.1.2.89

Van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. R. Matsumoto (Ed.), *Cross cultural research methods in psychology* (Vol. 1, pp. 17–45). New York, NY: Cambridge University Press.

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. https://doi.org/10.1177/109442810031002

Zuilkowski, S. S., McCoy, D. C., Serpell, R., Matafwali, B., & Fink, G. (2016). Dimensionality and the development of cognitive assessments for children in Sub-Saharan Africa. *Journal of Cross-Cultural Psychology, 47*, 341–354. https://doi.org/10.1177/0022022115624155

Received September 28, 2016
Revision received June 7, 2017
Accepted June 7, 2017
Published online December 15, 2017
EJPA Category/Section: Cognitive Science

Jan Rummel
Department of Psychology
Heidelberg University
Hauptstrasse 47-51
69117 Heidelberg
Germany
jan.rummel@psychologie.uni-heidelberg.de