Please use the following citation when referencing this work:

Benson, N. F., Beaujean, A. A., McGill, R. J., & Dombrowski, S. C. (2018). Revisiting Carroll's survey of factor-analytic studies: Implications for the clinical assessment of intelligence. *Psychological Assessment*. Advance online publication. doi: 10.1037/pas0000556

Revisiting Carroll's Survey of Factor-Analytic Studies: Implications for the Clinical Assessment of Intelligence

Nicholas F. Benson and A. Alexander Beaujean

Baylor University

Ryan J. McGill

William & Mary

Stefan C. Dombrowski

Rider University

Author note

Correspondence concerning this article should be addressed to Nicholas F. Benson, Department of Educational Psychology, Baylor University, One Bear Place #97301, Waco, TX 76798. E-mail: Nicholas_Benson@baylor.edu

Abstract

John Carroll's three-stratum theory—and the decades of research behind its development—are foundational to the contemporary practice of intellectual assessment. The present study addresses some limitations of Carroll’s work: specification, reproducibility with more modern methods, and interpretive relevance. We re-analyzed select datasets from Carroll's survey of factor analytic studies using confirmatory factor analysis as well as modern indices of interpretive relevance. For the majority of the datasets, we found that Carroll likely extracted too many factors representing Stratum II abilities. Moreover, almost all of factors representing Stratum II abilities had little-to-no interpretive relevance above and beyond that of general intelligence. We conclude by discussing implications of this research and some directions for future research.

Keywords: John Carroll, Intelligence, Three-stratum theory, Replication study

Revisiting Carroll's Survey of Factor-Analytic Studies: Implications for the Clinical Assessment of Intelligence

John Carroll's (1993) three-stratum (3S) theory of intelligence has been widely lauded (e.g., Lubinski, 2000, 2004). People have compared it to Dmitri Mendeleev's periodic table (Horn, 1998), Isaac Newton's Mathematical Principles of Natural Philosophy (McGrew, 2005), and a composer's magnum opus (Jensen, 2004). 3S theory is part of the Spearman-Holzinger-Burt "British" school that prioritizes a single general ability over other abilities and posits that cognitive abilities directly influence all tasks that require any aspect of cognitive functioning (Carroll & Schweiker, 1951). Like other theories in the British tradition, 3S is hierarchical with each stratum representing psychological abilities of an increasing level of abstraction. Stratum I (S1) consists of abilities that closely correspond to the surface-level characteristics of mental tasks (e.g., induction, sequential reasoning, verbal ability, and visualization). Stratum II (S2) consists of broad abilities that differ in the extent to which they are related to the content of test stimuli, cognitive processing demands, or response demands. There are eight S2 abilities: crystallized intelligence (Gc), fluid intelligence (Gf), broad retrieval ability (Gr), broad cognitive speediness (Gs), processing speed (Gt), broad auditory perception (Gu), broad visual perception (Gv), general memory and learning (Gy). Stratum III (S3) reflects a highly abstract and general ability (*g*) that affects all tasks requiring cognitive ability.

Carroll was not a clinician, so his primary purpose in creating the 3S theory was to "identify, catalog, and interpret the known abilities, without regard for their importance or validity" (1993, p. 693). He extrapolated about some possible clinical uses, but they were primarily centered on test construction. Nonetheless, very soon after publishing his work it became a cornerstone of clinical intellectual assessment—a role it has held on to for almost two decades (Keith & Reynolds, 2010).

The rapid infusion of Carroll's work into clinical practice was largely due to two major events. The first event was the development of the Cattell-Horn-Carroll (CHC) theory of cognitive abilities (McGrew, 2005). CHC is an integration of *Gf-Gc* theory (Cattell, 1943; Horn & Cattell, 1966) with 3S theory to help clinicians provide explanations of how and why people differ in their various cognitive abilities (Schneider & McGrew, 2012). Second, the publication of the third edition of the Woodcock-Johnson (WJ III; Woodcock, McGrew, & Mather, 2001). Not only was the WJ III the first test to be aligned with CHC theory, but also its technical manual provided the first comprehensive publication about it.

Distinctions from Cattell-Horn-Carroll Theory

While 3S theory shares many similarities with CHC theory, there are important distinctions. First, S1 and S2 abilities differ somewhat with respect to number and interpretation— likely due to the theories' different purposes, authors, and data used for their development. Initially, CHC contained 10 S2 abilities (compared to the 8 in 3S), but this increased to 16 and now includes 18 (Schneider & McGrew, in press). Moreover, the definition of some of the S2 abilities have changed as well (Beaujean, in press).

Second, Carroll was adamant that 3S aligned with the British tradition, so can be represented by a bifactor model (Beaujean, 2015). ~~In fact, Carroll (1993) heavily utilized the Schmid-Leiman (1957) procedure to approximate bifactor structure.~~ Conversely, CHC theory is more aligned with the American/Thurstonian tradition, so is better represented by a higher-order model. We discuss this more in the *Factor Model* section in the current manuscript, but there are two key differences between these models. First, in a bifactor model *g* is orthogonal to all other factors; thus, the group factors represent effects that are independent of *g*. In a higher-order model, *g* represents what is common to the group factors. Second, in a bifactor model, all common factors—including the general factor—have direct effects on measured abilities. In a higher-order model, since *g* is defined by group factors, it only has indirect effects on mental tasks meaning that lower-order factors fully mediate the effects of *g* on all mental tasks (Gignac, 2008).

A third difference between 3S and CHC is the importance placed on *g*. Carroll (2012) believed Spearman's two-factor theory of intelligence was essentially correct, but was incomplete since Spearman was hesitant about interpreting any broad ability. Thus, *g* is an indispensable aspect of 3S. In contrast, CHC theory is ambiguous about the role of *g* (Cucina & Howardson, 2017). While *g* is included, users are encouraged to ignore it if they do not believe that *g* has merit, particularly in applied clinical assessment contexts (Schneider & McGrew, 2012).

Despite the aforementioned differences, both Carroll's research and his 3S theory are used as pivotal support for CHC theory (e.g., Flanagan, Alfonso, & Ortiz, 2012; McGrew, 2005, 2009; Ortiz, 2015; Schneider & Flanagan, 2015; Schneider & McGrew, 2012). Thus, there is a need for clarity about how Carroll's findings relate to CHC. Such an investigation is especially appropriate given the numerous examples from psychology's "reproducibility crisis" (Pashler & Wagenmakers, 2012) that illustrate the benefits of reevaluating the evidence-base for widely accepted theories—or recommended application of those theories in practice—in light of new developments. One can only have confidence in a theory when it has been subjected to a 'risky' empirical test with due consideration to rival or alternative explanations for the data (Meehl, 1990; Platt, 1964; Popper, 2002).

Shortcomings of Carroll's Research

Carroll acknowledged important limitations of his research.

> Much work remains to be done in the factor-analytic study of cognitive abilities. The map of abilities provided by the three-stratum theory undoubtedly has errors of commission and omission, … [it] needs to be further validated by acquiring information about the importance and relevance of the various abilities it specifies. (Carroll, 2012, p. 889)

These limitations are often overlooked, however, in the effort to use his research to support the development of clinical tests or score interpretation systems. Three major limitations of 3s theory, which are detailed below, are inadequate specification of S1 and S2 factors, the need to replicate the factors using more modern methods, and lack of clarity about the factor model.

Inadequate specification. No single dataset that Carroll (1993) evaluated had sufficient breadth to evaluate fully abilities at all three strata of 3S theory. In fact, the majority of the datasets he used only contained a handful of S1 abilities and very few S2 abilities. Moreover, often these abilities were factorially complex (i.e., non-negligible cross-loadings). Thus, Carroll wrote that many of the factors representing 3S abilities are "inadequately specified, and many aspects of the three-stratum theory need to be tested and refined" (p. 688).

Notably, even though Carroll (1993) developed 3S based on factor-analysis, he did not determine the abilities' strata based on a factor order. Instead, he used his judgement of the ability's degree of generality. This was largely because his datasets contained a variety of breadth in the measured variables. Thus, S2 abilities could emerge as either a first- or second-order factor, and *g* could emerge as a first-, second- or third-order factor in different datasets. While he provided a detailed explanation for his decisions, relying on such qualitative judgements raises the question of replicability. Would independent scholars derive the same conclusions with the same data? Carroll himself even suggested that his findings should be viewed cautiously until others could verify his research.

Being able to replicate his results—especially concerning S2 abilities—is important. Significantly under- or over-factoring datasets can bias factor loadings and distort interpretation (Comrey, 1978). Moreover, over-factoring can yield unnecessarily complex theories that contain components of little theoretical or practical import (Fabrigar, Wegener, MacCallum, & Strahan, 1999), which some have argued is far too common in modern intelligence research (Frazier & Youngstrom, 2007).

Replication with modern techniques. Carroll (1993) relied exclusively on exploratory factor analysis (EFA) for all his original analyses because of its practicality, inexpensiveness, and ease of operation. Nonetheless, he recognized that his results need to be replicated using confirmatory techniques. In a few studies that followed his landmark study, he was able to utilize confirmatory factor analysis (CFA; Carroll, 1995, 1997, 2003). He concluded that the CFA results confirmed much of 3S, but also differed in important ways from what he found with his original EFAs. Thus, he suggested that the two methods should be used in combination (Carroll, 1995).

CFA has important characteristics that make it extremely useful when examining factor structure (Loehlin & Beaujean, 2016). Unlike EFA, common variance attributable to latent variables are partitioned from measurement error so that factor structure can be examined independent of these effects. Moreover, because CFA models are often over-identified, there are a variety of model-fit measures that can aid in a more accurate decision for factor retention than using EFA (Keith, Caemmerer, & Reynolds, 2015). Finally, stricter assumptions about relations among measured and latent variables can be specified *a priori*, which allows for stronger sets of hypotheses to evaluate.

Factor model clarity. When utilizing CFA as a confirmatory method, it is important to specify the models *a priori*. There are differing opinions about what model best represents the 3S theory, and Carroll himself was not always clear in his writings. In some places, the wording he used gives the impression that abilities at higher strata directly influence abilities at lower strata, but in other places he described the abilities at one stratum being independent of abilities at another stratum. To add to this confusion, in his EFA he performed high-order extraction of factors (when applicable), but then utilized an orthogonal transformation of the results to approximate a bifactor solution.

To gain clarity on this issue, it is informative to examine some of his writing about 3S following the publication of his landmark study. For his CFA studies, (Carroll, 1995, 1997, 2003), he always specified a bifactor model. Moreover, in his 1996 chapter he was very explicit that *g* should be ascribed primary importance in these models. Carroll contended that it is best to study abilities that are orthogonal to each other— partialing out their covariance of factors at other orders —and that abilities at each stratum had direct effects on the measured variables. All of these attributes are best expressed through a bifactor model (Beaujean, 2015). Thus, we believe a bifactor model best represents Carroll's intention in developing 3S theory.

Purpose of the Current Study

The present study is aimed at addressing some of the limitations of Carroll's (1993) work as well as investigating the support—or lack thereof—for CHC theory. The first aim was to better understand the specification and replicability of Carroll's S2 abilities using CFA with select datasets from Carroll's original factor analytic survey. Although much has been written about the reproducibility of psychological research in the past decade (e.g., Funder et al., 2014), little has focused on factor-analytic research. Yet, replication of factors—across datasets and methods—is paramount to the interpretation and use of factor analytic studies (Gorsuch, 2003; Osborne & Fitzpatrick, 2012).

The second aim of this study was to evaluate the interpretive relevance of test scores believed to reflect the factors representing S2 and S3 abilities. A test-derived score with interpretive relevance: (a) provides a good representation of the construct targeted for measurement; (b) is distinct from conceptually similar constructs; (c) is likely to be replicable across datasets and methods; and (d) has adequate unique, reliable variance such that it is statistically distinguishable from test-derived scores reflecting conceptually similar constructs (McGill, 2017).

We examined interpretive relevance because of the ubiquitous practice of interpreting and utilizing test scores measuring S2 abilities in applied psychology (e.g., Schrank, McGrew, & Mather, 2014; Wechsler, 2014). We believe that if clinicians are going to use test scores for applied purposes such as diagnosis and treatment planning, then there needs to be evidence to support the interpretation of these test scores (AERA/APA/NCME, 2014). Factor analysis can provide evidence indicating an ability exists, but such evidence is insufficient to demonstrate that scores reflecting this ability have interpretive relevance. Carroll (1993) was concerned mainly about knowing what cognitive abilities exist, not whether such abilities are clinically useful. Thus, he largely eschewed the topic of interpretive relevance and left such questions for others to investigate.

Method

Inclusion Criteria

We initially selected the 18 datasets from which Carroll (1993) originally extracted multiple factors representing S2 abilities (see Table 15.3, pp.585–589). We used these because his results suggested these datasets contain a sufficient number of measured variables and adequate variance for identifying multiple S2 abilities.

For each of the identified studies, we obtained the original manuscript. Unfortunately, some of these manuscripts either did not supply covariances or did not supply enough information to convert the supplied correlation matrix into a covariance matrix. While correlation matrices are appropriate for EFA studies, using them in CFA can be problematic (Cudek, 1989). Thus, we limited our study to a subset of nine studies from Carroll's (1993) project for which we could input a covariance matrix. For two of the studies (Sung & Dawis, 1981; Thurstone & Thurstone, 1941), some of our models did not converge—likely due to idiosyncratic characteristics of the cognitive tasks or sample. Instead of employing a lot of *post hoc* tinkering to force convergence, we just report results from the seven studies that did not exhibit estimation troubles.

Descriptive information regarding the participants in these studies is presented in Table 1. Of note, the majority of these studies used non-representative samples (e.g., prisoners, enlisted military personnel). The number of cognitive variables used in these studies range from 19–48, with an average of 27 (SD = 10.5). Detailed descriptions of the variables in these studies are provided in the original publications.

Data Analysis

Confirmatory factor analysis. We used Mplus (Version 8.0; Muthén & Muthén, 1998-2017) for all model comparisons. For each dataset, we systematically fit a series of bifactor models starting with Spearman's two-factor model. Next, we sequentially fit models containing one additional group factor representing a S2 ability. If adding a group factor yielded significant improvement of model fit, then we retained the group factor. We subsequently repeated this process for each additional group factor.

We used the following indices to evaluate model fit: comparative fit index (CFI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and Bayesian information criterion (BIC). The CFI, RMSEA, and SRMR have metrics that are directly interpretable, but the BIC does not. To facilitate interpretation of the BIC, we transformed the values to BIC weights (see formulae in Wagenmakers and Farrell, 2004). These weights range from 0–1 and can be interpreted as the probability that a candidate model is the best model among the set of candidate models examined.

Interpretative relevance. To examine interpretative relevance, we examined model-based indices of: (a) the dimensionality of the datasets, (b) the interpretability of factors, and (c) reliability of test-derived scores (Rodriguez, Reise, & Haviland, 2016a). To evaluate dimensionality, we used three indices: (a) percentage of uncontaminated correlations (PUC; Bonifay, Reise, Scheines, & Meijer, 2015); (b) explained common variance (ECV; Reise, Moore, & Haviland, 2010); and (c) average relative parameter bias (ARPB). PUC is an index of the extent to which covariance terms reflect a general factor and are uncontaminated by variance from group factors, with values ≥.80 suggesting unidimensionality (Reise, Scheines, Widaman, & Haviland, 2013). ECV is an index of general factor strength. As a frame of reference, values of .80 and .70 indicate that ignoring group factors would result in bias of less than 5% and 10%, respectively, in estimating a unidimensional general factor. ARPB is an estimate of the difference between a measured variable's factor loading in a unidimensional model and its *g* loading in a bi-factor model. Values < .15 indicate an acceptable amount of bias in interpreting a multidimensional dataset as unidimensional (Muthén, Kaplan, & Hollis, 1987). The aforementioned criteria for dimensionality can be used to inform decisions regarding the retention of factors, providing a safeguard against both under-factoring and over-factoring.

To evaluate the interpretability of the S2 ability factors, we examined factor determinacy (Beauducel, 2011) and construct replicability (H; Hancock & Mueller, 2001). Factor determinacy values reflect the uniqueness of factor scores. Values closer to 1.0 (generally .90 or higher) indicate that test-derived scores reflect individual differences in performance and may be of value for subsequent analyses or applied assessment purposes. H is a measure of how well the observed variables represent their intended constructs (i.e., *g* and S2 abilities) as well as the extent to which these factors are likely to replicate across studies. Values >.80 are usually considered indicative of a well-defined latent variable that is likely to replicate across datasets and methods.

Last, to evaluate the reliability of test-derived scores representing their latent constructs we calculated omega (ω; Lucke, 2005), as well as omega hierarchical (ωH) and omega hierarchical subscale (ωHS). ωH and ωHS, respectively, reflect the reliability of a general factor and the unique reliability of group factors after removing the variance due to a general factor (Reise, 2012; Rodriguez, Reise, & Haviland, 2016b). If factors do not possess sufficient unique, reliable variance then it is not possible to make reliable distinctions between them when interpreting profiles of scores. Moreover, given that measures of cognitive ability tend to be strongly correlated low ωHS values (e.g., < .30) portend limited incremental validity beyond *g* for predicting outcomes such as academic achievement or occupational performance.

Results

CFA results are presented in Table 2 for each study. In Table 3 we provide a summary of how the results comport with Carroll's (1993) original EFA results. In 43% of the studies, all the S2 abilities proposed by Carroll's analysis were confirmed, while in 57% of the studies we did not find support for at least one S2 ability. Thus, in over half of the studies we found that Carroll likely extracted too many S2 ability factors.

In Table 4 we provide interpretive relevance indices. *g* appears to adequately explain most of the individual differences in performance for 57% of the studies (i.e., Fogarty, 1987; Hakstian & Cattell, 1978; Horn & Stankov, 1982; Undheim, 1981), although a case could be made for possibly interpreting some S2 abilities in a few of the datasets. Specifically, broad auditory perception (*Gu)* may have some interpretive relevance in Fogarty's (1987) data, and broad visual perception (*Gv)* may have some relevance in Hakstian and Cattell's (1978) data.

A multidimensional structure appears to be needed for 29% of the datasets (i.e., Christal, 1958; Gustafsson, 1984). For Christal's data, general memory and learning (*Gy*) appears to have some interpretative relevance, while fluid reasoning (*Gf*) appears to have some interpretative relevance in Gustafsson 's data.

The data from Horn (1965) are somewhat equivocal. PUC suggests a unidimensional interpretation, but the ARPB and ECV do not. Moreover, the ωHS value for processing speed (*Gs*) is relatively high and very close to the ω and ωH values for *g*. Thus, it is indeterminate as to whether *g* adequately explains individual differences in performance, or if other abilities might also be relevant for this data.

Discussion

Three-stratum (3S) theory was developed by an eminent factor analyst based on a large body of factor-analytic research. Factor analytic results are constrained by the quality of datasets analyzed, however, and most of the studies Carroll (1993) used were from non-representative convenience samples (e.g., prisoners, military personnel) or were minimally multi-dimensional.

For our study, we purposely selected the datasets from which Carroll (1993) identified the most second-order factors representing S2 abilities; of these, we were able to re-analyze seven. For the majority of the datasets, we found that Carroll likely extracted too many factors representing S2 abilities. Moreover, we found that almost all of the S2 ability factors we could replicate had little-to-no interpretative relevance. Stated differently, *g* was the only ability that was consistently found to influence the measured variables and be clinically meaningful across all the datasets.

Interpretative Relevance

Although our results imply that the 3S theory may not be as parsimonious as it could be, this is not necessarily surprising. Carroll was not interested in model parsimony, factor importance, or practical applications as much as he was in identifying all known human cognitive abilities regardless of their importance. What is surprising is the emphasis clinicians and clinical researchers have placed on S2 abilities and their citation of Carroll's research as supporting this idea. A prime example of this is cross-battery assessment (Flanagan, Alfonso, Ortiz, 2012).

Cross-battery assessment was developed to enable clinicians to conduct intellectual assessments that "approximate the total range of cognitive and academic abilities and neuropsychological processes more adequately than what is possible with most collections of conormed tests" (Flanagan et al., 2012, p. 459). In other words, it is an approach to intellectual assessment that focuses on S2 (and some S1) abilities while all-but-excluding *g*. It has developed a large following since its initial development and currently many states allow—and some even require—examining the pattern of strengths and weaknesses in S2 abilities as part of their learning disability eligibility criteria (Maki, Floyd, & Roberson, 2015).

Although the developers of the cross-battery approach note its foundations in CHC theory, they cite Carroll's work ubiquitously as supporting the idea that clinical assessment should focus on S2 and S1 abilities. We do not disagree that S1 and S2 abilities exist—in fact, our re-analyses were able to replicate many of the same S2 abilities as Carroll (1993) found. At the same time, however, our study lends little credence to the notion that Carroll's data support the clinical interpretation of abilities other than *g*. Since we selected the datasets from Carroll's original study that had the highest likelihood of supporting of S2 abilities, we believe this conclusion probably applies to the entirety of Carroll's catalog of studies.

It could be argued that Carroll's (1993) datasets were not sufficiently broad to capture the actual importance of S2 abilities. Instead, a more rigorous examination would require collecting data on all the S2 abilities from a single sample, such as those from the norming data of the third and fourth editions of the Woodcock-Johnson cognitive abilities tests (Woodcock et al., 2001, Schrank et al., 2014). The authors of these tests designed them to measure many of the S2 abilities in CHC theory, which also encompasses most of the S2 abilities in the 3S theory. As with the results of the present study, structural analyses of the Woodcock-Johnson raise concerns about the clinical interpretability of the scores representing S2 abilities in those datasets (e.g., Dombrowski, 2013; Dombrowski, 2014a, 2014b; Dombrowski, 2015a; Dombrowski, McGill & Canivez, 2017a, 2017b; Dombrowski & Watkins, 2013).

Another argument could be made that factor analytic studies are not sufficient to examine interpretive relevance. Instead, studies need to be conducted that examine the diagnostic or instructional utility of the S2 abilities. In the two decades since Carroll published his masterwork, there is very limited evidence supporting diagnostic or treatment utility of interpreting scores representing S2 abilities (Kranzler, Floyd, Benson, Zaboski & Thibodaux, 2016; McGill & Busse, 2017). Carroll (2000) himself even noted this:

> What I find lacking is evidence of the value and predictive validity of the diagnoses and recommendations that a school psychologist might make on the basis of such a [cognitive profile] system (p. 454).

Clinical Utility

Incremental validity is a crucial issue regarding the utility of scores derived from tests of cognitive abilities. Incremental validity is supported when test scores improve prediction of important external criteria beyond other scores on the same test or scores on other established measures (Hunsley & Meyer, 2003). Although possessing reliable specific variance is not sufficient for incremental validity, it is a necessary condition for improving the prediction of external criteria beyond *g*. Using some of Carroll's (1993) most multidimensional data, our results suggest that the factors representing S2 abilities account for minimal reliable variance in test performance beyond *g*. This, in turn, suggests few of these factors would display incremental validity. Moreover, the moderate-to-strong intercorrelations among these factors cause multicollinearity issues when examining the effects of these abilities outside of latent variable models (e.g., multiple regression analyses with observed test scores, interpreting profiles of cognitive strengths and weaknesses).

Recently, Benson, Kranzler, and Floyd (2016) suggested that orthogonal factor scores can provide more precise representations of the targeted constructs than non-refined factor scores (i.e., scores calculated by simply summing scaled scores and assigning equal weights to all subtest scores that contribute to a composite) and maintain close correspondence with the cognitive–achievement relations observed for latent variables. Results from the present study support the use of some factor scores in subsequent analyses outside of latent variable models. Following the recommendation of Gorsuch (1983), factor score estimates should only be used when the factor determinacy index value exceeds .90. Using this criterion, it would be appropriate to use factor scores for about 11% of the S2 factors identified in this study. Although research (Benson et al., 2016, Kranzler, Benson, & Floyd, 2015) suggests that a few S2 abilities display incremental validity in predicting important outcomes such as academic achievement, evidence strongly supports emphasizing *g* when interpreting individual differences in test performance and predicting academic achievement.

Limitations

Our study is limited by some of the same factors that limited Carroll's work, particularly the quality of datasets examined. Also, we only analyzed a small subset of the studies Carroll analyzed—although we carefully selected them in an effort to maximize the identification of S2 abilities. While it is unlikely that results more supportive of these abilities would emerge from the studies we did not examine, it is possible. Finally, we must note that the existence of S2 abilities in addition to *g* is strongly supported by improvement in model fit. This is not necessarily a limitation of our work—we simply note this because it could be argued that scores representing these abilities should be interpreted based on evidence of improved model fit. As we have repeatedly stated throughout this article, the fact that an ability exists is not sufficient evidence of the ability's clinical importance. Even interpretive relevance is insufficient , since a score may have interpretive relevance yet lack utility for intended purposes such as diagnosis or improving treatment outcomes. Meaningful interpretation of test scores also requires evidence to support the relevance and utility of these scores for specific purposes in particular applied settings (AERA/APA/NCME, 2014; Messick, 1995).

Future Directions

Despite the absence of validity evidence to support score interpretations and uses, practitioners are routinely encouraged to assess S2—and even S1—abilities. This encouragement comes from both test publishers and general score interpretation systems. Carroll (1993) questioned the fact that test manuals often emphasize the interpretation of S2 abilities over *g*, and wrote that Frank (1983) was essentially correct in stating:

> The Wechsler tests are like the dinosaur, too large, cumbersome and ill-fitted and awkward in the age in which they developed, unable to remain viable in a psychometric age which has passed it by in conceptualization. As with the dinosaur it is time for the Wechsler test to become extinct (p. 126).

Although we do not necessarily espouse the sentiments of Frank, we do note that decision to align the fifth edition of Wechsler Intelligence Scale for Children with CHC theory has done little to improve the structural clarity of the instrument. Instead, it has increased its complexity (Canivez & Watkins, 2016; Dombrowski, Canivez & Watkins, 2017). Thus, perhaps it is time to revisit recent efforts to move the Wechsler scales and other clinical intelligence scales in the direction of permitting the interpretation of more S2 abilities.

Our results provide further evidence against the bloated nature of many current cognitive assessment practices that require administering many subtests and can take clinicians hours to complete, score, and interpret (e.g., Glutting, Watkins, & Youngstrom, 2003). Such assessment and interpretive complexity is done to enable clinicians to de-emphasis *g* in lieu of emphasizing abilities at lower strata. We suggest that test users consider evidence of interpretative relevance and clinical utility before engaging in these activities. The administration of superfluous cognitive tasks that yield scores of dubious clinical value consumes time that would be better spent on evidence-based practices (Yates & Taub, 2003). It is possible that stronger evidence supporting interpretation of S1 and S2 abilities may emerge, but given results from more than a century of psychometric research it is unlikely that different results will occur without a precipitating change in the construction of cognitive tests and conceptualization of cognitive assessment (Cucina & Howardson, 2017). Instead, it is likely best to follow what the extant research supports, which is to focus test score interpretation mainly on measures of *g* (e.g., Canivez, 2013; Dombrowski, 2015b; Kranzler & Floyd, 2013).

References

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

Beauducel, A. (2011). Indeterminacy of factor score estimates in slightly misspecified confirmatory factor models. *Journal of Modern Applied Statistical Methods*, *10*, 583-598.

Beaujean, A. A. (2015). John Carroll's views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence*, *3*, 121-136.

Beaujean, A. A. (in press). Cattell-Horn-Carroll theory of intelligence. In B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement and evaluation*. Thousand Oaks, CA: Sage.

Benson, N. F., Kranzler, J. H., & Floyd, R. G. (2016). Examining the integrity of measurement of cognitive abilities in the prediction of achievement: Comparisons and contrasts across variables from higher-order and bifactor models. *Journal of School Psychology*, *58*1-19. doi:10.1016/j.jsp.2016.06.001

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling*, *22*, 504-516. doi:10.1080/10705511.2014.938596

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Canivez, G. L. (2013). Psychometric versus actuarial interpretation of intelligence and related aptitude batteries. In D. H. Saklofske, C. R. Reynolds, & V. L. Schwean (Eds.), The

Oxford handbook of child psychological assessments (pp. 84–112). New York, NY: Oxford University Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY, US: Cambridge University Press. doi:10.1017/CBO9780511571312

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research*, *30*, 429-452. doi:10.1207/s15327906mbr3003_6

Carroll, J.B. (1997). Theoretical and technical issues in identifying a factor of general intelligence. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success: Scientists respond to the bell curve* (pp. 466-470). New York, NY: Springer-Verlag.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5-21). New York, NY: Pergamon.

Carroll, J. B. (2012). The Three-Stratum Theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 883-890). New York, NY, US: Guilford Press.

Carroll, J. B., & Schweiker, R. F. (1951). Factor analysis in educational research. *Review of Educational Research, 21*, 368-388. doi:10.3102/00346543021005368

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40*, 153-193. doi:10.1037/h0059973

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46*, 648-659. doi:10.1037/0022-006X.46.4.648

Christal, R. E. (1958). Factor analytic study of visual memory. *Psychological Monographs: General and Applied*, *72*(13), 1-24. doi:10.1037/h0093792

Cucina, J. M., & Howardson, G. N. (2017). Woodcock-Johnson III, Kaufman Adolescent and Adult Intelligence Test (KAIT), Kaufman Assessment Battery for Children (KABC), and Differential Ability Scales (DAS) support Carroll but not Cattell-Horn. *Psychological Assessment, 29,* 1001-1015. doi: 10.1037/pas0000389

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327. doi:10.1037/0033-2909.105.2.317

Dombrowski, S. C. (2013). Investigating the structure of the WJ-III at school age. *School Psychology Quarterly, 28,* 154-169.

Dombrowski, S. C. (2014a). Exploratory Bifactor Analysis of the WJ-III Cognitive in Adulthood via the Schmid–Leiman Procedure. *Journal of Psychoeducational Assessment, 32,* 330-341. *doi:10.1177/0734282913508243*

Dombrowski, S. C. (2014b). Investigating the structure of the WJ III Cognitive in early school age through two exploratory bifactor analysis procedures. *Journal of Psychoeducational Assessment, 32, 483-494. doi:10.1177/0734282914530838*

Dombrowski, S. C. (2015a). Exploratory bifactor analysis of the WJ III Achievement at school age via the Schmid-Leiman procedure. *Canadian Journal of School Psychology, 30, 34-50*.

Dombrowski, S. C. (2015b). *Psychoeducational assessment and report writing.* New York, NY: Springer Science.

Dombrowski, S. C., McGill, R. J., & Canivez, G. L. (2017a). Exploratory and hierarchical factor analysis of the WJ IV Cognitive at school age. *Psychological Assessment, 29*, 294-407. Advance online publication. http://dx.doi.org/10.1037/pas0000350

Dombrowski, S. C., McGill, R. J. & Canivez, G. L. (2017b). Exploratory and hierarchical factor analysis of the WJ IV Full Test battery. *School Psychology Quarterly*, doi:10.1037/spq0000221

Dombrowski, S. C. & Watkins, M. W. (2013). Exploratory and Higher Order Factor Analysis of the WJ-III Full Test Battery: A School-Aged Analysis. *Psychological Assessment, 25,* 442-455. *doi:10.1037/a0031335*

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299. doi:10.1037/1082-989X.4.3.272

Flanagan, D. P., Alfonso, V. C., & Ortiz, S. O. (2012). The cross-battery assessment approach: An overview, historical perspective, and current directions. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 459-483). New York, NY, US: Guilford Press.

Fogarty, G. (1987). Timesharing in relations to broad ability domains. *Intelligence*, *11*, 207-231. doi:10.1016/0160-2896(87)90007-9

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, *35*, 169-182. doi:10.1016/j.intell.2006.07.002

Frank, G. (1983). *The Wechsler enterprise*. New York, NY: Pergamon.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18*, 3-12. doi:10.1177/1088868313507536

Gignac, G. E. (2008). Higher-order models versus direct hierarchical models: g as superordinate or breadth factor? *Psychology Science*, *50*, 21-43.

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross-battery abilitiy assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed; pp. 343-374). New York: Guilford.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (2003). Factor analysis. In J. A. Schinka & F. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 143-164). Hoboken, NJ: John Wiley.

Gustafsson, J. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179-203. doi:10.1016/0160-2896(84)90008-4

Hakstian, A. R., & Cattell, R. B. (1978). Higher-stratum ability structures on a basis of twenty primary abilities. *Journal of Educational Psychology*, *70*, 657-669. doi:10.1037/0022-0663.70.5.657

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Joreskog* (pp. 195-216). Lincolnwood, IL: Scientific Software International.

Horn, J. L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities.* Unpublished doctoral dissertation, University of Illinois, Champaign.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock, *WJ-R technical manual* (pp. 197-232). Chicago: Riverside.

Horn, J. L. (1998). A basis for research on age differences in cognitive abilities. In J. J. McArdle & R. W. Woodcock, Eds.), H*uman cognitive abilities in theory and practice* (pp. 57-92). Mahwah, NJ: Erlbaum.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, *57*, 253-270.

Horn, J. L., & Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence*, *6*(2), 165-185. doi:10.1016/0160-2896(82)90012-5

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment*, *15*, 446 – 455.

Jensen, A. R. (2004). Obituary—John Bissell Carroll. *Intelligence*, *32*(1), 1-5.

Keith, T. Z., Caemmerer, J. M., & Reynolds, M. R. (2016). Comparison of methods for factor extraction for cognitive test-like data: Which overfactor, which underfactor? *Intelligence*, *54*, 37-54. doi:10.1016/j.intell.2015.11.003

Keith, T. Z., & Reynolds, M. R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, *47*, 635-650.

Kranzler, J. H., Benson, N., & Floyd, R. G. (2015). Using estimated factor scores from a bifactor analysis to examine the unique effects of the latent variables measured by the WAIS-IV on academic achievement. *Psychological Assessment, 27*, 1402-1416. doi:10.1037/pas0000119

Kranzler, J. H., & Floyd, R. G. (2013). *Assessing intelligence in children and adolescents: A practical guide.* New York, NY: Guilford Press.

Kranzler, J. H., Floyd, R. G., Benson, N., Zaboski, B., & Thibodaux, L. (2016). Cross-Battery Assessment pattern of strengths and weaknesses approach to the identification of specific learning disorders: Evidence-based practice or pseudoscience? *International Journal of School & Educational Psychology, 4*, 146-157. doi:10.1080/21683603.2016.1192855

Loehlin, J. C., & Beaujean, A. A. (2016). *Latent variable models: An introduction to factor, path, and structural equation analysis* (5th ed.). New York, NY: Routledge.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology, 51*, 405-444.

Lubinski, D. (2004). John Bissell Carroll (1916-2003). *American Psychologist, 59*, 43-44. doi:10.1037/0003-066X.59.1.43

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement, 29*, 65-81. doi:10.1177/0146621604270882

Maki, K. E., Floyd, R. G., & Roberson, T. (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly, 30*, 457–469.

McClain, A. L. (1996). Hierarchical analytic methods that yield different perspectives on dynamics: Aids to interpretation. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 229--240). Bingley, England: Emerald Group Publishing.

McGill, R. J. (2017). Re(examining) relations between CHC broad and narrow abilities and reading achievement. *Journal of Educational and Developmental Psychology*, *7*, 265-282. doi: 10.5539/jedp.v7nip265

McGill, R. J., & Busse, R. T. (2017). When theory trumps science: A critique of the PSW model for SLD identification. *Contemporary School Psychology*, *21*, 10-18.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (2nd ed., pp. 504–520). New York, NY: Guilford Press.

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*, 1-10. doi:10.1016/j.intell.2008.08.004

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108-141. doi:10.1207/s15327965pli0102_1

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. A*merican Psychologist*, *50*, 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431-462.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.) Los Angeles, CA: Muthén & Muthén.

Ortiz, S. O. (2015). CHC theory of intelligence. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Handbook of intelligence: Evolutionary theory, historical perspectives, and current concepts* (pp. 209-228). New York: Springer.

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation, 17 (15)*, 1-8.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528-530. doi: 10.1177/1745691612465253

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*, 347-352. doi: 10.1126/science.146.3642.347

Popper, K. (2002). *Conjectures and refutations: The growth of scientific knowledge*. New York: Routledge.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544-559. doi:10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling a bifactor perspective. *Educational and Psychological Measurement, 73,* 5-26.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*, 223-237. doi:10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137-150. doi:10.1037/met0000045

Schneider, W. J., & Flanagan, D. P. (2015). The relationship between theories of intelligence and intelligence tests. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Hanbook of intelligence: Evolutionary theory, historical perspectives, and current concepts* (pp. 317-340). New York: Springer.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan and P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 99-144). New York, NY, US: Guilford Press.

Schneider, W. J., & McGrew, K. S. (in press). 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.). New York, NY: Guilford Press.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock–Johnson IV.* Rolling Meadows, IL: Riverside.

Sung, Y. H., & Dawis, R. V. (1981). Level and factor structure differences in selected abilities across race and sex groups. *Journal of Applied Psychology*, *66*, 613-624. doi:10.1037/0021-9010.66.5.613

Thurstone, L. L., & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, No. 2.

Undheim, J. O. (1981). On intelligence: I. Broad ability factors in 15-year-old children and Cattell's theory of fluid and crystallized intelligence. *Scandinavian Journal of Psychology*, *22*, 171-179. doi:10.1111/j.1467-9450.1981.tb00391.x

Wagenmakers, E., & Farrell, S. (2004). *Psychonomic Bulletin & Review, 11,* 192-196.

Wechsler, D. (2014). *Wechsler intelligence scale for children* (5th ed.). Bloomington, MN: NCS Pearson.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised.* Chicago: Riverside.

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

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment*, *15,* 478-495. doi: 10.1037/1040-3590.15.4.478

Table 1

Descriptive Information for Samples

Study	**Number of Cognitive Variables**	**n**	**Mean Age (*SD*)**	**% Female**	**Description of Participants**
Christal (1958)	29	718	19.35 (2.32)	0*	Newly enlisted Air Force personnel
Fogarty (1987)	48	126	26 (9.9)	≈ 52	Australian adults/college students
Gustafsson (1984)	20	981	12 (NA)	≈ 51	Swedish students from 50 6th grade classrooms
Hakstian & Cattell (1978)	20	280	17 (.82)	≈ 51	Canadian students from six high schools
Horn (1965)	31	297	27.6 (10.6)	≈ 28	Prisoners and persons on unemployment rolls
Horn & Stankov (1982)	19	241	26 (NA)	0	Prisoners
Undheim (1981)	21	148	15 (NA)	≈ 70	Norwegian eighth- and ninth-grade students

*Notes.** While sex is not mentioned in this article, the number of females is likely zero or near zero given enlistment policies at the time this study was completed.

Table 2

Summary of Model Comparisons

Model	BIC	w_i (BIC)	X^2 (*df*)	CFI	RMSEA (90% CI)	SRMR
			Christal (1958)			
1. Single factor (*g*)	81616.27	<.001	3954.76 (377)	.605	.115 (.112-.118)	.098
2. Bifactor (g + *Gy*)	80095.44	<.001	2328.70 (361)	.783	.087 (.084-.091)	.064
3. Model 2 + *Gv*	79779.31	<.001	1979.69 (356)	.821	.080 (.076-.083)	.061
4. Model 3 + G*c*	79567.09	1.00*	1747.74 (353)	.846	.074 (.071-.078)	.058
			Fogarty (1987)			
1. Single factor (*g*)	29493.69	<.001	2287.09 (1080)	.618	.094 (.089-.100)	.092
2. Bifactor (g + *Gc*)	29474.01	<.001	2214.21 (1069)	.637	.092 (.087-.098)	.090
3. Model 2 + *Gf*	29438.40	<.001	2144.74 (1062)	.657	.090 (.084-.095)	.089
4. Model 3 + *Gv*	29368.44	<.001	2050.60 (1057)	.658	.086 (.081-.092)	.087
5. Model 4 + *Gu*	29215.64	1.00*	1859.11 (1049)	.743	.078 (.072-.084)	.083
			Gustafsson (1984)			
1. Single factor (*g*)	49663.10	<.001	3873.85 (170)	.628	.149 (.145-.153)	.093
2. Bifactor (g + *Gv*)	48946.65	<.001	3109.18 (163)	.704	.136 (.132-.140)	.086
3. Model 2 + *Gf*	48219.50	<.001	2375.15 (162)	.778	.118 (.114-.122)	.082
4. Model 3 + *Gc*	47333.46	1.00*	1440.89 (155)	.871	.092 (.088-.096)	.060
			Hakstian & Cattell (1978)			
1. Single factor (*g*)	32084.71	<.001	367.06 (170)	.802	.064 (.055-.073)	.062
2. Bifactor (g + *Gf*)	32055.40	<.001	309.58 (165)	.854	.056 (.046-.056)	.057
3. Model 2 + *Gr*	32034.04	<.001	260.04 (160)	.899	.047 (.037-.058)	.051
4. Model 3 + *Gv*	32022.83	<.001	243.20 (159)	.915	.043 (.032-.054)	.049
5. Model 4 + *Gc*	32017.04	<.001	231.77 (158)	.926	.041 (.029-.052)	.048
6. Model 5 + *Gy*	32008.71	1.00*	217.80 (157)	.939	.037 (.024-.049)	.047
7. Model 5 + *Gs*	32018.55	.007	205.11 (153)	.948	.035 (.021-.047)	.045
			Horn (1965)			
1. Single factor (*g*)	24571.62	<.001	1838.17 (434)	.568	.104 (.099-.109)	.098
2. Bifactor (*g* + *Gf*)	24541.82	<.001	1768.51 (427)	.587	.103 (.098-.108)	.098
3. Model 2 + *Gc*	24467.90	<.001	1666.12 (422)	.617	.100 (.095-.105)	.097
4. Model 3 + *Gv*	24414.55	<.001	1584.31 (417)	.641	.097 (.092-.02)	.096
5. Model 4 + *Gs*	24251.97	*1.00**	1393.26 (412)	.698	.090 (.084-.095)	.103
6. Model 5 + *Gr*†	-	-	-	-	-	-
			Horn & Stankov (1982)			
1. Single factor (*g*)	11949.47	<.001	698.05 (152)	.690	.122 (.113-.131)	.097
2. Bifactor (*g* + *Gf)*	11916.83	<.001	643.48 (148)	.719	.118 (.109-.127)	.095
3. Model 2 + *Gc*	11815.84	<.001	515.05 (143)	.789	.104 (.094-.114)	.090
4. Model 3 + *Gv*	11802.33	<.001	485.09 (140)	.804	.101 (.091-.111)	.087
5. Model 4 + *Gu*	11761.79	.622	428.10 (137)	.835	.094 (.084-.104)	.078
6. Model 5 + *Gy*	11763.37	.378	424.19 (136)	.836	.094 (.084-.104)	.077

Table 2 (Continued)

Model	**BIC**	w_i **(BIC)**	X^2 **(*df*)**	**CFI**	**RMSEA (90% CI)**	**SRMR**
			Undheim (1981)			
1. Single factor (*g*)	18782.94	<.001	595.62 (189)	.720	.121 (.110-.132)	.091
2. Bifactor (*g* + *Gf)*	18789.09	<.001	571.79 (183)	.732	.120 (.109-.1310	.089
3. Model 2 + *Gv*	18773.83	<.001	541.54 (180)	.751	.116 (.105-.128)	.087
4. Model 3 + *Gs*	18687.80	<.001	435.52 (176)	.821	.100 (.088-.112)	.078
5. Model 4 + *Gc*	18678.71	<.001	416.44 (174)	.833	.097 (.085-.109)	.074
6. Model 5 + *Gr*	18647.17	1.00*	364.91 (170)	.866	.088 (.076-.100)	.069

Notes. *g* = general intelligence, *Gc* = crystallized intelligence, *Gf* = fluid intelligence, *Gr* = broad retrieval ability, *Gs* = broad cognitive speediness, *Gu* = broad auditory ability, *Gv* = broad visual perception, *Gy* = general memory and learning. *Model was identified as best among the set of alternatives tested. †Model could not be identified as it was based on a single indicator/observed variable.

Table 3

Stratum Two Abilities Agreement Between Carroll's (1993) Exploratory Factor Analysis and Confirmatory Factor Analysis

	Gc		***Gf***		***Gr***		***Gs***		***Gu***		***Gv***		***Gy***	
Study	EFA	CFA	EFA	CFA	EFA	CFA	EFA	CFA	EFA	CFA	EFA	CFA	EFA	CFA
Christal (1958)	Y	**N**	-	-	-	-	-	-	-	-	Y	**N**	Y	Y
Fogarty (1987)	Y	Y	Y	Y	-	-	-	-	Y	Y	Y	Y	-	-
Gustafsson (1984)	Y	Y	Y	Y	-	-	-	-	-	-	Y	Y	-	-
Hakstian & Cattell (1978)	Y	Y	Y	Y	Y	Y	Y	**N**	-	-	Y	Y	Y	Y
Horn (1965)	Y	Y	Y	Y	Y	**N**	Y	Y	-	-	Y	Y	-	-
Horn & Stankov (1982)	Y	Y	Y	Y	-	-	-	-	Y	Y	Y	Y	Y	**N**
Undheim (1981)	Y	Y	Y	Y	Y	Y	Y	Y	-	-	Y	Y	-	-

Notes. EFA= Exploratory factor analysis, CFA=Confirmatory factor analysis, Y = factor supported, N = factor not supported, *Gc* = crystallized intelligence, *Gf* = fluid intelligence, *Gr* = broad retrieval ability, *Gs* = broad cognitive speediness, *Gu* = broad auditory perception, *Gv* = broad visual perception, *Gy* = general memory and learning.

Table 4

Interpretive Relevance Indices

Factor	ARPB	ECV	FD	H	ω	ωH	ωHS	PUC
				Christal (1958)				
Model	.256	.569	-	-	-	-	-	.660
g	-	-	.946	.925	.932	.704	-	-
Gc	-	-	.791	.504	.869	.314	.361	-
Gv	-	-	.841	.679	.853	.341	.399	-
Gy	-	-	.901	.820	.875	.547	.625	-
				Fogarty (1987)				
Model	.098	.710	-	-	-	-	-	.889
g	-	-	.972	.959	.960	.903	-	-
Gc	-	-	.804	.603	.913	.149	.163	-
Gf	-	-	.857	.585	.862	.237	.275	-
Gu	-	-	.901	.779	.847	.571	.674	-
Gv	-	-	.853	.661	.792	.482	.608	-
				Gustafsson (1984)				
Model	.185	.595	-	-	-	-	-	.853
g	-	-	.936	.907	.929	.800	-	-
Gc	-	-	.901	.777	.902	.437	.484	-
Gf	-	-	.887	.667	.895	.552	.617	-
Gv	-	-	.886	.682	.843	.294	.349	-
				Hakstian & Cattell (1978)				
Model	.098	.576	-	-	-	-	-	.879
g	-	-	.902	.841	.850	.750	-	-
Gc	-	-	.579	.310	.506	.269	.532	-
Gf	-	-	.708	.498	.752	.280	.372	-
Gr	-	-	.760	.534	.608	.337	.555	-
Gv	-	-	.657	.421	.441	.386	.875	-
Gy	-	-	.601	.328	.532	.284	.534	-
				Horn (1965)				
Model	.776	.575	-	-	-	-	-	.890
g	-	-	.946	.917	.880	.756	-	-
Gc	-	-	.587	.299	.706	.168	.238	-
Gf	-	-	.844	.654	.863	.240	.278	-
Gs	-	-	.864	.723	.786	.562	.715	-
Gv	-	-	.836	.659	.815	.482	.591	-

Table 4 (Continued)

Factor	ARPB	ECV	FD	H	ω	ωH	ωHS	PUC
				Horn & Stankov (1982)				
Model	.059	.624	-	-	-	-	-	.830
g	-	-	.939	.905	.901	.799	-	-
Gc	-	-	.820	.593	.864	.246	.285	-
Gf	-	-	.857	.528	.801	.204	.255	-
Gu	-	-	.843	.702	.763	.308	.404	-
Gv	-	-	.696	.464	.543	.385	.709	-
				Undheim (1981)				
Model	.059	.660	-	-	-	-	-	.852
g	-	-	.954	.928	.938	.859	-	-
Gc	-	-	.726	.426	.849	.185	.218	-
Gf	-	-	.604	.248	.852	.043	.050	-
Gr	-	-	.781	.576	.730	.446	.611	-
Gs	-	-	.842	.660	.816	.475	.582	-
Gv	-	-	.703	.411	.758	.278	.367	-

Notes. *g* = general intelligence, *Gc* = crystallized intelligence, *Gf* = fluid intelligence, *Gr* = broad retrieval ability, *Gs* = broad cognitive speediness, *Gu* = broad auditory ability, *Gv* = broad visual perception, *Gy* = general memory and learning, ARPB = average relative parameter bias, ECV = explained common variance for *g*, PUC = percent of uncontaminated correlations, FD = factor determinacy, H = construct replicability, **ω** = model-based estimate of internal consistency for unit-weighted composite score, **ω**H = percentage of variance attributable to *g*, **ω**HS = percentage of unique, reliable variance for group factor that is independent of *g*.