Intelligence 98 (2023) 101751

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

The cross-cultural generalizability of cognitive ability measures: A systematic literature review.

Christopher J. Wilson [a,b,*], Stephen C. Bowden [a], Linda K. Byrne [c], Nicole R. Joshua [b], Wolfgang Marx [d], Lawrence G. Weiss [e]

[a] Melbourne School of Psychological Sciences, The University of Melbourne, Parkville, VIC, Australia
[b] Pearson Clinical Assessment, Pearson, Richmond, VIC, Australia
[c] School of Psychology, Deakin University, Burwood, VIC, Australia
[d] iMPACT, School of Medicine, Deakin University, Geelong, Australia
[e] Test Development Consultant, San Antonio, TX, USA

ARTICLE INFO

Keywords:
Measurement invariance
Factorial invariance
Cross-culture
Cognitive assessment
Cattell-Horn-Carroll model
Factor analysis

ABSTRACT

Examining factorial invariance provides the strongest test of the generalizability of psychological constructs across populations and should be investigated prior to cross-cultural interpretation of cognitive assessments. The aim of this systematic review was to critically evaluate the current evidence regarding the factorial invariance and the generalizability of cognition models across cultures. The review was structured using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The literature search identified 57 original studies examining the factorial invariance of cognitive ability assessments across cultures. The results were strongly supportive of the cross-cultural generalizability of the underlying cognitive model. Ten studies found configural invariance, 20 studies found weak or partial weak factorial invariance, 12 found strong or partial strong factorial invariance, and 13 found strict factorial invariance. However, the quality of the factorial invariance analyses varied between studies, with some analyses not adopting the hierarchical approach to factorial invariance analysis, leading to ambiguous results. No study that provided interpretable results in terms of the hierarchical approach to factorial invariance found a lack of factorial invariance. Overall, the results of this review suggest that i) the factor analytic models of cognitive abilities generalize across cultures, ii) the use of the hierarchical approach to factorial invariance is likely to find strong or strict factorial invariance, iii) the results are compatible with well-established Cattell-Horn-Carroll constructs being invariant across cultures. Future research into factorial invariance should follow the hierarchical analytic approach so as not to misestimate factorial invariance. Studies should also use the Cattell-Horn-Carroll taxonomy to systematize intelligence research.

1. Introduction

Examining measurement invariance involves the simultaneous analysis of a measurement model across two or more groups. Extending the analysis of measurement invariance by evaluating the invariance of the factor structures across groups is referred to as 'factorial invariance' (Meredith, 1993; Widaman & Olivera-Aguilar, 2023). Establishing measurement or factorial invariance can be understood as a test for the presence of bias, either in terms of scaling or intercept bias, by demonstrating that two individuals with the same ability but from different populations, when undertaking the same test, will achieve the same test score (Widaman & Reise, 1997). Finding invariance (i.e., equality) in the measurement of cognitive abilities across populations is required to generalize construct validity evidence from research in one population to application in different populations (AERA et al., 2014; Bowden, Lange, Weiss, & Saklofske, 2008). Further, factorial invariance is the strongest test of the generalizability of psychological constructs across different populations (Bowden, Petrauskas, Bardenhagen, Meade, & Simpson, 2013; Bowden, Weiss, Holdnack, & Lloyd, 2006; Horn & McArdle, 1992; Widaman & Reise, 1997). Establishing factorial invariance is necessary for meaningful interpretation of the construct validity evidence underlying a test used for diagnostic and classification

* Corresponding author at: Melbourne School of Psychological Sciences, The University of Melbourne, Parkville 3010, VIC, Australia.
E-mail address: wilsonc4@student.unimelb.edu.au (C.J. Wilson).

https://doi.org/10.1016/j.intell.2023.101751
Received 13 September 2022; Received in revised form 14 March 2023; Accepted 18 March 2023
Available online 24 March 2023

decisions across populations (Meredith, 1993; Widaman & Reise, 1997).

A failure of factorial invariance of a test used across cultures can lead to misinterpretation of scores from any individual or group when interpreted in terms of research or norms derived from another culture. Failure of factorial invariance may also lead to misinterpretation of the comparison of group means. Therefore, cross-cultural factorial invariance must be investigated prior to cross-cultural interpretations of test scores or convergent and discriminant validity evidence (Chen, Keith, Weiss, Zhu, & Li, 2010; Putnick & Bornstein, 2016). Evidence of factorial invariance in cognitive ability tests has been found across age, gender, diagnostic groups, language and over time (Bowden et al., 2006; Bowden et al., 2008; Goodman et al., 2021; Jewsbury, Bowden, & Duff, 2016; McArdle, Fisher, & Kadlec, 2007; Savage-McGlynn, 2012; Watkins & Canivez, 2001). While differences in cognition have been shown to be mediated by socio-economic, educational, occupational, physical and mental health, among other factors, the demonstration of factorial invariance allows for the generalization of construct validity evidence, essential for theoretical generalizations (Bontempo & Hofer, 2007; Weiss & Saklofske, 2020).

1.1. The hierarchical approach to factorial invariance

The process of establishing measurement or factorial invariance involves a series of steps. Each step seeks to reject the null hypothesis of invariance or numerical equivalence of the measurement model parameters across populations. Factorial invariance is evaluated on a hierarchy of increasingly restrictive analyses using multiple-group confirmatory factor analysis (CFA; Bontempo & Hofer, 2007; Meredith, 1993; Meredith & Teresi, 2006; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Published guidelines are available for multi-group CFA and are appropriate for large-scale cross-cultural analysis (Alkemade, Bowden, & Salzman, 2015; Bowden, Saklofske, van de Vijver, Sudarshan, & Eysenck, 2016; Meredith, 1993; Meredith & Teresi, 2006; Millsap, 2011; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Exploratory factor analysis does not allow tests of the restrictive parameter estimation required for measurement or factorial invariance (Brown, 2015).

A full CFA measurement model is explained in relation to five matrices calculated from a multi-group CFA for continuous or interval-scale indicators or variables (Brown, 2015; Vandenberg & Lance, 2000). The first three matrices make up the measurement model and are i) the factor loadings (λ), ii) the vector of observed variable intercepts (τ), and iii) the residual variances (θ), which together are the measurement components of the latent-variable model (see Fig. 1) and permit algebraic estimation of the latent variables in terms of the observed variables. The final two matrices in the multiple-group CFA make up the structural model and are iv) the vectors of latent means in each group and v) the matrix of variances and covariances between factors or latent variables. Together, matrices iv and v are termed the structural components as they represent the values of, and relationships between, the latent variables (Widaman & Reise, 1997). The numerical equality of the values of the first three of these five matrices is required for the assumption of measurement or factorial invariance (Meredith, 1993). When representing measurement invariance or factorial structures within a CFA model across groups, Widaman and Reise (1997) outline four forms of factorial invariance, (i) configural invariance, (ii) weak factorial invariance, (iii) strong factorial invariance, and (iv) strict factorial invariance. Their systematic process outlines an interpretive sequence that begins with a freely estimated model in two samples, apart from identification constraints, with the incremental addition of constraints at each step (see Fig. 1). The increasingly restrictive hierarchy of factorial invariance tests, if shown to demonstrate strong or strict factorial invariance, provides evidence for direct group comparisons and allows for the identification of the source of any potential non-invariance (Meredith & Teresi, 2006). See Table 1 for a summary of the factorial invariance hierarchy and steps, common nomenclature, and

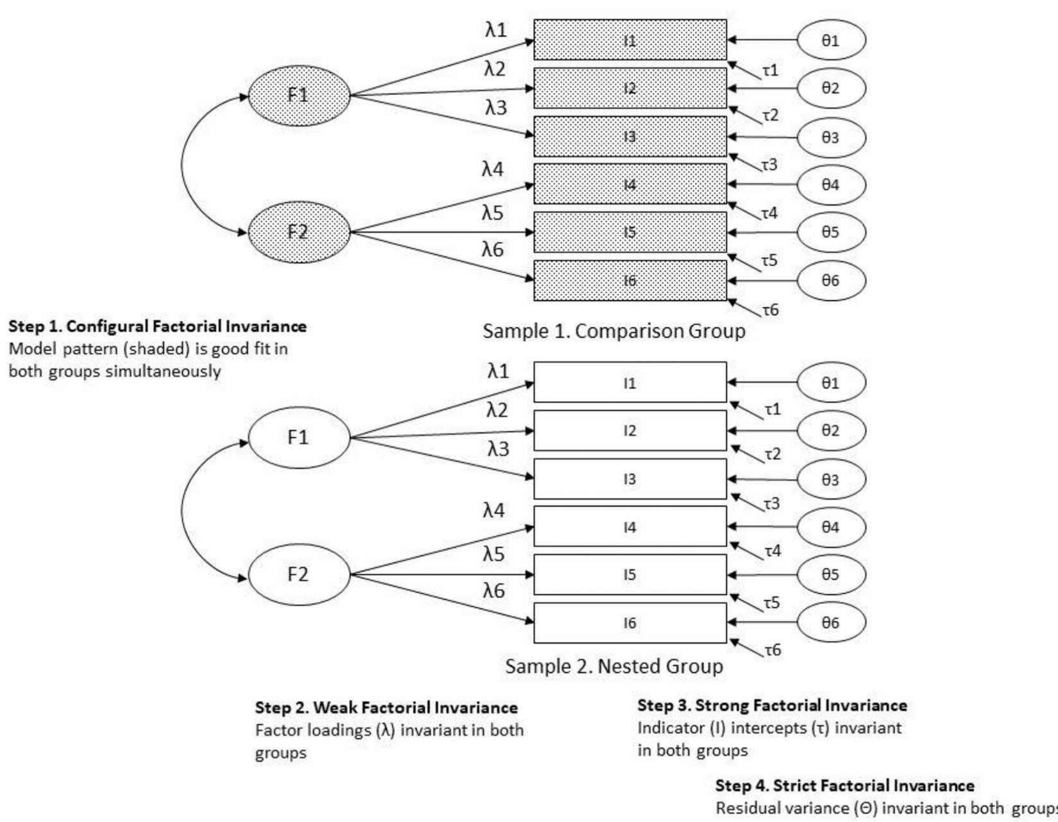

Fig. 1. Factorial Invariance Hierarchical Steps.

Table 1
The Hierarchical Approach and Factorial Invariance Classifications.

Variable	Factorial Invariance			
Hierarchical Approach	Step 1 Nonmetric Configural	Step 2 Weak Factorial	Step 3 Strong Factorial	Step 4 Strict Factorial
Also Known As	Baseline model Equal Form	Equal Factor Loadings Pattern invariance Metric invariance	Scalar Equal intercepts Thresholds	Residual Unique invariances Invariant uniqueness
Algebraic Matrix	Variance/covariance matrix	Lambda (λ) Factor loading matrix	Tau (τ) Vector of the observed-variable intercepts	Theta (Θ) Error variance–covariance matrix
Restrictions/ Constraints	Holding only the pattern of factor loadings invariant	Factor loadings invariant across groups; intercepts and residuals unconstrained	Factor loadings and intercepts invariant across groups; residuals unconstrained	Factor loadings, intercepts, and residuals invariant across groups
Invariance Interpretation	Pattern of constructs is identical across groups. Serves as a baseline of model fit for comparison for more restrictive models.	Permits comparisons of factor variances and covariances but not factor means. The unit of measurement is identical across groups.	The absolute scores of the observed variables can be directly compared across groups. The latent factors have the same meaning across the groups.	The residual variance, a combination of reliable unique variance and unreliable (error) variance is equivalent across groups. Common factors the cause of any group differences in means and variances.

factorial invariance interpretation.

1.2. The hierarchical sequence

1. Configural invariance is the most basic form of factorial invariance. Configural invariance can be understood as 'nonmetric' invariance and involves observation of the same factor-loading pattern of the cognitive ability model across groups (Widaman & Reise, 1997). Configural invariance requires that the same indicators (i.e., test items or subtests) load on the same latent variables (i.e., factors or constructs) across groups. Configural invariance demonstrates that the same baseline CFA provides the best fit to the data in the groups being examined and is the starting point for the examination of factorial invariance. When configural invariance is found, the same number of factors and the overall factor-loading pattern of the cognitive ability model is identical across groups. Configural invariance must be established in order for subsequent tests to be meaningful (Horn & McArdle, 1992). Establishing configural invariance provides evidence that the same psychological constructs and factor pattern are the best fit in both groups.
2. Weak factorial invariance adds the additional criterion that the same factor structure and numerical values of factor loadings are identical across groups. When weak factorial invariance is established, the comparison of factor variances and covariances across groups is interpretable in a straightforward manner. Weak factorial invariance implies that the unit of measurement of the factor(s) is identical across groups and allows for direct comparisons across groups in terms of convergent and discriminant construct validity. A finding of weak factorial invariance infers that the indicators (i.e., subtests) have equivalent relationships with the factors (i.e., constructs) across groups. Note, establishing weak factorial invariance is not sufficient to allow for comparisons of mean scores across groups.
3. Strong factorial invariance, also known as scalar invariance, is more restrictive. Strong factorial invariance requires the equality of the factor structure and factor loadings and adds the additional constraint that the indicator intercepts are numerically identical across groups. A finding of strong factorial invariance is required for meaningful interpretations of group mean differences.
4. Strict factorial invariance is the most restrictive and requires the same factor structure, factor loadings, and intercepts and adds the constraint of the identical numerical value of the residual variances for items for the model under analysis. Strict factorial invariance suggests that "group differences in means and variances on the measured variables are a function only of group differences in means and variances on the common factors" (Widaman & Reise, 1997, p. 296) and that "group differences in manifest variable means and variances are accounted for by group differences in common factor means and dispersion matrices" (Meredith & Teresi, 2006, p. 7).

Demonstration of factorial invariance at any level, that is configural, weak, strong, or strict factorial invariance, allows for some level of scientifically justifiable interpretation (Horn & McArdle, 1992; Widaman & Reise, 1997). Configural invariance provides evidence that the same factor structure displays a good fit across populations and may be sufficient to justify cultural and language adaptation and local norming of an imported test (Bowden et al., 2016).

Establishing weak factorial invariance is sufficient for theoretical and applied construct validity interpretations in terms of convergent and discriminant validity. Partial weak factorial invariance may also allow for the generalization of construct validity research (Horn & McArdle, 1992; Meredith, 1993; Meredith & Teresi, 2006; Widaman & Reise, 1997). However, a finding of strong factorial invariance is required to allow for a straightforward comparison of group means and covariances (correlations in the standardized form) of the latent variables and comparison of observed mean scores between groups (Widaman & Reise, 1997). Strict factorial invariance further assumes equivalent residual variances. Residual variances reflect a combination of reliable unique variance and unreliable (error) variance but may be overly restrictive and unnecessary for establishing the generalization of construct validity (Byrne, 2004; Horn & McArdle, 1992; Millsap, 2011; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Further, strict factorial invariance is less important where group differences are indicative of real-world differences such as those found across cultures (Meredith & Teresi, 2006). Indeed, it is often reasonable for residual variances to differ across populations (Horn & McArdle, 1992; Widaman & Reise, 1997).

An alternate analytic approach to the increasingly restrictive hierarchical method is to move directly from configural invariance to strict factorial invariance (i.e., the most constrained model). If strict factorial invariance is established, testing for weak and strong factorial invariance is redundant. This approach has the benefit of reducing the Type I error rate or multiple-test problem (Vandenberg & Lance, 2000). However, as noted, it is often unreasonable to expect strict factorial invariance (Brown, 2015; Meredith & Teresi, 2006). Another approach to the increasingly restrictive hierarchical method is the step-down approach, whereby the starting model is the most constrained model, and the subsequent models are evaluated for improved fit as factorial invariance restrictions are relaxed.

When the test of the null hypothesis of factorial invariance is rejected at the weak, strong, or strict level, partial factorial invariance can be investigated (Byrne, Shavelson, & Muthén, 1989). Under partial factorial invariance, only some of the parameter estimates in the loadings, intercepts, or residual matrices may be constrained to invariance across groups, with the remaining allowed to vary freely (Millsap & Kwok, 2004). A failure of weak or strong factorial invariance may still allow for the comparison of constructs across cultures. Indeed, provided at least the same three indicators of the measured construct satisfy the requirements of weak or strong factorial invariance, comparisons of constructs across cultures can still be performed without limitations (Alkemade et al., 2015; Bowden et al., 2016; Millsap, 2011). However, partial factorial invariance does not allow for comparisons of the subset of indicators in which factorial invariance failed, only on the subset that was found invariant.

While there are many methods to estimate the common factor model in CFA, the most common estimation method is maximum likelihood. Maximum likelihood is a full information estimator which allows for an evaluation of how well the factor solution reproduces the observed variances and covariances among the indicators (Brown, 2015). If the indicators are dichotomous or categorical, estimators such as weighted least squares are more appropriate, however, they may require larger samples (Brown, 2015; Browne, 1984).

When examining factorial invariance, the measurement models are compared using model fit. Model fit is determined by examining multiple indices representing a variety of fit criteria. Configural invariance, the first step in the hierarchical approach, is determined by evaluating the overall model fit. While weak, strong, and strict factorial invariance are determined by comparing the fit of the nested models with the added constraints. Any differences between models can then be ascribed to the added constraints. The two common ways to assess model fit include the chi-square test and fit indexes. The chi-square test has been found to be overly sensitive to reject the null hypothesis of invariance when applied in large samples (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008; Vandenberg & Lance, 2000). It is therefore recommended to use fit indices which are less sensitive to sample size (Hu & Bentler, 1999).

Currently, there are no universally agreed statistical rules for interpreting fit indices for the evaluation of measurement or factorial invariance (Brown, 2015). However, there are empirical studies to guide invariance researchers. First, when conducting baseline CFA model estimation in the respective samples, Hu and Bentler (1999) recommend an SRMR below 0.08, RMSEA below 0.06, and CFI and TLI greater than 0.95 as a demonstration of good model fit. Next, when comparing the nested models estimated simultaneously in the respective sample for measurement invariance analysis, Cheung and Rensvold (2002) recommend using a change in CFI value of less than or equal to 0.01 to retain the assumption of invariance. Additionally, Chen (2007) recommends a change of less than or equal to 0.01 in CFI, supplemented by a change of less than or equal to 0.015 in RMSEA or a change of less than or equal to 0.03 in SRMR to retain the assumption of invariance.

Recent research, however, has reported that these most widely used cut-offs for the fit indices have inconsistent Type I error rates and instead suggest permutation tests (Jorgensen, Kite, Chen, & Short, 2018). Permutation testing creates an empirical approximation of a sampling distribution for which models can be compared. Importantly, sample sizes of less than 100 per comparison group will likely not have the power to detect a lack of invariance (Meade et al., 2008), however, the addition of groups will add to the power. Otherwise, achieving factorial invariance has been shown to be unrelated to sample size (Rutkowski & Svetina, 2013).

1.3. The Cattell-Horn-Carroll Model

Describing the factor structure underlying a cognitive assessment battery is important in providing evidence for the construct validity of the cognitive assessment. The most prominent and current theory of intelligence, the Cattell–Horn–Carroll (CHC) model, is a factor analytic model which consists of broad and narrow factors or abilities, with the broad abilities described as the most relevant for neuropsychological assessment (Jewsbury, Bowden, & Strauss, 2016; Schneider & McGrew, 2018). Analysis of cognitive data sets has shown that the CHC framework is consistently found across different tests and populations (Agelink van Rentergem et al., 2020; Caemmerer, Keith, & Reynolds, 2020; Gross, Khobragade, Meijer, & Saxton, 2020; Jewsbury, Bowden, & Duff, 2016; Jewsbury, Bowden, & Strauss, 2016; Reynolds, Keith, Flanagan, & Alfonso, 2013; Schneider & McGrew, 2018) and the CHC model has been proposed as a common nomenclature for describing theoretical frameworks of human cognitive abilities (McGrew, 2009). When investigating the generalizability of cognitive tests across cultures, it is of interest to determine whether invariant measures may be described by CHC broad abilities, although original study authors may not have used this nomenclature. Such evidence would provide additional support for the generality of cognitive assessments (Bryan & Mayer, 2020).

1.4. Aims of the current study

Evidence suggests that culturally adapted cognitive ability test scores are reliable and valid measures (van de Vijver & Leung, 1997). However, with significant population movements, increased English-language-learners in English-speaking countries, rising refugee populations, and larger numbers of ethnic minorities, there is a growing concern that assessments may misestimate the cognitive ability of individuals from cultures in which the assessment was not developed (Wicherts, 2016). Comparing diverse cultures requires confidence in the generalizability of the construct validity evidence. Demonstrating invariance of the measurement model underlying the cognitive assessment enables clinicians and researchers to generalize validity information from research across populations (Bowden, Saklofske, & Weiss, 2010). Despite the theoretical importance of this question, to date, there has been no systematic review of the factorial invariance of assessments of cognitive ability across cultures.

The first aim of this review was to critically evaluate the current evidence relating to the factorial invariance of cognitive assessments across cultures. To this end, this paper will present a summary of the current literature reporting factorial invariance findings and the associated analytic approaches. The second aim of this study is to evaluate the quality of the factorial invariance analytic approaches reported in published studies and to evaluate whether the analytic approach to factorial invariance influences the outcome of the analysis. Lastly, this review will describe the factor structure of the studies to evaluate the utility of the CHC model.

2. Method

The current study was registered with the international register for systematic review, PROSPERO, registration number CRD42018091461. The search strategy employed in this review was guided by the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Moher et al., 2009). This review included published peer-reviewed journal articles, books, and dissertations using PubMed and PsychINFO (Ovid) databases, with the search performed in November 2022 (see Appendix in supplementary materials). The date range was limited to 1993 to November 2022, as this earlier date includes some of the first accessible empirical studies that report the details of the measurement invariance analysis (Horn & McArdle, 1992; Meredith, 1993). For earlier exposition and description of factorial invariance analyses, see McGaw and Jöreskog (1971) and Sörbom (1974), and for earlier examples of factorial invariance of covariance structure models, see Meredith (1964a), Meredith (1964b), and Joreskog (1970).

The search strategy was structured with two main criteria for study inclusion. The first criterion required cross-cultural measurement using any assessment of cognitive ability. The second criterion required factorial invariance analysis across cultural groups. Search fields were limited to abstracts, keywords, and titles of publications. The search strategy only included studies published in English. See the Appendix for the full search strategy.

References were imported for screening using the Covidence systematic review management platform (www.covidence.org). All article titles and abstracts were double-screened for eligibility by two authors (CW and WM). Any study deemed potentially relevant by one author but not the other was included in the full-text assessment for eligibility. The full-text reviews were undertaken by the same two authors, and any conflicts on the inclusion of a study were decided by a third author (SB). Studies that met the inclusion criteria after full-text reviews were critically reviewed by two authors (CW and SB) to determine the quality of the analysis.

Any study that deviated from the hierarchical approach to factorial invariance outlined in Table 1 and Fig. 1 was noted for further evaluation. For each study, the cognitive ability measure, the comparison groups, the sample characteristics, the hierarchical order of the factorial invariance analysis, the interpretable results, and the reported factorial invariance result were extracted and are listed in Table 2.

After studies were extracted for inclusion, we sought to classify the factor names in the original studies to the CHC model (see Table 3). Original study assessment details were collected, including the names of the assessment instruments, the original factor names reported in the factorial invariance analysis, descriptions of the original factors, the factor-indicator loading pattern, and descriptions of the test-indicators. Using the definitions found in Table 4 below, we adopted a similar methodology to Jewsbury, Bowden, and Duff (2016), who successfully classified a priori over 250 test scores using the CHC model without the need for any post hoc revision. We classified each original study factor to a CHC broad ability. For example, the current empirical evidence on the psychological construct of ‘Executive Function’ suggests that most factors assessing Executive Function may be described in the CHC model as Gs or Gwm, depending on the test-indicators used (Jewsbury, Bowden, & Strauss, 2016).

Classification under the CHC terminology was supported by the definitions of the empirically verified neuropsychological tests listed by Jewsbury, Bowden and Duff (2016, Table 5 pp 558). We also used the assessment recommendations for each CHC narrow ability and corresponding broad ability described by Schneider and McGrew (2018) and the detailed operational definitions at *www.iqscorner.com/2017/07/cattell-horn-carroll-chc-theory-of.html* (last accessed, Oct 11, 2022). Complex factors were classified into their most likely predominant CHC broad abilities for the purpose of this study. Full agreement on CHC alignment was established independently by the first and second authors, and differences in 10 out of 216 factor definitions (5%) were resolved by consensus.

3. Results

3.1. Search summary

The combined search identified 5747 studies for potential inclusion in the review (PsychINFO, $n = 3909$, PubMed, $n = 3125$), with 1287 omitted as duplicates. A total of 5397 studies were excluded during initial screening for not investigating the factorial invariance of a cognitive assessment across cultures. In total, 350 articles were reviewed in full text for inclusion. Of these, 296 were excluded for reasons noted in the PRISMA flow chart, see Fig. 2. In total, 54 articles consisting of 57 original empirical studies were included in the systematic review.

3.2. Sample characteristics

A total pool of 125,576 participants was included in the review, with sample characteristics consisting of standardization or norming data sets ($n = 22$), community samples ($n = 19$), student groups ($n = 9$), nationally representative samples ($n = 5$), and clinical representative or consecutive samples ($n = 4$). See Table 2 for the list of assessments and comparison groups. For clarity, we will use the original study author’s terminology for the description of race, culture, or ethnicity.

3.3. Cross-cultural factorial invariance

Critical evaluation of the included 57 studies found that 39 studies used the hierarchical approach to the evaluation of factorial invariance (see Table 2). Of the 18 studies that deviated from the hierarchical approach, 16 were able to be interpreted to some degree in terms of the hierarchical approach, making 55 interpretable studies in total. All 57 studies are listed in Table 2.

The results from the interpretable studies were highly supportive of the cross-cultural generalizability of the measurement models underlying cognitive assessment batteries. All 55 interpretable factorial invariance studies found a level of invariance across cultures. Ten studies were interpreted as finding configural invariance across cultures, 20 were interpreted as finding weak or partial weak factorial invariance across cultures, 12 as finding strong or partial strong factorial invariance across cultures, and 13 as finding strict factorial invariance across cultures (see Table 2). Note that 13 studies did not undertake a strict factorial invariance analysis but found weak or strong factorial invariance. As noted above, strict factorial invariance is not necessary for the generalization of construct validity across cultures.

Notably, all but one of the studies which was interpreted as finding strong or strict factorial invariance analyzed samples that were from the same country (e.g., Whites & African Americans) or culturally similar samples (e.g., U.S. and Australia). The one exception examined factorial invariance across U.S. and Italian samples (Naglieri, Taddei, & Williams, 2013). Conversely, 14 of the 30 studies which were interpreted as finding configural or weak factorial invariance analyzed culturally diverse samples (e.g., US & China, or German & North American). The results thus suggest that culturally similar samples are more likely to find strong or strict factorial invariance. The lack of detail in reported studies precludes the evaluation of alternative explanations for this pattern of results which may arise because of differences in construct validity or because of differences in test score means giving rise to differences in intercept values. The latter would lead to the failure of strong factorial invariance but has quite different implications for the generalization of constructs to the former. Only further detailed studies will illuminate this question.

3.4. The hierarchical approach to factorial invariance

Two studies were interpreted as ambiguous or incomplete. One factorial invariance analysis was classified as incomplete as it described a partial test of structural invariance (Edwards & Oakland, 2006). The second study that was interpreted as ambiguous described a two-step approach whereby the authors fitted a covariance model with unrestricted means and fixed factor variances across groups to a fully factorial invariance model where the intercepts were set to equal across groups, and the factor means fixed to zero in one group and estimated in the other (Lubke, Dolan, Kelderman, & Mellenbergh, 2003). This study was not described in sufficient detail to interpret evidence of invariance.

Other interpretable studies included two that reported factorial invariance results, which were reclassified for the purpose of the current review from weak factorial invariance to configural invariance to conform to the hierarchical approach (Holding et al., 2018; Mungas, Reed, Crane, Haan, & Gonzalez, 2004). Another study was reclassified from strict factorial invariance to finding weak factorial invariance as the authors demonstrated factor loading invariance at the final two steps described in the analysis, with the prior steps of the factorial invariance analysis differing from the hierarchical sequence (Keith, Quirk, Schartzer, & Elliott, 1999). A further seven studies, reported from six articles that undertook the factorial invariance analysis differing from the hierarchical sequence, were able to be interpreted in line with the authors’ conclusions (Dolan, 2000; Dolan & Hamaker, 2001; Dolan, Roorda, & Wicherts, 2004; Keith et al., 1995; Pandolfi, 1998; Wicherts & Dolan, 2010). The authors of one study concluded a lack of factorial invariance, however, this result was reclassified to configural invariance as the

Table 2
Summary Table of Included Studies. For clarity the sample description reported by original study authors is reiterated here.

Reference	Assessment	Groups	Sample Characteristics	Analysis in Hierarchical Order	Reported Result	Interpreted Result
Avila et al. (2019)	Neuropsychological test battery	Non-Hispanic White, Black, & Hispanic	Community sample	Yes	Partial strong invariance	Partial strong invariance
Barnes et al. (2016)	18-test cognitive battery	White & African American	Older adults without known dementia	Yes	Strict invariance	Strict invariance
Beaujean and McGlaughlin (2014)	RIAS	Black & White	Referred students for special education services	Yes	Strict invariance	Strict invariance
Benson, Kranzler, and Floyd (2018)	UNIT-2	Asian or Pacific Islanders, Black, & White	Norming samples	Yes	Configural invariance	Configural invariance
Benson et al. (2018)	UNIT-2	Hispanic & non-Hispanic	Norming samples	Yes	Weak invariance	Weak invariance
Bertola et al. (2020)	Neuropsychological test battery	White & Non-white	Community samples	Yes	Strong invariance	Strong invariance
Blankson and McArdle (2015)	6-test cognitive battery	Black, White, & Hispanic	Nationally representative older samples	Yes	Strict invariance	Strict invariance
Bowden, Lange, et al. (2008)	WAIS-III	U.S. & Canadian	Standardization samples	Yes	Strict invariance	Strict invariance
Bowden, Lissner, McCarthy, Weiss, and Holdnack (2007)	WAIS-III	U.S. & Australian	Standardization sample & general community sample	Yes	Strict invariance	Strict invariance
Bowden, Saklofske, and Weiss (2011)	WAIS-IV	U.S. & Canadian	Standardization samples	Yes	Strict invariance	Strict invariance
Bowden, Weiss, Holdnack, Bardenhagen, and Cook (2008)	WAIS-III	U.S. & Australian	Standardization sample & clinical sample	Yes	Strict invariance	Strict invariance
Bowden et al. (2010)	WAIS-IV	U.S. & Canadian	Standardization samples	Yes	Strict invariance	Strict invariance
Chen et al. (2010)	WISC-IV	China, Hong Kong, Macau, & Taiwan	Census matched samples	Yes	Strict invariance	Strict invariance
Cockcroft, Alloway, Copello, and Milligan (2015)	WAIS-III	UK & South African	University students	Yes	Configural invariance	Configural invariance
Damas (2002)	WISC-R	U.S. & China	Standardization samples	No	Lacks measurement invariance	Configural invariance
Deng and Georgiou (2015)	D-N CAS	Canada & China	Community samples	Yes	Weak invariance	Weak invariance
Dolan (2000)	WISC-R	Whites & Blacks	Stratified samples	No	Strict invariance	Strict invariance
Dolan and Hamaker (2001)	WISC-R & K-ABC	Whites & Blacks	Matched pair	No	Strict invariance	Strict invariance
Dolan et al. (2004)	GATB	Dutch, Surinamese, Dutch Antilleans, North Africans, & Turks	Blue collar job applicants	No	Weak invariance	Weak invariance
Dolan et al. (2004)	JAT	Whites, Blacks, & Indians	Adolescents	No	Configural invariance	Configural invariance
Edwards and Oakland (2006)	WJ III	Caucasian American & African American	Standardization samples	No	Configural invariance	Incomplete
Gavett, Stypulkowski, Johnson, Hall, and O'Bryant (2018)	Neuropsychological Test Battery	Hispanic (English Language), Hispanic (Spanish Language), & non-Hispanic	Community sample	Yes	Strict invariance	Strict invariance
Gygi, Fux, Grob, and Hagmann-von Arx (2016)	RIAS	Germans with and without a Migration Background	Standardization sample	Yes	Strong invariance	Strong invariance
Hajovsky and Chesnut (2022)	WJ IV	Non-Hispanic & Hispanic	Standardization sample	Yes	Strong invariance	Strong invariance
Hajovsky and Chesnut (2022)	WJ IV	Asian, African American, & Caucasian	Standardization sample	Yes	Strong invariance	Strong invariance
Holding et al. (2018)	Cognitive Battery	Tanzania, Ghana, & Bangladesh	Relative poverty/rural sample	No	Configural invariance	Weak invariance
Karr, Scott, Aghvinian, and Rivera Mindt (2022)	NIH Cognitive Battery	English-speaking & Spanish-speaking (Latinx)	Standardization sample	Yes	Strict invariance	Strict invariance
Keith et al. (1995)	K-ABC	Black & White	Standardization sample & general community sample	No	Weak invariance	Weak invariance
Keith et al. (1999)	DAS	Black, White, & Hispanic US Children	Standardization and bias oversample	No	Strict invariance	Weak Invariance
Khoo (2021)	MoCA	White non-Hispanic, Black, non-Hispanic, & Hispanic	Older adult sample	Yes	Weak invariance	Weak invariance

(continued on next page)

Table 2 (*continued*)

Reference	Assessment	Groups	Sample Characteristics	Analysis in Hierarchical Order	Reported Result	Interpreted Result
Reference	Assessment	Groups	Sample Characteristics	Analysis in Hierarchical Order	Reported Result	Interpreted Result
Lakin (2012)	CogAT, Form 6	Hispanic ELL, Hispanic non-ELL, & non-Hispanic	Students	No	Weak invariance	Weak invariance
Lubke et al. (2003)	Primal Mental Ability Test	African & Caucasian Americans	Students	No	Strict invariance	Ambiguous
Mungas et al. (2004)	SENAS	Caucasian & Hispanic (English Language), Hispanic (Spanish Language)	Community dwelling	No	Weak invariance	Configural invariance
Mungas et al. (2011)	SENAS	White, African American, Hispanic (English Language) & Hispanic (Spanish Language)	Community dwelling	Yes	Partial weak invariance	Partial weak invariance
Naglieri et al. (2013)	CAS	USA & Italian	Matched standardization samples	Yes	Strict invariance	Strict invariance
Neumann, Peterson, Underwood, Morton, and Waldie (2021)	Cognitive Battery	European, Māori, Pacific Peoples, & Asian	Community sample of Children	Yes	Configural invariance	Configural invariance
Omura and Sugishita (2004)	WMS-R	USA & Japanese	Standardization sample	No	Configural invariance	Configural invariance
Pandolfi (1998)	WISC-III	White, Black, & Hispanic	Standardization sample	No	Measurement invariance	Weak invariance
Papadopoulos, Georgiou, Deng, and Das (2018)	D-N CAS	Canada, China, & Cyprus	University students	Yes	Weak invariance	Weak invariance
Rawlings et al. (2016)	Neuropsychological Battery	White & Black	Community based	Yes	Configural invariance	Configural invariance
Reference	Assessment	Groups	Sample Characteristics	Analysis in Hierarchical Order	Reported Result	Interpreted Result
Reed (2001)	WISC-III	African American & Caucasian Americans	Students identified for assessment	Yes	Weak invariance	Weak invariance
Reverte et al. (2015)	WISC-IV	French & French speaking Swiss	Standardization sample & stratified sample	Yes	Partial weak invariance	Partial weak invariance
Rummel et al. (2019)	Working Memory Capacity Battery	German and North American	University Students	Yes	Weak invariance	Weak invariance
Rushton, Skuy, and Bons (2004)	Raven's Advanced Progressive Matrices	African & non-African South Africans	First-year University Students	No	Weak invariance	Weak invariance
Scheiber (2016a)	KABC-II & KTEA-II	Caucasian, Black, & Hispanic	Standardization sample	Yes	Strong invariance	Strong invariance
Scheiber (2016b)	WISC-V	Caucasian, African American & Hispanic	US Norming sample	Yes	Strong invariance	Strong invariance
Siedlecki et al. (2010)	Neuropsychological Battery	English Speakers & Spanish Speakers	Community based healthy adults	Yes	Partial strong invariance	Partial strong invariance
te Nijenhuis, Evers, and Mur (2000)	DAT	Majority & Minority Dutch Groups	Secondary School Students	No	Partial weak invariance	Partial weak invariance
te Nijenhuis, Tolboom, Resing, and Bleichrodt (2004)	RAKIT	Dutch & Immigrant Groups	Primary School Children	No	Weak invariance	Weak invariance
Tennant et al. (2022)	NIH Cognitive Battery	Jamaican & African-American	Adult community sample	Yes	Weak invariance	Weak invariance
Trundt et al. (2017)	DAS-II	African American, Asian, Hispanic, & Caucasian US Children	Standardization Sample	Yes	Partial strong invariance	Partial strong invariance
Reference	Assessment	Groups	Sample Characteristics	Analysis in Hierarchical Order	Reported Result	Interpreted Result
Tuokko et al. (2009)	Neuropsychological battery	Canadian English & Canadian French	Stratified random, community dwelling & institutional older adults	Yes	Partial strong invariance	Partial strong invariance
White and Greenfield (2017)	Executive Functioning Battery	US Spanish-English, English, & Spanish language	Low SES Children enrolled in Head Start	Yes	Strong invariance	Strong invariance
Wicherts and Dolan (2010)	RAKIT	Immigrants from Turkey, Morocco, & the former Dutch colonies.	Immigrant children	No	Weak invariance	Weak invariance
Woods (2017)	WJ-IV	Black, Hispanic, White	Standardization sample	Yes	Weak invariance	Weak invariance
Wray et al. (2020)	NIH Cognitive Battery	Guatemala, Philippines, & South Africa	Adult community sample	Yes	Configural invariance	Configural invariance
Xu, Ellefson, Ng, Wang, and Hughes (2020)	Executive Functioning Battery	China, Hong Kong, & UK	Children community sample	Yes	Partial strong invariance	Partial strong invariance

Note. RIAS = Reynolds Intellectual Assessment Scales; UNIT = Universal Nonverbal Intelligence Test; WAIS = Wechsler Adult Intelligence Scale; WISC = Wechsler Intelligence Scale for Children; D—N CAS = Das-Naglieri Cognitive Assessment System; GATB = General Aptitude Test Battery; JAT = Junior Aptitude Test; W-J = Woodcock-Johnson; DAS = Differential Ability Scales; CogAT = Cognitive Abilities Test; SENAS = Spanish and English Neuropsychological Assessment Scales; CAS = Cognitive Assessment System; WMS = Wechsler Memory Scale; KABC = Kaufman Assessment Battery for Children; KTEA = Kaufman Test of Educational Achievement; DAT = Differential Aptitude Test; RAKIT = Revisie Amsterdamse Kinder Intelligentie Test; MoCA = Montreal Cognitive Assessment; NIH = National Institute on Aging.

study described undertaking a configural invariance two-factor model showing good fit to the data (Damas Jr., 2002).

A chi-square test was performed to examine the relationship between studies using the hierarchical approach to factorial invariance and study findings of strong or strict factorial invariance (i.e., the level at which factor covariances and means, respectively, can be compared across cultures). The relationship was significant, X^2 $(1, N = 57) = 12.62, p < .001$, suggesting that the use of the hierarchical approach to factorial invariance outlined in Table 1 was more likely to lead to finding strong or strict factorial invariance.

3.5. Aligning reported factors to the Cattell-Horn-Carroll Model

Referring to Table 3, it is evident that the studies assessed a wide range of cognitive ability factors, and the CHC correspondence column shows the factors identified using the CHC taxonomy. A total of 216 factors reported across the 57 studies could readily be identified as a well-established and empirically supported CHC construct. Seven studies reported factors that directly aligned with the CHC framework (Edwards & Oakland, 2006; Hajovsky & Chesnut, 2022; Reverte, Golay, Favez, Rossier, & Lecerf, 2015; Scheiber, 2016a, 2016b; Trundt, Keith, Caemmerer, & Smith, 2017; Woods, 2017). A total of 11 CHC broad abilities were measured in addition to 'g'. The most measured CHC broad abilities were Gc ($n = 39$), Gf and Gwm ($n = 38$), Gv ($n = 36$), Gs ($n = 31$), and Gl ($n = 18$). The other CHC broad abilities measured were Ga ($n = 3$), Gq and Gr ($n = 2$), and Gp and Grw ($n = 1$), while a single Spearman's 'g' factor, or similar global cognitive ability score, was measured seven times.

The results of this review, therefore, showed that a range of CHC constructs are generalizable across cultures. Further, the results demonstrate that post hoc alignment to the CHC model can be successfully reported alongside the original test nomenclature. Overall, the fact that we were able to readily classify reported factors in terms of CHC constructs reflects the comprehensive reach of the CHC model, a classification process that we report to encourage scrutiny and replication.

4. Discussion

The aim of this study was, firstly, to systematically review the evidence for factorial invariance of cognitive ability tests across cultures. The second aim was to critically review approaches to the evaluation of factorial invariance in the 57 studies published from 1993 to 2022 that sought to examine the generalizability of cognitive ability models across cultures. The findings of this review were highly supportive of the generalizability of cognitive ability constructs across cultures. Ten studies found configural invariance, which is the most basic form of factorial invariance and is sufficient for theoretical and applied claims of similar construct measurement (Horn & McArdle, 1992; Meredith, 1993; Meredith & Teresi, 2006; Widaman & Reise, 1997). Further, 20 studies found weak or partial weak factorial invariance, 12 found strong or partial strong factorial invariance, and 13 found strict factorial invariance. No study that provided interpretable results compared to the hierarchical sequence found a lack of invariance.

When the hierarchical sequence of factorial invariance analysis, outlined above (Table 1), was applied to all the studies in the review, 55 studies were able to be interpreted in terms of conformity to the hierarchical approach. Evaluation of these studies found that 18 undertook the factorial invariance analysis in a sequence that differed from the hierarchical order (see Table 2). However, 16 of the 18 studies which deviated from the hierarchical approach nevertheless could be interpreted with confidence. Overall, the evidence from this review suggests that the use of the hierarchical approach to factorial invariance is more likely to find strong or strict factorial invariance when comparing cognitive assessment scores across cultures.

A failure to find strict factorial invariance has been interpreted as a failure of the generalization of cognitive constructs across populations by some authors (DeShon, 2004; Wu, Li, & Zumbo, 2007). The finding of strict factorial invariance implies that the residual item variances (uniqueness and item reliability*)* are equivalent across groups and that common factors are the cause of any group differences in means and variances. However, it is unreasonable to expect strict factorial invariance across all populations and is not necessary for the generalization of construct validity evidence (Horn & McArdle, 1992; Millsap, 2011). Variance in the residuals, tested at the strict invariance level, may reflect the real-world differences in the unique or reliable variance components across groups (Mungas, Widaman, Reed, & Tomaszewski Farias, 2011). Further, if the cognitive assessment has not been adapted or standardized for any particular population or culture, test items can have different meanings and therefore be measuring different content as a consequence of cultural differences, socioeconomic status, or education (Van de Vijver, 1997). Importantly, given the restrictive parameters that a strict factorial invariance analysis places on the models, it was notable that 13 studies did find evidence for strict factorial invariance across cultures (see Table 2). This is more surprising as 13 studies did not undertake a strict factorial invariance test as it was deemed unnecessary, a view in line with recent conventions and reporting (Putnick & Bornstein, 2016).

One hypothesis may be that weaker invariance (i.e., configural invariance or weak factorial invariance) findings are associated with studies comparing more culturally distant or discrepant groups. Indeed, the data in this review suggests that there are patterns of weaker invariance and greater cultural distance between the samples. However, many weaker invariance findings also involved samples from the same norming data set suggesting that a finding of configural invariance or weak factorial invariance may additionally be reflective of i) the factorial invariance methodology, ii) the sample characteristics (e.g., representative standardization sample versus convenience community sampling), and iii) the cognitive tool itself, rather than differences between the cross-cultural samples.

The studies included in this review reported a wide variety of fit indexes, often employing different criteria to identify the acceptability of fit. While every study included in this review reported chi-square, this has been shown to be overly sensitive to misfit when applied in large samples such as those required to meaningfully undertake factorial invariance analysis (Meade et al., 2008; Vandenberg & Lance, 2000). Therefore, approaches that emphasize increments in fit indices may be less likely to reject the null hypothesis of invariance. As a result, it has been recommended to report specific fit indices such as the RMSEA and the change in CFI and Mc (Hu, & & Bentler, P. M., 1999; Meade et al., 2008) or use permutation tests (Jorgensen et al., 2018). Additionally, sample size varied across studies, with some samples smaller than recommended (see Wolf, Harrington, Clark, & Miller, 2013) for adequate power to detect a lack of invariance (Dolan & Hamaker, 2001; Lubke et al., 2003; White & Greenfield, 2017). An alternative method to the evaluation of factorial invariance is the concept of approximate factorial invariance, which uses Bayesian Structural Equation Modelling. Only one study in this review undertook Bayesian Structural Equation Modelling (Rummel, Steindorf, Marevic, & Danner, 2019).

An additional aim of the review was to employ the CHC taxonomy to classify factors identified in the included studies. If a comprehensive

Table 3
CHC Correspondence with Factors Reported from Original Studies.

Author(s)	Assessment	Factor Structure	Factors Named in Study	CHC equivalent
Avila et al. (2019)	Neuropsychological battery	3 factor model	Memory, Language, Visuo-Spatial	Gl, Gc, Gv/Gf
Barnes et al. (2016)	Cognitive battery	5 factor model	Semantic Memory, Working Memory, Perceptual Speed, Visuo Spatial Abilities, Episodic Memory;	Gc, Gwm, Gs, Gv/Gf, Gl,
Beaujean and McGlaughlin (2014)	RIAS	1 factor	g	g
Benson et al. (2018)	UNIT-2	Bifactor with g and 3 factors	Memory, Quantative, Reasoning	Gwm, Gf, Gv,
Bertola et al. (2020)	Neuropsychological battery	2 factor model	Episodic Memory, Executive Function	Gl, Gc/Gs
Blankson and McArdle (2015)	Cognitive battery	2 factor model	Episodic Memory factor, Mental status factor	Gl, g,
Bowden, Lange, et al. (2008)	WAIS-III	4 factor model	Verbal Comprehension, Perceptual Organization, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Bowden et al. (2007)	WAIS-III	4 factor model	Verbal Comprehension, Perceptual Organization, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Bowden et al. (2010)	WAIS-IV	4 factor model	Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Bowden, Weiss, et al. (2008)	WAIS-III	4 factor model	Verbal Comprehension, Perceptual Organization, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Bowden et al. (2011)	WAIS-IV	4 factor model	Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Chen et al. (2010)	WISC-IV	4 factor model	Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Cockcroft et al. (2015)	WAIS-III	4 factor model	Verbal Comprehension, Perceptual Organization, Working Memory, Processing Speed	Gc, Gf/Gv, Gwm, Gs
Damas (2002)	WISC-R	2 factor model	Verbal factor, Performance factor	Gc, Gf/Gv
Deng and Georgiou (2015)	D–N CAS	4 factor model	Planning, Attention, Simultaneous, Successive	Gv/Gs, Gs, Gf, Gwm
Dolan (2000)	WISC-R	3 factor model	Verbal, Performance, Memory	Gc, Gf/Gv, Gwm
Dolan and Hamaker (2001)	WISC-R & K-ABC	3 factor model	Verbal, Memory, Spatial	Gc, Gwm, Gf/Gv
Dolan et al. (2004)	GATB	3 factor model	Fluid & Crystallized, Visual, Speed	Gc/Gf, Gv, Gwm
Dolan et al. (2004)	JAT	3 factor model	Factor 1 (Spatial), Factor 2 (Memory), Factor 3 (Language)	Gf/Gv, Gwm, Gc,
Edwards and Oakland (2006)	W-J III	7 factor model	Gf, Gc, Glr, Gsm, Ga, Gs, Gv	Gf, Gc, Gl, Gwm, Ga, Gs, Gv

(continued on next page)

Table 3 (*continued*)

Author(s)	Assessment	Factor Structure	Factors Named in Study	CHC equivalent
Gavett et al. (2018)	Neuropsychological Test Battery	5 factor model	Memory, AEPS (attention, executive functioning, and processing speed) Language, Visuospatial, Motor	Gl, Gs/Gwm Gc, Gv, Gp
Gygi et al. (2016)	Reynolds Intellectual Assessment Scales German (RIAS)	2 factor model	Verbal, Non-verbal	Gc, Gf
Hajovsky and Chesnut (2022)	WJ-IV	7 factor model	Gc, Gf, Gwm, Gs, Ga, Glr, Gv	Gc, Gf, Gwm, Gs, Ga, Gl, Gv
Author(s)	Assessment	Factor Structure	Factors Named in Study	CHC equivalent
Holding et al. (2018)	Battery measuring General Intelligence, Executive Function and Achievement	1 factor model	Single factor of General Intelligence, Executive Function and Achievement	*g*
Karr et al. (2022)	Cognitive Battery	2 factor model	Verbal, Non-verbal	Gc, Gf
Keith et al. (1995)	K-ABC	2 factor model	Simultaneous, Successive	Gf, Gwm
Keith et al. (1995)	DAS	1 factor model	*g*	*g*
Khoo (2021)	MoCA	1 factor model	*Unnamed*	*g*
Lakin (2012)	CogAT, Form 6	3 factor model	Verbal, Quantitative, Non-verbal	Gc, Gq, Gf
Lubke et al. (2003)	Primal Mental Ability Test	1 factor model	*unnamed*	*g*
Mungas et al. (2004)	Spanish and English Neuropsychological Assessment Scales (SENAS	6 factor model	Conceptual thinking, Semantic memory, Attention span, Episodic memory, Non-verbal spatial abilities, Verbal abilities.	Gf, Gc, Gwm, Gl, Gv, Gc
Mungas et al. (2011)	Spanish and English Neuropsychological Assessment Scales (SENAS	5 factor model	Episodic mem, Semantic/language, Spatial, Attention/working memory, Fluency	Gl, Gc, Gf/Gv, Gwm, Gc/Gs
Naglieri et al. (2013)	CAS	4 factor model	Planning, Attention, Simultaneous, Successive	Gv/Gs, Gs, Gf, Gwm
Author(s)	Assessment	Factor Structure	Factors Named in Study	CHC equivalent
Omura and Sugishita (2004)	WMS-R	3 factor model	Attention/Concentration, Immediate Memory, Delayed Recall	Gwm, Gl, Gl,
Pandolfi (1998)	WISC-III	4 factor model	Processing Speed, Freedom From Distractibility, Perceptual Organization, Verbal Comprehension	Gs, Gwm, Gf/Gv, Gc
Papadopoulos et al. (2018)	D-N CAS	4 factor model	Planning, Attention, Simultaneous Processing, Successive Processing	Gv/Gs Gs, Gf, Gwm
Rawlings et al. (2016)	ARIC-NCS	3 factor model	Memory, Language, SAPS (sustained attention and processing)	Gl, Gc, Gs/Gwm
Reed (2001)	WISC-III	2 factor model	Verbal, Performance	Gc, Gv
Reverte et al. (2015)	WISC-IV	5 factor model	Gc, Gv, Gf, Gsm, Gs	Gc, Gv, Gf, Gwm, Gs
Rummel et al. (2019)	Working Memory Capacity Battery	1 factor model	Working-memory capacity	Gwm
Rushton et al. (2004).	Raven's Advanced Progressive Matrices	1 factor model	*g*	Gf/Gv

(*continued on next page*)

Table 3 (*continued*)

Author(s)	Assessment	Factor Structure	Factors Named in Study	CHC equivalent
Scheiber (2016a)	KABC-II & KTEA-II	7 factor model	Grw, Gq, Gc, Glr, Gv, Gf, Gsm	Grw, Gq, Gc, Gl, Gv, Gf, Gwm
Scheiber (2016b)	WISC-V	5 factor model	Gc, Gv, Gf, Gsm, Gs	Gc, Gv, Gf, Gwm, Gs
Siedlecki et al. (2010)	Neuropsychological Battery	4 factor model	Memory, Language, Visual-spatial, Processing speed	Gl, Gc, Gv, Gs
te Nijenhuis et al. (2000)	DAT	3 factor model	Crystallized, Fluid, Broad Visual Perception	Gc, Gf, Gv
te Nijenhuis et al. (2004)	RAKIT	4 factor model	Hybrid factor, Visual factor, Memory factor, Retrieval factor	Gc/Gf, Gv, Gwm/Gl, Gr
Tennant et al. (2022)	NIH Toolbox Cognitive Battery	1 factor model	Fluid cognition	*g*
Trundt et al. (2017)	DAS-II	6 factor model	Gc, Gf, Gv, Glr, Gsm, Gs	Gc, Gf, Gv, Gl, Gwm, Gs
Tuokko et al. (2009)	CSHA neuropsychological battery	3 factor model	Verbal Ability, Visuospatial Ability, Long-term Retrieval	Gc, Gv, Gl
Wicherts and Dolan (2010)	RAKIT	4 factor model	Hybrid factor, Visual factor, Memory factor, Retrieval factor	Gc/Gf, Gv, Gwm/Gl Gr
Woods (2017)	WJ-IV	5 factor model	Ga, Gc, Gs, Gsm, Gf	Ga, Gc, Gs, Gwm, Gf
Wray et al. (2020)	NIH Cognitive Battery	1 factor model	Executive function	Gf/Gs/Gwm
Xu et al. (2020)	Executive Functioning Battery	1 factor model	Executive function	Gf/Gs/Gwm

Note. g = general intelligence; Gf = fluid reasoning; Gc = comprehension-knowledge; Gv = visuospatial processing; Gwm = working memory capacity; Gl = learning efficiency; Gr = retrieval fluency; Gs = processing speed; Gps = psychomotor speed; RIAS = Reynolds Intellectual Assessment Scales; UNIT = Universal Nonverbal Intelligence Test; WAIS = Wechsler Adult Intelligence Scale; WISC = Wechsler Intelligence Scale for Children; D—N CAS = Das-Naglieri Cognitive Assessment System; GATB = General Aptitude Test Battery; JAT = Junior Aptitude Test; W-J = Woodcock-Johnson; DAS = Differential Ability Scales; CogAT = Cognitive Abilities Test; SENAS = Spanish and English Neuropsychological Assessment Scales; CAS = Cognitive Assessment System; WMS = Wechsler Memory Scale; KABC = Kaufman Assessment Battery for Children; KTEA = Kaufman Test of Educational Achievement; DAT = Differential Aptitude Test; RAKIT = Revisie Amsterdamse Kinder Intelligentie Test; MoCA = Montreal Cognitive Assessment; NIH = National Institute on Aging.

taxonomy is available, it should help to systematize an approach to invariance studies and encourage future researchers to adopt a systematic taxonomy. Because most authors of studies included in this review did not aim to measure CHC constructs in any systematic or comprehensive way, the range of CHC constructs reported is limited and ad hoc. Nevertheless, to the extent the studies covered CHC abilities, a wide range of well-established constructs were shown to be invariant across cultures. No CHC constructs were shown to consistently fail a test of factorial invariance, including general ability measures aimed at evaluating Spearman's '*g*' or similar global ability scores. Despite the post hoc application of the CHC taxonomy to reported factors, the reviewed studies provide a demonstration of the utility of the CHC constructs because the CHC model could be applied to all the results and by inference, the invariance of the CHC constructs across cultures. This review also lends support to the results of a recent reanalysis of non-Western, non-industrialized nation samples, which used exploratory factor analysis to conclude that Spearman's 'g' is evident across diverse cultures (Warne & Burningham, 2019).

It should be noted that the conclusions of the current review regarding the level of factorial invariance achieved in the individual studies are based on the factorial invariance modelling results reported by the original study authors. While we reinterpreted the results of some studies to conform to the hierarchical approach to factorial invariance analysis, all interpretations were based on the original analytic models reported in each study. No reanalysis of the original data was performed. Therefore, it is possible that further analysis of respective study data would show stronger levels of factorial invariance, for example, if models displaying better fit were found and compared across groups.

A further important implication of the finding of factorial invariance in most of the studies reviewed above relates to the interpretation of test score 'dissociations' in cognitive neuroscience and neuropsychological research. Dissociations, or differential patterns of performance on

Table 4
Summary Description of Cattell-Horn-Carroll Broad Abilities Names in Table 3.

CHC Broad Ability	Description
General Intelligence (*g*)	The positive correlation of all mental abilities.
Auditory Processing (Ga)	The ability to discriminate, remember, and reason creatively on auditory stimuli.
Comprehension-Knowledge (Gc)	The ability to comprehend and communicate culturally valued knowledge.
Visual Processing (Gv)	The ability to use mental imagery to perceive, discriminate, and manipulate mental images to solve problems.
Fluid Reasoning (Gf)	The use of deliberate and controlled mental processes requiring focused attention to solve novel problems.
Working Memory Capacity (Gwm)	The ability to maintain and manipulate information in active attention.
Processing Speed (Gs)	The ability to control attention to automatically, quickly, and fluently perform simple repetitive cognitive tasks.
Learning Efficiency (Gl)	The ability to learn, store, and consolidate new information.

Note. Descriptions of the CHC broad abilities are adapted from Schneider and McGrew (2018).

different cognitive tests by people assumed to be drawn from different populations, has been used as evidence to infer different patterns of cognitive constructs underlying test score performance in the different populations. Such dissociations are typically used to interpret patterns of test scores, for example, in patients with different patterns of brain lesions (Bates, Appelbaum, Salcedo, Saygin, & Pizzamiglio, 2003; Delis, Jacobson, Bondi, Hamilton, & Salmon, 2003; Martin & Allen, 2012; Shallice, 1988). Dissociations are typically defined as differential patterns of correlations between the same test scores in different populations or differential patterns of test performance in selected patients, in effect, extreme quadrants in bivariate scatter plots (Bates et al., 2003; Delis et al., 2003; Shallice, 1988).

However, the evidence of the generalizability of cognitive constructs, based on the common observation of factorial invariance, illustrates that cognitive-dissociation evidence may be ambiguous or even misleading (Bowden, 2004; Bowden, Gregg, et al., 2008; Widaman & Reise, 1997). Observed dissociations may reflect (a) real differences in cognitive constructs in different populations, or differences in (b) factor covariances or (c) item reliabilities (Bowden, Gregg, et al., 2008). Both (b) and (c) may be observed in the presence of factorial invariance across populations. Therefore, the latter two conditions (b and c) need not lead to inferences of different cognitive constructs across populations despite observations of 'dissociations.' In other words, the observation of

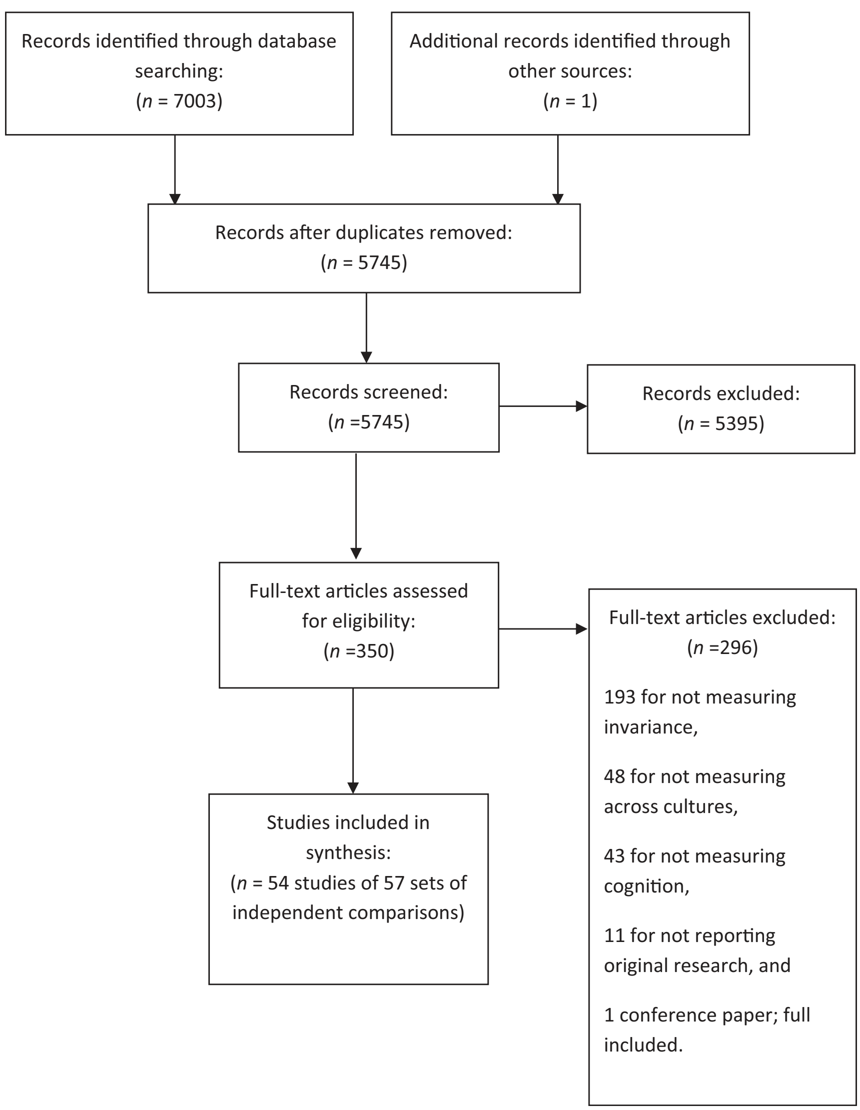

Fig. 2. PRISMA Flow Diagram.

dissociations in test scores across populations assumed to have different brain lesions or cognitive deficits cannot be used to infer different patterns of construct validity until factorial invariance has been formally examined and shown not to apply. In small-sample and single-case studies, where dissociations are assumed to indicate different patterns of construct validity, it is not possible to test factorial invariance, so scientific caution would encourage retention of the assumption of factorial invariance.

4.1. Future research

A key finding of this review is that a proportion of published studies on factorial invariance have not followed the hierarchical sequence testing of factorial invariance outlined above, which led to ambiguous results. Further, the included studies reported a wide variety of fit indexes and cut-off scores. The authors of this review recommend that future research use the hierarchical factorial invariance approach outlined in Table 1 and illustrated in Fig. 1, and first described by Widaman and Reise (1997), (cf. Horn & McArdle, 1992; Meredith, 1993; Meredith & Teresi, 2006; Vandenberg & Lance, 2000). It is also recommended that invariance testing studies adopt the fit guidelines outlined by Hu, and Bentler, P. M. (1999), Chen (2007) and Meade et al. (2008) to ensure consistency and interpretability of results.

Further, the authors of this review encourage the use of the CHC taxonomy to systematize research on cognitive abilities. A finding of this current study is that alignment to the CHC model can be successfully reported alongside original test nomenclature, supportive of previous research recommendation on the harmonization of cognitive ability nomenclature (Jewsbury, Bowden, & Duff, 2016). This process has resulted in a detailed and transparent classification of reported factors from the studies reviewed and is, in principle, no different to the methodology adopted in many systematic reviews where diverse outcomes and study characteristics are classified according to an objective taxonomy. The ease with which we were able to do this speaks to the comprehensive nature of the CHC model and should encourage the further application of this taxonomy for better scientific evaluation.

Future research should include more cross-national populations as cognitive tests become more widely used internationally. Additionally, while some research has investigated the factor structure and factorial invariance of specific subtests of cognitive ability measures (Bowden et al., 2013), future research could explore the generalizability of these specific subtests across cultures.

4.2. Conclusion

In this systematic review of studies published between 1993 and 2022, every study that examined factorial invariance across cultures found some level of invariance. Most commonly, weak or strong factorial invariance was observed. Therefore, researchers and clinicians can be confident that the cognitive assessments in the cross-cultural literature can be used to explore patterns of convergent and discriminant validity and that such research reflects broadly generalizable cognitive traits in the populations represented in this review.

Until factorial invariance has been demonstrated at the strong or strict level, however, comparison of observed mean scores should always be interpreted with the strong caveats of local linguistic, educational, socio-economic, health status and myriad other effects that may give rise to differences in test score attainment (Weiss & Saklofske, 2020). Although the results of this review are supportive of cross-cultural invariance and the generalizability of cognitive assessments, researchers and clinicians must use caution when comparing population means, as factorial invariance should not be assumed to apply in any new population until demonstrated with appropriate research. As noted, some studies included in this review did not use the hierarchical approach to factorial invariance and, as a consequence, led to ambiguous results and unwarranted speculation regarding the failure of factorial invariance in the populations examined. Future research into factorial invariance should follow the recommended hierarchical approach so as to minimize the risk of misinterpretation of findings.

Lastly, every factor measured in the current review could be successfully aligned a priori to a CHC broad ability. The results provide strong support of the generalizability of CHC constructs across cultures and support the continued use and development for the CHC model as a common nomenclature for researchers and test developers.

The current study was preregistered with the international register for systematic review, PROSPERO, registration number CRD42018091461; see [www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42018091461].

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The authors would like to acknowledge the helpful contribution of three reviewers.

Declaration of Competing Interest

Christopher J. Wilson, Nicole R. Joshua, & Lawrence G. Weiss were involved in the research and development of numerous psychological tests, including the Wechsler scales as employees of Pearson.

Data availability

This a review paper that used previously published results. The search code is available in the appendix.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2023.101751.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association. https://www.testingstandards.net/.

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., Staaks, J. P. C., Consortium, A., & Huizenga, H. M. (2020). The factor structure of cognitive functioning in cognitively healthy participants: a meta-analysis and meta-analysis of individual participant data. *Neuropsychology Review, 30*(1), 51–96. https://doi.org/10.1007/s11065-019-09423-6

Alkemade, N., Bowden, S. C., & Salzman, L. (2015). Scoring correction for MMPI-2 Hs scale with patients experiencing a traumatic brain injury: A test of measurement invariance. *Archives of Clinical Neuropsychology, 30*(1), 39–48. https://doi.org/10.1093/arclin/acu058

Avila, J. F., Rentería, M. A., Witkiewitz, K., Verney, S. P., Vonk, J. M. J., & Manly, J. J. (2019). Measurement invariance of neuropsychological measures of cognitive aging across race/ethnicity by sex/gender groups. *Neuropsychology.*. https://doi.org/10.1037/neu0000584

Barnes, L. L., Yumoto, F., Capuano, A., Wilson, R. S., Bennett, D. A., & Tractenberg, R. E. (2016). Examination of the factor structure of a global cognitive function battery across race and time. *Journal of the International Neuropsychological Society, 22*(1), 66–75. https://doi.org/10.1017/S1355617715001113

Bates, E., Appelbaum, M., Salcedo, J., Saygin, A. P., & Pizzamiglio, L. (2003). Quantifying dissociations in neuropsychological research. *Journal of Clinical and Experimental Neuropsychology, 25*(8), 1128–1153. https://doi.org/10.1076/jcen.25.8.1128.16724

Beaujean, A. A., & McGlaughlin, S. M. (2014). Invariance in the Reynolds intellectual assessment scales for black and White referred students. *Psychological Assessment, 26*(4), 1394–1399. https://doi.org/10.1037/pas0000029

Benson, N., Kranzler, J. H., & Floyd, R. G. (2018). Exploratory and confirmatory factor analysis of the universal nonverbal intelligence test-second edition: Testing dimensionality and invariance across age, gender, race, and ethnicity. *Assessment.* https://doi.org/10.1177/1073191118786584, 1073191118786584.

Bertola, L., Benseñor, I. M., Lotufo, P. A., Barreto, S. M., Moreno, A. B., Griep, R. H., … Suemoto, C. K. (2020). Measurement invariance of neuropsychological tests across different sociodemographic backgrounds in the Brazilian longitudinal study of adult health (ELSA-Brasil). *Neuropsychology, 34*(2), 227–234. https://doi.org/10.1037/neu0000597

Blankson, A. N., & McArdle, J. J. (2015). Measurement invariance of cognitive abilities across ethnicity, gender, and time among older Americans. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences, 70*(3), 386–397. https://doi.org/10.1093/geronb/gbt106

Bontempo, D. E., & Hofer, S. M. (2007). *Assessing factorial invariance in cross-sectional and longitudinal studies*. Oxford University Press.

Bowden, S. C. (2004). The role of factor analysis in construct validity: Is it a myth? *Journal of the International Neuropsychological Society, 10*(7), 1018–1019. author reply 1020-1011 https://doi.org/10.1017/s1355617704107091.

Bowden, S. C., Gregg, N., Bandalos, D., Davis, M. D., Coleman, C., Holdnack, J. A., & Weiss, L. G. (2008). Latent mean and covariance differences with measurement equivalence in college students with developmental difficulties versus the Wechsler adult intelligence scale-III/Wechsler memory scale-III normative sample. *Educational and Psychological Measurement, 68*(4), 621–642. https://doi.org/10.1177/0013164407310126

Bowden, S. C., Lange, R. T., Weiss, L. G., & Saklofske, D. H. (2008). Invariance of the measurement model underlying the Wechsler adult intelligence scale-III in the United States and Canada. *Educational and Psychological Measurement, 68*(6), 1024–1040. https://doi.org/10.1177/0013164408318769

Bowden, S. C., Lissner, D., McCarthy, K. A., Weiss, L. G., & Holdnack, J. A. (2007). Metric and structural equivalence of core cognitive abilities measured with the Wechsler adult intelligence scale-III in the United States and Australia. *Journal of Clinical and Experimental Neuropsychology, 29*(7), 768–780. https://doi.org/10.1080/13803390601028027

Bowden, S. C., Petrauskas, V. M., Bardenhagen, F. J., Meade, C. E., & Simpson, L. C. (2013). Exploring the dimensionality of digit span. *Assessment, 20*(2), 188–198. https://doi.org/10.1177/1073191112457016

Bowden, S. C., Saklofske, D. H., van de Vijver, F. J. R., Sudarshan, N. J., & Eysenck, S. B. G. (2016). Cross-cultural measurement invariance of the Eysenck personality questionnaire across 33 countries. *Personality and Individual Differences, 103*, 53–60. https://doi.org/10.1016/j.paid.2016.04.028

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2010). Invariance of the measurement model underlying the Wechsler adult intelligence scale-IV in the United States and Canada. *Educational and Psychological Measurement, 71*(1), 186–199. https://doi.org/10.1177/0013164410387382

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2011). Augmenting the core battery with supplementary subtests: Wechsler adult intelligence scale–IV measurement invariance across the United States and Canada. *Assessment, 18*(2), 133–140. https://doi.org/10.1177/1073191110381717

Bowden, S. C., Weiss, L. G., Holdnack, J. A., Bardenhagen, F. J., & Cook, M. J. (2008). Equivalence of a measurement model of cognitive abilities in U.S. standardization and Australian neuroscience samples. *Assessment, 15*(2), 132–144. https://doi.org/10.1177/1073191107309345

Bowden, S. C., Weiss, L. G., Holdnack, J. A., & Lloyd, D. (2006). Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale–III. *Psychological Assessment, 18*(3), 334–339. https://doi.org/10.1037/1040-3590.18.3.334

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second Edition (2nd ed.)). The Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Bryan, V. M., & Mayer, J. D. (2020). A meta-analysis of the correlations among broad intelligences: Understanding their relations. *Intelligence, 81*. https://doi.org/10.1016/j.intell.2020.101469

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(2), 272–300. https://doi.org/10.1207/s15328007sem1102_8

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Caemmerer, J. M., Keith, T. Z., & Reynolds, M. R. (2020). Beyond individual intelligence tests: Application of Cattell-Horn-Carroll theory. *Intelligence, 79*. https://doi.org/10.1016/j.intell.2020.101433

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, H., Keith, T. Z., Weiss, L., Zhu, J., & Li, Y. (2010). Testing for multigroup invariance of second-order WISC-IV structure across China, Hong Kong, Macau, and Taiwan [article]. *Personality and Individual Differences, 49*(7), 677–682. https://doi.org/10.1016/j.paid.2010.06.004

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance [periodical]. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. https://doi.org/10.1207/s15328007sem0902_5

Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between south African and British students on the Wechsler adult intelligence scales third edition (WAIS-III). *Frontiers in Psychology, 6*, 297. https://doi.org/10.3389/fpsyg.2015.00297

Damas, A., Jr. (2002). A factorial invariance analysis of intellectual abilities in children across different age and cultural groups using the Wechsler intelligence scale for children. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 62*(10–B), 4832.

Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M., & Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: Lessons from memory assessment. *Journal of the International Neuropsychological Society, 9*(6), 936–946. https://doi.org/10.1017/S1355617703960139

Deng, C., & Georgiou, G. K. (2015). Establishing measurement invariance of the cognitive assessment system across cultures. In *Cognition, intelligence, and achievement: A tribute to J P Das* (pp. 137–148). Elsevier Academic Press; US.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science, 46*(1), 137–149.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*(1), 21–50. https://doi.org/10.1207/S15327906MBR3501_2

Dolan, C. V., & Hamaker, E. L. (2001). Investigating black-White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In *, Vol. 6. Advances in psychology research* (pp. 31–59). Nova Science Publishers; US.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman's hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence, 32*(2), 155–173. https://doi.org/10.1016/j.intell.2003.09.001

Edwards, O. W., & Oakland, T. D. (2006). Factorial invariance of woodcock-Johnson III scores for African Americans and Caucasian Americans. *Journal of Psychoeducational Assessment, 24*(4), 358–366. https://doi.org/10.1177/0734282906289595

Gavett, B. E., Stypulkowski, K., Johnson, L., Hall, J., & O'Bryant, S. E. (2018). Factor structure and measurement invariance of a neuropsychological test battery designed for assessment of cognitive functioning in older Mexican Americans [article]. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10*, 536–544. https://doi.org/10.1016/j.dadm.2018.08.003

Goodman, Z. T., Llabre, M. M., Perera, M. J., Bainter, S. A., González, H. M., Vásquez, P. M., & Zeng, D. (2021). Testing measurement equivalence of neurocognitive assessments across language in the Hispanic community health study/study of Latinos. *Neuropsychology, 35*(4), 423–433. https://doi.org/10.1037/neu0000725

Gross, A. L., Khobragade, P. Y., Meijer, E., & Saxton, J. A. (2020). Measurement and structure of cognition in the longitudinal aging study in India-diagnostic assessment of dementia. *Journal of the American Geriatrics Society, 68*(Suppl. 3), S11–S19. https://doi.org/10.1111/jgs.16738 (S3).

Gygi, J. T., Fux, E., Grob, A., & Hagmann-von Arx, P. (2016). Measurement invariance and latent mean differences in the Reynolds intellectual assessment scales (RIAS): Does the German version of the RIAS allow a valid assessment of individuals with a migration background? *PLoS One, 11*(11), Article e0166533. https://doi.org/10.1371/journal.pone.0166533

Hajovsky, D. B., & Chesnut, S. R. (2022). Examination of differential effects of cognitive abilities on reading and mathematics achievement across race and ethnicity: Evidence with the WJ IV. *Journal of School Psychology, 93*, 1–27. https://doi.org/10.1016/j.jsp.2022.05.001

Holding, P., Anum, A., van de Vijver, F. J. R., Vokhiwa, M., Bugase, N., Hossen, T., … Gomes, M. (2018). Can we measure cognitive constructs consistently within and across cultures? Evidence from a test battery in Bangladesh, Ghana, and Tanzania. *Applied Neuropsychology: Child, 7*(1), 1–13. https://doi.org/10.1080/21622965.2016.1206823

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3–4), 117–144. https://doi.org/10.1080/03610739208253916

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

Jewsbury, P. A., Bowden, S. C., & Duff, K. (2016). The Cattell–Horn–Carroll model of cognition for clinical assessment. *Journal of Psychoeducational Assessment, 35*(6), 547–567. https://doi.org/10.1177/0734282916651360

Jewsbury, P. A., Bowden, S. C., & Strauss, M. E. (2016). Integrating the switching, inhibition, and updating model of executive function with the Cattell-Horn-Carroll model. *Journal of Experimental Psychology. General, 145*(2), 220–245. https://doi.org/10.1037/xge0000119

Joreskog, K. G. (1970). Simultaneous factor analysis in several populations. *Psychometrika.*. https://doi.org/10.1007/BF02291366

Jorgensen, T., Kite, B., Chen, P., & Short, S. (2018). *Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis* (Vol. 23, pp. 708–728).

Karr, J. E., Scott, T. M., Aghvinian, M., & Rivera Mindt, M. (2022). Harmonization of the English and Spanish versions of the NIH toolbox cognition battery crystallized and fluid composite scores. *Neuropsychology.*. https://doi.org/10.1037/neu0000822

Keith, T. Z., Fugate, M. H., DeGraff, M., Diamond, C. M., Shadrach, E. A., & Stevens, M. L. (1995). Using multi-sample confirmatory factor analysis to test for construct Bias: An example using the K-ABC. *Journal of Psychoeducational Assessment, 13*(4), 347–364. https://doi.org/10.1177/073428299501300402

Keith, T. Z., Quirk, K. J., Schartzer, C., & Elliott, C. D. (1999). Construct bias in the differential ability scales? Confirmatory and hierarchical factor structure across three ethnic groups. *Journal of Psychoeducational Assessment, 17*(3), 249–268. https://doi.org/10.1177/073428299901700305

Khoo, Y. M. (2021). *Psychometric evaluation and empirical application of the Montreal cognitive assessment for older adults Saint Louis university*. Saint Louis University: ProQuest Dissertations Publishing.

Lakin, J. M. (2012). Multidimensional ability tests and culturally and linguistically diverse students: Evidence of measurement invariance. *Learning and Individual Differences, 22*(3), 397–403. https://doi.org/10.1016/j.lindif.2011.12.003

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*(6), 543–566. https://doi.org/10.1016/S0160-2896(03)00051-5

Martin, R. C., & Allen, C. (2012). Case studies in neuropsychology. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, K. J. Sher, … K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. (pp. 633–646)*. American Psychological Association. https://doi.org/10.1037/13620-033.

McArdle, J. J., Fisher, G. G., & Kadlec, K. M. (2007). Latent variable analyses of age trends of cognition in the health and retirement study, 1992-2004. *Psychology and Aging, 22*(3), 525–545. https://doi.org/10.1037/0882-7974.22.3.525

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology, 24*(2), 154–168. https://doi.org/10.1111/j.2044-8317.1971.tb00463.x

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology, 93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Meredith, W. (1964a). Notes on factorial invariance. *Psychometrika, 29*(2), 177–185. https://doi.org/10.1007/BF02289699

Meredith, W. (1964b). Rotation to achieve factorial invariance. *Psychometrika, 29*(2), 187–206. https://doi.org/10.1007/BF02289700

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/bf02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(11 Suppl 3), S69–S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis Group.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93–115. https://doi.org/10.1037/1082-989X.9.1.93

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ, 339*, Article b2535. https://doi.org/10.1136/bmj.b2535

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & Gonzalez, H. (2004). Spanish and English neuropsychological assessment scales (SENAS): Further development and psychometric characteristics. *Psychological Assessment, 16*(4), 347–359. https://doi.org/10.1037/1040-3590.16.4.347

Mungas, D., Widaman, K. F., Reed, B. R., & Tomaszewski Farias, S. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology, 25*(2), 260–269. https://doi.org/10.1037/a0021090

Naglieri, J. A., Taddei, S., & Williams, K. M. (2013). Multigroup confirmatory factor analysis of U.S. and Italian children's performance on the PASS theory of intelligence as measured by the cognitive assessment system. *Psychological Assessment, 25*(1), 157–166. https://doi.org/10.1037/a0029828

Neumann, D., Peterson, E. R., Underwood, L., Morton, S. M. B., & Waldie, K. E. (2021). Exploring the factor structure of the NIH toolbox cognition battery in a large sample of 8-year-old children in Aotearoa New Zealand. *Journal of the International Neuropsychological Society, 27*(8), 825–834. https://doi.org/10.1017/S1355617720001265

te Nijenhuis, J., Evers, A., & Mur, J. P. (2000). Validity of the differential aptitude test for the assessment of immigrant children. *Educational Psychology, 20*(1), 99–115. https://doi.org/10.1080/014434100110416

te Nijenhuis, J., Tolboom, E., Resing, W., & Bleichrodt, N. (2004). Does cultural background influence the intellectual performance of children from immigrant groups? The RAKIT intelligence test for immigrant children. *European Journal of Psychological Assessment, 20*(1), 10–26. https://doi.org/10.1027/1015-5759.20.1.10

Omura, K., & Sugishita, M. (2004). Simultaneous confirmatory factor analysis of the wechsler memory scale - revised for two standardization samples: A comparison of groups from Japan and the United States. *Journal of Clinical and Experimental Neuropsychology, 26*(5), 645–652. https://doi.org/10.1080/13803390409609788

Pandolfi, V. (1998). Assessment of factor models underlying the WISC-III in White, black, and Hispanic subgroups of the standardization sample. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 58*(8–B), 4464.

Papadopoulos, T. C., Georgiou, G. K., Deng, C., & Das, J. P. (2018). The structure of speed of processing across cultures. *Advances in Cognitive Psychology, 14*(3), 112–125. https://doi.org/10.5709/acp-0243-7

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. https://doi.org/10.1016/j.dr.2016.06.004

Rawlings, A. M., Bandeen-Roche, K., Gross, A. L., Gottesman, R. F., Coker, L. H., Penman, A. D., … Mosley, T. H. (2016). Factor structure of the ARIC-NCS neuropsychological battery: An evaluation of invariance across vascular factors and demographic characteristics. *Psychological Assessment, 28*(12), 1674–1683. https://doi.org/10.1037/pas0000293

Reed, C. L. (2001). An investigation of measurement invariance in the wisc iii: Examining a sample of referred African American and Caucasian students. *Dissertation Abstracts International Section A: Humanities and Social Sciences, 61*(8–a), 3055.

Reverte, I., Golay, P., Favez, N., Rossier, J., & Lecerf, T. (2015). Testing for multigroup invariance of the WISC-IV structure across France and Switzerland: Standard and CHC models. *Learning and Individual Differences, 127*. https://doi.org/10.1016/j.lindif.2015.03.015

Reynolds, M. R., Keith, T. Z., Flanagan, D. P., & Alfonso, V. C. (2013). A cross-battery, reference variable, confirmatory factor analytic investigation of the CHC taxonomy. *Journal of School Psychology, 51*(4), 535–555. https://doi.org/10.1016/j.jsp.2013.02.003

Rummel, J., Steindorf, L., Marevic, I., & Danner, D. (2019). A validation study of the German complex-span tasks and some general considerations on task translation procedures in cognitive psychology. *European Journal of Psychological Assessment, 35*(5), 725–736. https://doi.org/10.1027/1015-5759/a000444

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's advanced progressive matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12*(3), 220–229. https://doi.org/10.1111/j.0965-075X.2004.00276.x

Rutkowski, L., & Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's standard progressive matrices plus [article]. *Personality and Individual Differences, 53*(2), 137–141. https://doi.org/10.1016/j.paid.2011.06.013

Scheiber, C. (2016a). Do the Kaufman tests of cognitive ability and academic achievement display construct bias across a representative sample of black, Hispanic, and Caucasian school-age children in grades 1 through 12? *Psychological Assessment, 28*(8), 942–952. https://doi.org/10.1037/pas0000236

Scheiber, C. (2016b). Is the Cattell–Horn–Carroll-Based Factor Structure of the Wechsler Intelligence Scale for Children—Fifth Edition (WISC-V) Construct Invariant for a Representative Sample of African–American, Hispanic, and Caucasian Male and Female Students Ages 6 to 16 Years? *Journal of Pediatric Neuropsychology, 2*(3–4), 79–88. https://doi.org/10.1007/s40817-016-0019-7

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan, & E. M. McDonough (Eds.), *Contemporary intellectual assessment* [*Theories, Tests, and Issues*] (4th ed.). The Guilford Press.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press. https://doi.org/10.1017/CBO9780511526817

Siedlecki, K. L., Manly, J. J., Brickman, A. M., Schupf, N., Tang, M. X., & Stern, Y. (2010). Do neuropsychological tests have the same meaning in Spanish speakers as they do in English speakers? *Neuropsychology, 24*(3), 402–411. https://doi.org/10.1037/a0017515

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Tennant, I. A., Hull, D. M., Fagan, M. A., Casaletto, K. B., Heaton, R. K., James Bateman, C., … Boyne, M. (2022). Assessment of cross-cultural measurement invariance of the NIH toolbox fluid cognition measures between Jamaicans and African-Americans. *Applied Neuropsychology. Adult, 1-9*. https://doi.org/10.1080/23279095.2022.2126939

Trundt, K. M., Keith, T. Z., Caemmerer, J. M., & Smith, L. V. (2017). Dont use testing for construct Bias in the differential ability scales, second edition: A comparison among African American, Asian, Hispanic, and Caucasian children. *Journal of Psychoeducational Assessment, 36*(7), 670–683. https://doi.org/10.1177/0734282917698303

Tuokko, H. A., Chou, P. H., Bowden, S. C., Simard, M., Ska, B., & Crossley, M. (2009). Partial measurement equivalence of French and English versions of the Canadian study of health and aging neuropsychological battery. *Journal of the International Neuropsychological Society, 15*(3), 416–425. https://doi.org/10.1017/S1355617709090602

Van de Vijver, F. J. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28*(6), 678–709. https://doi.org/10.1177/0022022197286003

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*(1), 4–70. https://doi.org/10.1177/109442810031002

van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (p. c1997). Thousand Oaks: Sage Publications.

Warne, R. T., & Burningham, C. (2019). Spearman's g found in 31 non-Western nations: Strong evidence that g is a universal phenomenon. *Psychological Bulletin, 145*(3), 237–272. https://doi.org/10.1037/bul0000184

Watkins, M. W., & Canivez, G. L. (2001). *Longitudinal Factor Structure of the WISC-III among Students with Disabilities. 38* pp. 291–298). https://doi.org/10.1002/pits.1019

Weiss, L. G., & Saklofske, D. H. (2020). Mediators of IQ test score differences across racial and ethnic groups: The case for environmental and social justice. *Personality and Individual Differences, 161*, Article 109962. https://doi.org/10.1016/j.paid.2020.109962

White, L. J., & Greenfield, D. B. (2017). Executive functioning in Spanish and English-speaking head start preschoolers. *Developmental Science, 20*(1), 1–14. https://doi.org/10.1111/desc.12502

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist, 30*(7), 1006–1016. https://doi.org/10.1080/13854046.2016.1205136

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29*(3), 39–47. https://doi.org/10.1111/j.1745-3992.2010.00182.x

Widaman, K. F., & Olivera-Aguilar, M. (2023). Investigating measurement invariance using confirmatory factor analysis. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (2nd ed., pp. 367–384).

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant,

& M. Windle (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, Bias, and solution propriety. *Educational and Psychological Measurement, 76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Woods, I. L. (2017). *Do the WJ-IV tests of cognitive abilities predict Reading equally across groups?* The University of Memphis. /z-wcorg/ https://umwa.memphis.edu/etd/index.php/view/home/5998.

Wray, C., Kowalski, A., Mpondo, F., Ochaeta, L., Belleza, D., DiGirolamo, A., … Cohorts.. (2020). Executive functions form a single construct and are associated with schooling: Evidence from three low- and middle- income countries. *PLoS One, 15*(11), Article e0242936. https://doi.org/10.1371/journal.pone.0242936

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data [article]. *Practical Assessment, Research & Evaluation, 12*(3), 1–26. https://doi.org/10.7275/mhqa-cd89

Xu, C., Ellefson, M. R., Ng, F. F., Wang, Q., & Hughes, C. (2020). An east-west contrast in executive function: Measurement invariance of computerized tasks in school-aged children and adolescents. *Journal of Experimental Child Psychology, 199*, Article 104929. https://doi.org/10.1016/j.jecp.2020.104929