Online Cognitive Assessment in the Era of COVID-19: Examining the Validity of the MEZURE

Stefan C. Dombrowski[1]
Rider University

A. Alexander Beaujean[2]
Baylor University

Ryan J. McGill[3]
William & Mary

Ryan L. Farmer[4]
The University of Memphis

Author Note

[1]Stefan C. Dombrowski https://orcid.org/0000-0002-8057-3751
[2]A. Alexander Beaujean https://orcid.org/0000-0001-7007-7968
[3]Ryan J. McGill https://orcid.org/0000-0002-5138-0694
[4]Ryan L. Farmer https://orcid.org/0000-0003-1409-7555

Author Contribution Statement: Stefan Dombrowski was responsible for the design, data analysis, conceptualization, and writing of the manuscript. All additional authors were involved in the conceptualization, interpretation of data, and writing of the manuscript.

Conflict of Interest Disclosure: None.

Study was not preregistered.

Data are available in Clinical Manual. Code is widely available in many resources on structural validity analyses/factor analysis.

Correspondence concerning this article should be addressed to Stefan C. Dombrowski, Ph.D., Department of Graduate Education, Leadership and Counseling, 2083 Lawrenceville Road, Lawrenceville, NJ 08648. Email: sdombrowski@rider.edu.

Abstract

Developed more than two decades ago, the MEZURE (Assessment Technologies, Inc, 1995-2020; https://www.mezure.com/) has received increased attention as a result of the COVID-19 pandemic. It is the first individualized test of cognitive ability created to use an online (local or remote) assessment modality. The MEZURE claims to be aligned both with extended Gf-Gc theory as well as the Cattell-Horn-Carroll (CHC) model of abilities. Whereas the test publisher claims it used exploratory factor analysis (EFA) to investigate the instrument's factor structure, only the subtest factor loadings on the Gf-Gc factors were furnished. No other structural validity information was provided suggesting that users of the instrument should interpret the scores produced by the MEZURE with caution. Accordingly, the present study used exploratory and confirmatory factor analysis (CFA) to more fully investigate the structural validity of the MEZURE. The results revealed that the MEZURE contains a combined perceptual reasoning [i.e., (Gf/Gv)/working memory (Gwm)] group factor, a verbal ability group factor, and a relatively weak general factor that is dominated by perceptual reasoning. The finding of a paltry general factor that is weakly loaded by verbal subtests is inconsistent with the broader research on traditional cognitive ability assessment and could be related to the online administration format of the test. Future research is required to better understand this finding.

Keywords: COVID-19; Factor Analysis; General Intelligence; Remote Testing; IQ Testing; Telehealth; Teleassessment

Statement

The MEZURE has received increased attention following the COVID-19 pandemic and this study produced findings that psychologists should more fully understand (i.e., the general factor is reflective of nonverbal content, not verbal abilities, and is weak). The results defy historical precedent in cognitive assessment, could be related to the online/remote modality of the instrument, and suggest that the instrument be cautiously interpreted primarily as a measure of nonverbal abilities.

Online Cognitive Assessment in the Era of COVID-19: Examining the Validity of the MEZURE

As a consequence of the social distancing recommendations for containing the COVID-19 pandemic, psychologists necessarily began increasing the assessments they provided via telehealth technology (Farmer et al., 2020). This raised a number of ethical, legal, and psychometric issues (Farmer et al., 2021). Authors of many commonly employed instruments implemented technology to allow for remote use despite the lack of norming for telehealth. Other instruments were specifically designed for the purpose of remote administration, but have been largely neglected in the research literature. One instrument that falls in this latter class is the MEZURE^{TM} (Assessment Technologies, Inc, 1995-2020; https://www.mezure.com/).[1]

Created over two decades ago, the MEZURE is an automated computer administered intelligence instrument that is administered either locally or remotely through the Internet. A previous review noted some positive aspects of the instrument (Dombrowski, Engle & Lennon, 2022), but questions remain about whether the instrument measures its intended constructs.

Overview of the MEZURE

The MEZURE's publisher is somewhat ambiguous about the purpose of the instrument. One the one hand the publisher claims it is designed for measurement. On the other hand the publisher claims the scores can be used for the pragmatic purpose of classification/diagnosis (Assessment Technologies, Inc, 2020a, p. 3). While the two purposes are not mutually exclusive, they do require different forms evidence to support the claims (Hand, 2004; Newton, 2017). Since measurement involves depicting an attribute's manifestations, support of measurement claims requires validity evidence (Borsboom et al., 2004). Pragmatic purposes involve making

[1] Although the publisher does not discuss the origin of the appellation, MEZURE, it appears to be a non-standard spelling of the term *measure* rather than an acronym—despite always being presented in all uppercase letters. Additionally, the MEZURE is also known on the publisher's website as the C.O.M.I.T but they are the same instruments containing the same items, User's Manual, and Clinical Manual.

decisions (e.g., diagnostic decisions), so support of these claim requires evidence of score utility (e.g., sensitivity, cost-benefit).

The utility evidence the MEZURE publisher provided in the Clinical Manual consists solely of basic descriptive statistics (e.g., means, standard deviation) of the scores for respondents classified as having an intellectual disability, giftedness, or a specific learning disability (Assessment Technologies, 2020a, pp. 50–52).[2] While interesting, this information is insufficient to support claims that the MEZURE scores have diagnostic utility (McFall & Treat, 1999). This aspect of validity deserves further investigation, but will not be evaluated within this article. Instead, this article focuses on evaluating the evidence supporting the publisher's measurement claims via an explication of structural validity.

MEZURE as a Measurement Instrument

The MEZURE is designed to measure "…a broad range of cognitive abilities as represented in current theories of human intelligence" (Assessment Technologies, 2020a, p. 3). It consists of 14 subtests organized into two batteries: screening and standard. The screening battery consists of four subtests and is intended to be an "expeditious measure of general functioning" (Assessment Technologies, Inc, 2020a, p. 6). The standard battery consists of three additional subtests (seven total) and is designed to be a "comprehensive measure of an individual's current intellectual functioning in both fluid and crystallized domains . . . appropriate for clinical and psychoeducational purposes" (Assessment Technologies, Inc, 2020a, p. 7). In addition to the two batteries' subtests, the MEZURE contains five additional subtests for children and seven additional subtests for adults that reflect "distinct processing modalities that may prove helpful in in-depth psychoeducational, neuropsychological, or clinical assessments" (Assessment Technologies, Inc, 2020a, p. 54). Since the publisher of the MEZURE discusses the

[2] The MEZURE's publisher included the instrument's technical information (e.g., validity/reliability) and normative characteristic in its "clinical manual."

standard battery extensively in the Clinical Manual, and does not include the additional subtests in the correlation matrix furnished, this article focuses on the standard battery's subtests and the scores derived from them.

The publisher's measurement claims regarding the standard battery are confusing. On the one hand, the publisher states that the instrument provides a "comprehensive measure of general intelligence" (Assessment Technologies, Inc, 2020a, p. 2). On the other hand, the publisher states that it created the instrument based on extended Gf-Gc theory and Cattell-Horn-Carroll theory (Assessment Technologies, Inc, 2020a, p. 3). The creators of both theories are explicit in eschewing the concept of general intelligence (Beaujean & Benson, 2019). Thus, it would be theoretically inconsistent to employ those theories to create an instrument intended to measure general intelligence.

The MEZURE's publisher claims that the seven subtests of the standard battery capture four intellective abilities: fluid reasoning, crystallized reasoning, short-term memory, and visual processing (Assessment Technologies, Inc, 2020a, pp. 8-9).[3] The subtests and the abilities they are designed to capture are presented in Table 1 while subtest descriptions are available in the manual. Except for the Visual Memory subtest, the subtests' items all have a selected response format in which respondents select from among a set of four-to-six response options. Second, there are no subtests designed to capture fluid reasoning (Gf), visual processing (Gv), or short-term memory (Gwm) alone (see Table 1). Instead, the subtests are a blend of each of these attributes.

Although the MEZURE's publisher allows for interpreting the individual subtests, they suggest it should only be done when there are statistically significant performance differences

[3] There is no concept called c*rystallized reasoning* in either the extended Gf-Gc theory or the Cattell-Horn-Carroll theory. The MEZURE publisher's definition of the term is consistent with the *acculturation knowledge* concept in extended Gf-Gc theory and the *comprehension-knowledge* concept in Cattell-Horn-Carroll theory.

across the subtests (Assessment Technologies, Inc, 2020a, pp. 26-29). If that is not the case, then the publisher recommends interpreting Fluid IQ Scale and Crystallized IQ Scale scores[4]. If the differences between the Fluid IQ Scale and Crystallized IQ Scale scores is not statistically different either, then the publisher recommends interpreting the Composite IQ score because it is "the most reliable and valid measure of a youngster’s global cognitive functioning" (p. 29). However, there is no mention in the MEZURE’s Clinical Manual of the reliability of these subtest-to-subtest or composite-to-composite difference scores; thus, it is not immediately clear whether such differences are robust to standard error in the measures.

The composition of the MEZURE's Fluid IQ Scale, Crystallized IQ Scale, and Composite IQ scores is presented in Table 2. Their creation and meaning, however, is confusing. First, although the publisher recommends interpreting the Composite IQ score in the presence of minimal score differences, the publisher does not state it is a measure of general intelligence. Instead, it is described as simply a "summative index of general intellectual functioning" (Assessment Technologies, Inc, 2020a, p. 29). While this is consistent with extended Gf-Gc (EFC; Horn, 1965b) and Cattell–Horn–Carroll (CHC; Schneider & McGrew, 2018) theories, it means the score's interpretation is limited to pragmatic purposes (Spearman, 1931); that is, it does not measure a clear theory-determined construct, but instead may only serve as data to facilitate decision-making processes.

Second, the Fluid IQ Scale and Crystallized IQ Scale scores appear to be derived empirically rather than based on the meaning of the Gf or Gc concepts. Specifically, the publisher states, "assignment of each subtest to either the Fluid or Crystallized Scales is based on empirical validation via subtest intercorrelations and subsequent factor analysis" (Assessment

[4] Please note that the practice of eschewing interpretation of the general factor in the presence of a statistically significant discrepancy between index scores is considered inappropriate (see Dombrowski et al., 2022; McGill et al., 2018).

Technologies, Inc, 2020a, p. 9). This is concerning. While it is legitimate to use empirical investigations for testing hypotheses about the instrument scores or creating pragmatic scores, empirical investigations cannot tell us what instrument scores represent (Guttman, 1977). The meaning of an instrument's scores is a conceptual issue (Krause, 2012).

Accordingly, it appears that the MEZURE’s publisher employed factor analysis for testing hypotheses about the standard battery's scores rather than to determine the meaning of scores. Specifically, it appears the publisher tested the following two hypotheses: (1) whether performance on the subtests designed to measure fluid reasoning cohere to the degree that Fluid IQ Scale represents fluid reasoning with fidelity; and (2) whether performance on the subtests designed to measure crystallized reasoning cohere to the extent that the Crystallized IQ Scale represents crystallized reasoning with fidelity.

The evidence the MEZURE’s publisher used to test the above hypotheses is insufficient. The publisher conducted an “exploratory factor analysis with rotation” using the correlations among the seven standard battery subtests for the entire norming sample (Assessment Technologies, Inc, 2020a, p. 10). The publisher then presented the loadings from a single solution with two factors, which they interpreted as corroborating the two hypotheses. Absent from the presentation is important information, such as the form of extraction (e.g., principal axis, least squares), the criteria for determining the number of factors to extract (e.g., Horn’s Parallel Analysis; minimum average partial test), and the criterion for the oblique rotation (e.g., promax, oblimin; Preacher & MacCallum, 2003; Rodriguez et al., 2016). Decisions in all of these areas can strongly influence interpretation, which subsequently influences whether an investigator believes the hypotheses are corroborated or falsified (Loehlin & Beaujean, 2016).

Given the totality of the omissions in the MEZURE’s Clinical Manual, the present study sought to understand the theoretical and applied structure of the MEZURE using best practice in factor analytic methodology. Considering the increased emphasis on online or remote administration following the COVID-19 pandemic, a more thorough evaluation will likely benefit users as well as stakeholders interested in selecting and using instruments with strong validity evidence.

Two exploratory factor analytic (EFA) procedures that are useful for investigating the structure of cognitive ability instruments include principal axis factoring followed by the Schmid-Leiman procedure (SL, Schmid & Leiman, 1957) and exploratory bifactor analysis (EBFA; Jennrich & Bentler, 2011). Both methodological approaches handle well traits that are correlated and permit an in-depth understanding of an instrument's latent structure by disclosing the relationship between measured and latent variables through the partitioning of variance between the hypothesized general and group factors (Dombrowski, McGill, Canivez et al., 2021). From an applied perspective, this elucidates how much interpretive weight should be placed upon each construct and their associated scores. EFA is often an important first step when developing a new scale or investigating an instrument whose theoretical structure has not been previously subjected to formal empirical analysis. Of course, evidence from EFA is not sufficient in isolation. Other types of evidence, including confirmatory factor analysis and evaluation of relations between scores on the target instrument with other instruments, are necessary. With CFA, competing models (e.g., oblique, higher order, bifactor, unidimensional) may be tested to determine which model produces the best fit with the data informed by the results furnished by previous EFA studies. Additional statistics may be ascertained via methods that offer further insight for users of commercial ability measures. These include omega estimates, *H*, and percentage of uncontaminated correlations which, along with variance apportionment, can be used to ascertain how much interpretive emphasis should be placed upon the general and group factors; and ultimately whether and how the MEZURE should *actually* be interpreted in applied practice. These statistics were not furnished in the MEZURE's Clinical Manual likely because are not commonly calculated by test publishers though their calculation would be helpful (Dombrowski, 2020).

Method

Participants

The MEZURE's Clinical Manual provides the correlation matrix for the seven subtests of the standard battery for the entire norming sample (Assessment Technologies, Inc, 2020a, Table 2.1, p. 10). The study was not preregistered and was considered exempt from institutional review board approval as it is based upon a de-identified correlation matrix. The code used to analyze the correlation matrix is widely available in many resources on structural validity/factor analysis. The data from which the correlations were calculated come from a nationally representative sample of 4,184 individuals from the age of six through adult (Assessment Technologies, Inc, 2020a, p. 39).[5] They were selected to match the 1998 US Census information with respect to parental education, geographic region, ethnicity, sex, residence location (urban, rural), school grade, and school type (e.g., private, public).

Procedure

Although this study sought to evaluate two specific hypotheses, there is substantial confusion about the attributes captured by the MEZURE. Thus, this study initially employed an unrestricted (sometimes called exploratory) factor analyses (see McDonald [1999]). First, Bartlett's Test of Sphericity (Bartlett, 1954) and the Kaiser–Meyer–Olkin (KMO; Kaiser, 1974) statistic were calculated to verify that the correlation matrix is suitable for factor analytic evaluation. Second, multiple empirical criteria (Gorsuch, 2003) as well as factor interpretability were examined to determine the number of factors to extract. Specifically, the number of

[5] The MEZURE's Clinical Manual provides little information about the adult group except that those in the group were over 17 years of age when assessed.

eigenvalues > 1 (Kaiser, 1974), the visual scree test (Cattell, 1966), Horn's parallel analysis (HPA; Horn, 1965), minimum average partials (MAP; Velicer, 1976), and the Bayesian information criterion (BIC) were examined. Pattern coefficient values of .30 or higher were deemed sufficiently salient for interpretation (Child, 2006; Schmitt, 2011)

Third, factors were extracted using principal axis factoring (Cudeck, 2000; Fabrigar, Wegener, MacCallum, & Strahan, 1999). Extraction was followed by a communality sensitivity analysis to ensure the initial communality value specified for the extraction did not have an undue influence on the final loadings (e.g., Dombrowski et al., 2019). This required specifying initial communality starting values (.10 to 1.0 in increments of .10 along with the squared multiple correlation), and then comparing the loadings and final communality estimates across solutions.

Fourth, assuming there was a sufficient rationale for extracting more than one factor, the extraction was rotated to aid in interpretation. Two different rotations were used: bi-geomin (Jennrich & Bentler, 2011) and promax ($k = 4$; Tataryn, Wood, & Gorsuch, 1999). These two rotations were undertaken because they each follow two different traditions in intelligence research (Beaujean & Benson, 2019).

The bi-geomin rotation is designed to comport with the bi-factor model in which a set of indicators (i.e., recorded observations such as subtest scores) are specified to result from both broader and narrower attributes operating more or less independently of each other (Holzinger, Swineford, & Harman, 1937). This model was originally created as an extension of Spearman's two-factor theory (TFT) so it includes one general factor representing an attribute in common with all the indicators along with multiple group factors representing more narrow attributes. Consequently, employing the bi-geomin rotation requires extracting one additional factor (i.e., the general factor) beyond what any criteria indicate extracting. This should not be construed as extracting three group factors but rather one general and two group factors.

The promax rotation follows the Thurstone approach of modelling the intelligence sphere as if it is comprised of a set of multiple primary abilities that is developed and employed independently, but are related to each other. If the correlation among these factors is non-negligible, then it is often possible to extract an additional set of factors, which produces a higher-order model suggesting a superordinate psychological dimension. In the higher-order model (a) the set of indicators are specified to result from multiple related attributes; and (b) those attributes are specified to be the unobserved (unmeasured) effects of more super-ordinate attributes.

Interpreting super-ordinate factors in a higher-order model is notoriously difficult—akin to interpreting shadows of the shadows of mountains rather than the mountains themselves (McClain, 1996). The loadings from higher-order models can be transformed in different ways to aid interpretation of the super-ordinate factors. This study deployed the Schmid-Leiman transformation (Schmid & Leiman, 1957), a commonly used, elegant approach to partition variance among higher and lower order factors.

Following EFA, confirmatory factor analysis (i.e., constrained factor analysis) was used to fit models consistent with TFT, extended Gf-Gc theory, and CHC theory. The TFT models include the following: (1) A single general factor model; (2) a bi-factor model with one group factor; (3a) a bi-factor model with two group factors; and (3b) a bi-factor model with two group factors and Categorization loading solely on the general factor. The extended Gf-Gc/CHC models include the following: (4) a two-factor oblique model; and (5) a two-factor higher order model.

For each model, CFA with maximum likelihood estimation was used. This allowed for the calculation of fit indices for each model. Overall model fit was evaluated using the $\chi 2$ statistic, comparative fit index (CFI), standardized root mean squared residual (SRMR), Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), the Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC). In general, models with better fit have higher CFI and TLI values and lower the $\chi 2$, SRMR, RMSEA, AIC, and BIC values. Typically, models should only be considered if they have CFI and TLI values $\geq .90$ along with SRMR $\leq .09$ and RMSEA $\leq .08$ (Hu & Bentler, 1999). The AIC and BIC do not have a meaningful scale. Within a set of models evaluated with the same data, models with smaller AIC and BIC values are most likely to replicate (Kline, 2016). All EFA and CFA analyses were conducted using **R** (R Core Team, 2022) or Mplus (Muthen & Muthen, 1998-2022). Watkin's (2013) Omega program was used to calculate omega estimates, *H* and PUC.

Results

Results from Bartlett's Test of Sphericity (Bartlett, 1954) indicated that the correlation matrix was significantly different from an identity matrix ($\chi^2 = 13{,}819.03$, $df = 153$, $p < .0001$). The Kaiser–Meyer–Olkin (Kaiser, 1974) statistic was .867, well above the minimum standard of .60 for conducting a factor analysis suggested by Kline (2016). Measures of sampling adequacy for each variable were also within reasonable limits. Thus, the correlation matrix was judged to be appropriate for the factor analytic procedures that were employed.

Factor Extraction Criteria Comparison

The MAP criterion recommended extraction of one factor, while all other criteria suggested two factors. This was interpreted to mean that two factors were required to be extracted for the promax rotation and three for the bi-factor rotation (one general and two group). Next, a two- and three-factor solution was extracted using principal axis factoring and followed by a communality sensitivity analysis for both solutions (Table 3). The three-factor solution would not converge when SMC and .10 were specified as start values. All other three-factor solutions converged, but produced unstable final communality estimates for Categorization ranging from .199 to .998 depending upon start value (see Table 3 for the sensitivity analysis results). This range in final communality estimates is unusual and likely suggests the Categorization subtest is a problematic indicator. The two-factor solution produced stable communality estimates for all start values, but several subtests had unusually low final communality estimates suggesting that they have little in common with other subtests in the MEZURE. Specifically, the Visual Closure and Vocabulary subtests produced final communality estimates of approximately .22 while the Categorization subtest had a final communality estimate of .084 again suggesting it is problematic.

Promax Rotation

Online Supplement Table A1 presents the results from the two-factor extraction with promax rotation. The first factor accounted for approximately 35% of the variance, while the second factor accounted for approximately 17%. The correlation between the two factors is .40, which indicates a higher-order extraction may be possible.

A single second-order factor was able to be extracted using principal axis extraction. This higher-order model was subsequently transformed using the Schmid-Leiman procedure. The results are presented in Table 4. The second-order factor captures approximately 15% of the total variance and 47% of the common variance. This was approximately the same total and common variance as that accounted for by the first factor in the original rotation (13% and 40%, respectively) and higher than that accounted for by second factor in the original rotation (4% and 13%, respectively). The second-order factor accounted for a small (i.e., < 10%) proportion of the individual subtest variance in four of the subtests (Visual Closure, Information, Vocabulary and

Categorization), and somewhat higher amounts in the Auditory Memory, Visual Memory and Analogies subtests (21, 22 and 26%, respectively). The individual subtest variance attributable to group factors ranged from 12% to 28% for factor one and 13% for factor two. As expected, Categorization's low communality suggested it did not have salient loadings on any factor indicating that this subtest has little in common with the rest of the MEZURE's subtests.

There are two things to note about Online Supplement Table A1 and Table 4. First, while factor two comports with the meaning of Gc, factor one does not comport with the consensus meaning of Gf. Instead, it is an amalgam of Gv, Gf, and Gwm. For that reason, this factor was labelled Perceptual Reasoning (PR)/Gwm. Second, the loadings in the MEZURE Clinical Manual match neither the loadings in Online Supplement Table A1 nor Table 4. This is concerning because it is important for third parties to be able to replicate the results of original research (National Academies of Sciences, Engineering, and Medicine, 2019). Thus, the potential reasons for these discrepant results were a target for investigation within this article and are next described.

Replication of Proposed Publisher Structure

The MEZURE's Clinical Manual stated that it used an exploratory (unconstrained) factor analysis with oblique rotation, but neither disclosed the method of factor extraction nor the type of rotation. In attempt to replicate the structure and provide additional factor analytic results the structure presented in the Clinical Manual was replicated by assuming that a common default option in statistical application software (e.g., principal components analysis [PCA] with oblimin rotation) might be the statistical choice. This was indeed found to be the case and the additional results of the publisher's approach are presented in Online Supplement Table A2. Whereas the Clinical Manual only reported the loadings on the two group factors Online Supplement Table A2 provides additional, relevant information from the publisher's presumed analysis including structure coefficients, communalities, uniqueness, eigenvalues, and total variance. It also included the first unrotated loadings, which are considered an approximate proxy for the general factor loadings though PCA is not the best methodological choice for latent variable analysis. The results of our analyses match those in the MEZURE Clinical Manual within rounding error. If it is true that the MEZURE publisher used PCA, then this is troublesome. PCA and EFA are not mathematically exchangeable approaches as they contain different algorithms and different analytical assumptions (Widaman, 2007).

Bi-factor Model

Results of the bi-geomin rotation for a three- factor extraction (i.e., general plus two group factors) are presented in Online Supplement Table A3. This solution is concerning. First, the general factor is not general. It is akin to the Gf/Gv/Gwm factor from the promax rotation. Second, the Gc subtests are spread across the two group factors. One group factor is defined solely by the Categorization subtest, while the other is defined by the Vocabulary and Information subtests.

Given the unexpected results from the three-factor solution, two factors (one general plus one group) were extracted and subsequently rotated using the bi-factor rotation. The results are presented in Online Supplement Table A4. Again, the general factor is not a *general* factor associated commonly with psychometric g but similar to the Gf/Gwm/Gv factor from the promax rotation. The group factor is a verbal factor defined by the Vocabulary and Information subtests. The Categorization subtest had a loading of approximately .20 on both factors.

Conclusions from Exploratory (Unrestricted) Factor Analyses

The unrestricted factor analysis produced somewhat ambiguous results regarding the hypotheses about the MEZURE's subtests. First, the subtests comprising the Fluid IQ scale coalesce onto a single factor. At the same time, across rotations, the subtests primarily requiring memory (Auditory Memory, Visual Memory) contribute approximately equally to the factor as the subtest primarily requiring reasoning (Analogies). Moreover, the subtest primarily requiring visual processing (Visual Closure) contributed strongly to the factor. Thus, the results did not corroborate the hypothesis that the Fluid IQ Scale represents fluid reasoning with fidelity.

Second, the subtests designed to measure crystallized reasoning do not cohere well. While performance on the Information and Vocabulary subtests tend to go together, performance on the Categorization subtest does not, which is disconcerting. In fact, performance on the Categorization subtest is not really related to performance on any other subtest in the standard battery. Thus, the results failed to corroborate performance on the subtests designed to measure crystallized reasoning suggesting that the Crystallized IQ scale likely does not represent crystallized reasoning with fidelity.

It is noted that an exploratory analysis[6] is defined more by an orientation rather than a specific basked of techniques (Tukey, 1977). It is an approach to data in which scientists metaphorically take on the role of detective in order to identify patterns in a dataset rather than test particular hypotheses. Given that this study's analyses failed to corroborate hypotheses about the MEZURE's standard battery subtests, a confirmatory analysis was subsequently conducted to aid in understanding the instrument's structure and to test competing models.

Confirmatory (Restricted) Factor Analysis

CFA fit statistics results are presented in Table 5. The bi-factor model with two group factors and the higher-order model with two group factors did not converge. Of the models that converged, the one with a single factor fit the worst across all indices. The other three models fit the data similarly, although the bi-factor model where Categorization was specified to load only

[6] It is noted that EFA is rarely truly exploratory while CFA is rarely completely confirmatory; hence, the terms unrestricted and restricted may be a more accurate description of the techniques.

on the general factor fit slightly better than the other two. The loadings for this model are presented in Table 6.

In contrast to the unconstrained factor analyses presented in Table 4, the general factor in Table 6 is actually a general factor, but it is weak with Visual Analogies having the only loading much above .50. The two group factors are consistent with the unconstrained factor analysis. The PR/Gwm factor is defined by a mixture of memory, reasoning, and visual analysis tasks, while the so-called crystalized ability (verbal) factor is equally defined by Information and Vocabulary. While Categorization does load onto the general factor, the loading is relatively weak. Please see Figure 1 for a visual depiction of the Table 6 results.

Metrics of Scale Interpretability

Various metrics of scale interpretability (e.g., Omega-hierarchical, omega-hierarchical subscale, *H*, PUC, FDI) are furnished for all the models that were estimated. From Tables 4 and 6, omega estimates for the general, PR/Gwm and Verbal Ability (Gc) factors were as follows: general (.43 and .41), PR/Gwm (.40 and .47), and Verbal Ability (.13 and .14), respectively. This suggests that the general factor and PR/Gwm factors are below the level (.50 to .75) suggested for confidant clinical interpretation of those measures (Reise et al., 2013). By contrast, the ω_{hs} coefficients for the MEZURE Gc group factor was considerably lower and also lacked sufficient unit-weighted variance for confident clinical interpretation (Reise, 2012; Reise et al., 2013).

Discussion

The MEZURE is a fully online (in-person or remote) test of cognitive ability for children and adults. It was the first test of cognitive ability on the marketplace to offer this type of administration modality and has been in existence for over two decades. As a result of the

COVID-19 pandemic, interest in the use and development of such measures has increased considerably (Farmer et al., 2021).

To address the analytical limitations of the Clinical Manual, this study provides an enhanced understanding of the theoretical/factor structure of the MEZURE using recommended exploratory and confirmatory factor analysis methods. Paradoxically, the EFA results from the Schmid-Leiman procedure and the best fitting CFA model (*g* plus two group factors with Categorization on *g* only) produced a weak general factor saliently loaded only by the PR/Working Memory subtests, and two group factors (PR/Working Memory and Verbal Ability). The Categorization subtest had nothing in common with any other subtests on the MEZURE which resulted in both small general and group factor loadings overall. Equally puzzling, the Verbal Ability subtests had small loadings on the general factor, a finding that is inconsistent with both theory and the history of IQ test structural validity results going back to Spearman (1927). History suggests that the verbally loaded subtests of information and vocabulary tend have some of the highest *g* loadings on contemporary measures of cognitive ability (Sattler, 2018). Further, recent structural validity research supports a dominant general factor whose variance allotment dwarfs that of lower order factors by a magnitude of five to 15, a finding that is well replicated within the psychometric literature (Dombrowski, McGill & Morgan, 2021). Across the two group factor solutions (e.g., Tables 4 and 6), the variance assigned to the general factor was considerably lower than what would have been expected from the empirical literature. From a psychometric perspective, it is clear that the verbal subtests (e.g., Categorization, Information, Vocabulary) have little in common with the PR/Working Memory subtests contributing to a lower general factor loading. Categorization is particularly problematic as it has little in common with any subtest on the MEZURE. Likely, Categorization should have been eliminated from the MEZURE and should be a target for elimination in future editions of the test. What is less straightforward is the rationale for such weak loadings of the verbal ability subtests on the general factor with the MEZURE. Although it can only be speculated, one intriguing hypothesis relates to the online modality of administration. Perhaps the MEZURE does not provide optimal measurement of verbal abilities when presented in an online format, and does not capture the full flavor of such abilities in the same way that traditional tests of cognitive ability do. As recent equivalence studies between online and in-person administrations of other cognitive measures provide preliminary evidence that there is likely no significant difference between administration modality for verbal tests (e.g., Gilbert et al., 2021; Wright; 2020), it raises the question of whether this finding is an artifact of the measure in question. Further, subsequent large-scale efforts to examine and validate similar hypothesized constructs have been successful (i.e., International Cognitive Ability Resource [ICAR], Revelle et al., 2020). Finally, it appears that the Clinical Manual may have potentially mislabeled the Fluid Reasoning factor as this dimension is complexly determined containing subtest content attributable to visual-spatial processing, fluid reasoning *and* working memory. For that reason, this factor was labelled a combined Perceptual Reasoning (PR) and Working Memory (Gwm) factor. Conversely, the labeling of the Verbal Ability factor as Gc appears theoretically appropriate. Nevertheless, the present results illustrate well that factor analytic methodologies that presume the presence of a general factor are not biased (e.g., Dombrowski et al. 2021) given the relatively weak general factor loadings observed.

What does this mean for the interpretation of the MEZURE? When two group factors are extracted, the instrument may be viewed as having two group factors and a general factor. However, variance partitioning and omega estimates generally suggest a weak general factor (relative to historical standards), a weak verbal ability factor, and a modestly strong PR/Working Memory factor. Metrics of interpretability suggest caution with interpretation of the verbal ability factor and even the general factor in the final adopted model. Although the best fitting model contains two group factors and a general factor, the low communalities of Information, Vocabulary and Categorization suggest that the MEZURE is, in some respects, an instrument dominated by the PR/Working Memory subtests. This is not surprising given the inherent cognitive complexity of these subtests (McGrew et al., 2014). This finding is further supported by a review of the single group bifactor (1 *g* plus 1 group) model (Online Supplement Table A4), which provides evidence for the presence of a general factor dominated by the PR/Gwm subtests along with a group factor comprising the Gc subtests (e.g., Information and Vocabulary). Although CFA fit statistics (Table 5) suggest that the two group bifactor solution with Categorization loading only on *g* (consistent with the SL model; Table 4) is superior to the one group bifactor solution just described, an inspection of local fit may well suggest that the single group bifactor model (Online Supplement Table A4) is also plausible and perhaps even more parsimonious. In totality, users of the MEZURE are to exercise caution when making clinical decisions as the MEZURE struggles to measure well any of its purported factors at the group or general level and since the resulting structure is different then what is proposed by the publisher.

Limitations

This study is limited by the small number of variables (i.e., subtests) available to model. Ideally, all the cognitive ability instruments available in the marketplace would contain a greater number of subtests permitting more accurate modeling of linkage with theory but this must be weighed against practical constraints (e.g., administration time). This study is also limited by the use of the same samples for both EFA and CFA. Ideally, the sample would have been cross-validated to avoid capitalization on chance. Additionally, access to the actual normative sample would have permitted the investigation of further characteristics of the sample including the distribution of performance on the measures and potential departures from the assumption of normally distributed traits given the measurement modality employed. This would have allowed for a determination of whether maximum likelihood or some other estimation method for CFA should have been used. Accordingly, replication of these results in clinical samples to determine whether these results are unique to the normative sample data would greatly benefit users of the test.

Conclusion

This study provides additional structural validity insight into the MEZURE, an online computer administered test of cognitive ability that is conducted entirely over the Internet. The MEZURE was developed decades before the COVID-19 pandemic and may be considered ahead of its time. However, the MEZURE's Clinical Manual provides only surface validity information so the field's understanding of the instrument is incomplete. In totality, since the MEZURE is an instrument dominated by perceptual reasoning and working memory its composite IQ score should be interpreted as primarily a PR/Gwm composite. In some respects, the PR/Gwm (i.e., Gf) group factor represents a redundant specific factor that also measures the same construct as the general factor. For that reason, perhaps the structure presented in Online Supplement A4 (g plus Gc) is also tenable although it does not contain the best global fit (see Table 5). Conversely, the MEZURE does a poor job of measuring verbal ability as reflected by the low general factor loadings of the verbal ability subtests. Whether this finding is an artifact of the administration modality, the measure in question, or a relevant departure from established test development theory remains unanswered but deserves increased research attention, as online/remote cognitive ability assessment will likely remain a fixture in our assessment ecosystem given the potential paradigm shift toward teleassessment following the COVID-19 pandemic.

References

Assessment Technologies, Inc (1995-2020). *MEZURE* [Computer program]. Author.

Assessment Technologies, Inc (2020a). *MEZURE clinical manual*. Author.

Bartlett, M. S. (1954). A further note on the multiplying factors for various X^2 approximations in factor analysis. *Journal of the Royal Statistical Society, 16,* 296-298. **https://doi.org/10.1111/j.2517-6161.1954.tb00174.x**

Beaujean, A. A., & Benson, N. F. (2019). The one and the many: Enduring legacies of Spearman and Thurstone on intelligence test score interpretation. *Applied Measurement in Education, 32*(3), 198-215. https://doi.org/10.1080/08957347.2019.1619560

Bennett, M. R., & Hacker, P. M. S. (2021). *Philosophical foundations of neuroscience* (2nd ed.). Blackwell.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics.* Cambridge University Press. https://doi.org/10.1017/CBO9780511490026

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2)*,* 245-276. https://doi.org/10.1207/s15327906mbr0102_10

Child, D. (2006). *The essentials of factor analysis* (2nd ed.). Continuum.

Cudeck, R. (2000). Exploratory factor analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of multivariate statistics and mathematical modeling* (pp. 265–296). Academic Press.

Dombrowski, S. C. (2020). A newly proposed framework and a clarion call to improve practice. In S. C. Dombrowski, *Psychoeducational assessment and report writing* (2nd. ed., pp. 9-59). Springer Nature. https://doi.org/10.1007/978-3-030-44641-3

Dombrowski, S. C., Beaujean, A. A., Schneider, J. W. & McGill, R. J. & Benson, N. (2019). Using exploratory bifactor analysis to understand the latent structure of multidimensional psychological measures: An applied example featuring the WISC-V. *Structural Equation Modeling: A Multidisciplinary Journal, 26*(6), 847-860. https://doi.org/10.1080/10705511.2019.1622421

Dombrowski, S. C., Engel, S., & Lennon, J. (2022). Test Review: MEZURE. *Journal of Psychoeducational Assessment, 40*(4), 559–565. https://doi.org/10.1177/07342829211072399

Dombrowski, S. C., McGill, R. J., Canivez, G. L., Watkins, M. W., & Beaujean, A. A. (2021). Factor analysis and variance partitioning in intelligence test research: Clarifying misconceptions. *Journal of Psychoeducational Assessment, 39*(1), 28-38. https://doi.org/10.1177/0734282920961952

Dombrowski, S. C., J. McGill, R., Farmer, R. L., Kranzler, J. H., & Canivez, G. L. (2022). Beyond the rhetoric of evidence-based assessment: A framework for critical thinking in clinical practice. *School Psychology Review, 51*(6), 771-784. https://doi.org/10.1080/2372966X.2021.1960126

Dombrowski, S. C., McGill, R. J. & Morgan, G. W. (2021). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation and theory. *Assessment, 38*(3), 977-993. https://doi.org/10.1177/1073191119869828

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford University Press.

Farmer, R. L., McGill, R. J., Dombrowski, S. C., Benson, N. F., Smith-Kellen, S., Lockwood, A. B., Powell, S., Pynn, C., & Stinnett, T. A. (2021). Conducting psychoeducational assessments during the covid-19 crisis: The danger of good intentions. *Contemporary*

School Psychology, 25, 27-32. https://doi.org/10.1007/s40688-020-00293-x

Farmer, R. L., McGill, R. J., Dombrowski, S. C., McClain, M. B., Harris, B., Lockwood, A. B., Powell, S. L., Pynn, C., Smith-Kellen, S., Loethen, E., Benson, N. F., & Stinnett, T. A. (2020). Teleassessment with children and adolescents during the Coronavirus (COVID-19) pandemic and beyond: Practice and policy implications. *Professional Psychology: Research and Practice, 51*(5), 477–487. https://doi.org/10.1037/pro0000349

Farmer, R. L., McGill, R. J., Dombrowski, S. C., Benson, N. F., Smith-Kellen, S., Lockwood, A. B., . . . Stinnett, T. A. (2021). Conducting psychoeducational assessments during the COVID-19 crisis: The danger of good intentions. *Contemporary School Psychology*, *25*(1), 27-32. https://doi.org/10.1007/s40688-020-00293-x

Gilbert, K., Kranzler, J. H., & Benson, N. F. (2021). An independent examination of the equivalence of the standard and digital administration formats of the Wechsler Intelligence Scale for Children-5th Edition. *Journal of School Psychology,* 85*,* 113-124. https://doi.org/10.1016/j.jsp.2021.01.002

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

Guttman, L. (1977). What is not what in statistics. *Journal of the Royal Statistical Society. Series D (The Statistician), 26*, 81-107. https://doi.org/10.2307/2987957

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification*. Edward Arnold

Holzinger, K. J., Swineford, F., & Harman, H. H. (1937). *Student manual of factor analysis: An elementary exposition of the bi-factor method and its relation to multiple-factor methods*. University of Chicago Department of Education.

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

Horn, J. L. (1965a). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179-185. https//doi.org/ 10.1007/BF02289447

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligence. *Journal of Educational Psychology, 57*(5)*,* 253-270. https://doi.org/10.1037/h0023816

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

Krause, M. S. (2012). Measurement validity is fundamentally a matter of definition, not correlation. *Review of General Psychology, 16*, 391-400. https://doi.org/10.1037/a0027701

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika,* 39, 31-36. https://doi.org/10.1007/BF02291575

Keith, T. Z., & Kranzler, J. H. (1999). The absence of structural fidelity precludes construct validity: Rejoinder to Naglieri on what the cognitive assessment system does and does not measure. *School Psychology Review, 28*(2), 303-321. https://doi.org/10.1080/02796015.1999.12085967

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory Bi-factor Analysis. *Psychometrika, 76*(4), 537-549. https://doi.org/10.1007/s11336-011-9218-4

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

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). Emerald Group Publishing.

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

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241. https://doi.org/10.1146/annurev.psych.50.1.215

McGill, R. J., Dombrowski, S. C., & Canivez, G. L. (2018). Cognitive profile analysis in school psychology: History, issues, and continued concerns. *Journal of school psychology*, *71*, 108-121.

McGrew, K. S., LaForte, E. M., & Shrank, F. A. (2014). *Woodcock-Johnson IV* [Clinical Manual]. Riverside Publishing.

Michell, J. (1990). *An introduction to the logic of psychological measurement*. Erlbaum.

Muthén, L. K., & Muthén, B. O. (1998 –2022). *Mplus user's guide* (7th ed.) Author.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. The National Academies Press. https://doi.org/doi:10.17226/25303

Newton, P. E. (2017). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice, 36*, 5-15. https://doi.org/10.1111/emip.12146

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2*(1), 13-43. https://doi.org/10.1207/S15328031US0201_02

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

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

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140. https://doi.org/10.1080/00223891.2012.725437

Revelle, W., Dworak, E. M., & Condon, D. (2020). Cognitive ability in everyday life: The utility of open-source measures. *Current Directions in Psychological Science, 29*(4), 358-363. https://doi.org/10.1177/0963721420922178

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137-150. https://doi.org/10.1037/met0000045

Sattler, J. M. (2018). *Assessment of children and adolescents: Cognitive functions and applications* (6th ed.). Sattler Publishing.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53-61. https://doi.org/ 10.1007/BF02289209

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321. https://doi.org/10.1177/0734282911406653

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., pp. 73--163). Guilford.

Spearman, C. (1927). *The abilities of man*. Macmillan.

Spearman, C. E. (1931). Our need of some science in place of the word `intelligence.'. *Journal of Educational Psychology*, *22*(6), 401-410. https://doi.org/10.1037/h0070599

Tataryn, D. J., Wood, J. M., & Gorsuch, R. L. (1999). Setting the value of *k* in promax: A Monte Carlo study. *Educational and Psychological Measurement, 59*(3), 384-391. https://doi.org/10.1177/00131649921969938

Tukey, J. W. (1977). *Exploratory data analysis*. Addison Wesley.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician*, *34*(1), 23-25. https://doi.org/10.1080/00031305.1980.10482706

Velicer, W. F. (1976). Determining the number of components form the matrix of partial correlations. *Psychometrika, 31,* 321-327. doi: 10.1007/BF02293557

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177--203). Erlbaum.

Watkins, M. W. (2013). *Omega* [Computer software]. Ed & Psych Associates.

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology, 44*(3), 219-246. https://doi.org/10.1177/0095798418771807

Wright, A. J. (2020). Equivalence of remote, digital administration and traditional, in-person administration of the Wechsler Intelligence Scale for Children, Fifth Edition (WISC-V). *Psychological Assessment, 32*(9), 809-817. https://doi.org/10.1037/pas0000939

Table 1

Subtests of MEZURE Standard Battery

Score	Attributes Captured
Analogies	Fluid Reasoning Visual Processing
Auditory Memory	Fluid Reasoning Short-term Memory
Categorization	Crystallized Reasoning
Information	Crystallized Reasoning
Vocabulary	Crystallized Reasoning
Visual Memory	Fluid Reasoning Short-term Memory
Visual Closure	Fluid Reasoning Visual Processing

Table 2

Composite Scores produced from MEZURE Standard Battery

Score	Subtests	Attribute Represented
Fluid IQ	Analogies Auditory Memory Visual Closure Visual Memory	Fluid Reasoning
Crystallized IQ	Categorization Information Vocabulary	Crystallized Reasoning
Composite IQ	Categorization Information Vocabulary Analogies Auditory Memory Visual Closure Visual Memory	Summative index of general intellectual Functioning.

Table 3

Communality Estimates Following Initial Two and Three Factor Principal Axis Extraction

Three Factor Extraction

Start	**VC**	**AN**	**IN**	**CA**	**VM**	**VO**	**AM**	**Fit**	**Objective**
SMC	NC								
0.1	NC								
0.2	.227	.483	.226	.199	.491	.234	.503	.669	.008
0.3	.228	.482	.230	.300	.488	.235	.500	.685	.008
0.4	.227	.481	.226	.400	.490	.234	.504	.698	.008
0.5	.227	.481	.227	.499	.490	.233	.504	.709	.008
0.6	.227	.481	.228	.599	.490	.232	.504	.718	.008
0.7	.227	.481	.231	.699	.490	.230	.504	.726	.008
0.8	.227	.481	.232	.798	.490	.228	.504	.731	.008
0.9	.227	.481	.236	.898	.490	.225	.504	.734	.009
1	.227	.481	.241	.998	.490	.221	.504	.735	.009
Range	.001	.002	.015	.799	.002	.014	.001	.066	.001

Two Factor Extraction

Start	**VC**	**AN**	**IN**	**CA**	**VM**	**VO**	**AM**	**Fit**	**Objective**
SMC	.227	.484	.226	.084	.488	.231	.503	.648	.009
0.1	.227	.484	.227	.084	.488	.231	.503	.648	.009
0.2	.227	.484	.227	.084	.488	.230	.502	.648	.009
0.3	.228	.485	.229	.084	.487	.230	.501	.648	.009
0.4	.227	.484	.230	.084	.488	.229	.503	.648	.009
0.5	.227	.484	.231	.084	.488	.228	.503	.648	.009
0.6	.227	.484	.232	.084	.488	.227	.503	.648	.009
0.7	.227	.484	.234	.084	.488	.225	.503	.648	.009
0.8	.227	.484	.236	.084	.488	.223	.503	.648	.009
0.9	.227	.484	.238	.084	.488	.222	.503	.648	.009
1	.227	.484	.241	.084	.488	.219	.503	.648	.009
Range	.001	.001	.011	.000	.001	.009	.002	.648	.009

Note: SMC=Squared Multiple Correlation; NC=Nonconvergence; VC=Visual Closure AN=Visual Analogies; IN=Information; CA=Categorization; VM=Visual Memory; VO=Vocabulary; AM=Auditory Memory

Table 4

Schmid-Leiman Transformation of Higher-Order Factor Solution of MEZURE Standard Battery Subtests

	General		PR/Gwm		Gc			
	b	S^2	b	S^2	b	S^2	h^2	u^2
Auditory Memory	.46	.21	**.53**	.28	-.05	.00	.50	.50
Visual Memory	.47	.22	**.52**	.27	-.03	.00	.49	.51
Visual Analogies	.51	.26	**.47**	.22	.06	.00	.48	.52
Visual Closure	.32	.10	**.35**	.12	-.02	.00	.23	.77
Information	.31	.10	-.04	.00	**.36**	**.13**	.23	.77
Vocabulary	.32	.10	-.03	.00	**.36**	**.13**	.23	.77
Categorization	.23	.05	.10	.01	.14	.02	.08	.92
Total Variance		.15		.13		.04	.32	.680
Explained Common Variance		.47		.40		.13	1.00	
ω_{h}/ω_{hs}		.43		.39		.19		
H		.56		.54		.24		
FDI		.75		.74		.49		
PUC		.57						

Note. *b* = standardized loading of subtest on factor, S^2 = variance explained, h^2 = communality, u^2 = uniqueness, ω_H = Omega-hierarchical (general factor), ω_{HS} = Omega-hierarchical subscale (group factors), H = construct reliability or replicability index, FDI = factor determinancy index, PUC = percentage of uncontaminated correlations. PR = Perceptual Reasoning, Gwm = Working Memory. Gf=Fluid Reasoning, Gv=Visual-Spatial, Gc=Crystallized Ability.

Table 5

CFA Model Fit Statistics for the MEZURE Standard Battery

Model	χ^2	*df*	CFI	TLI	SRMR	RMSEA	RMSEA 90% CI	BIC	AIC
General Factor only	276.45	14	.937	.906	.039	.067	(0.060-0.074)	143704	143570
Bi-factor (general + 2 group factors)	Does not converge								
Higher-Order (1 second-order factor + 2 first-order factors)	Does not converge								
Oblique (2 group factors)	105.7	13	.978	.964	.024	.041	(0.034-0.049)	143402	143541
Bi-factor (general factor + 1 group factor)	62.54	11	.988	.977	.014	.033	(0.026-0.042)	143515	143363
Bi-factor (g + 2 factors with CA on general only)	35.14	9	.994	.985	.010	.026	(0.018-0.036)	143504	143339

Note: X^2= Chi-Square; *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean square; RMSEA = root mean square error of approximation; BIC = Bayesian Information Criterion; AIC = Akaike's Information Criterion*;* *g* = general intelligence; CFA = confirmatory factor analysis; CA = Categorization.

Table 6

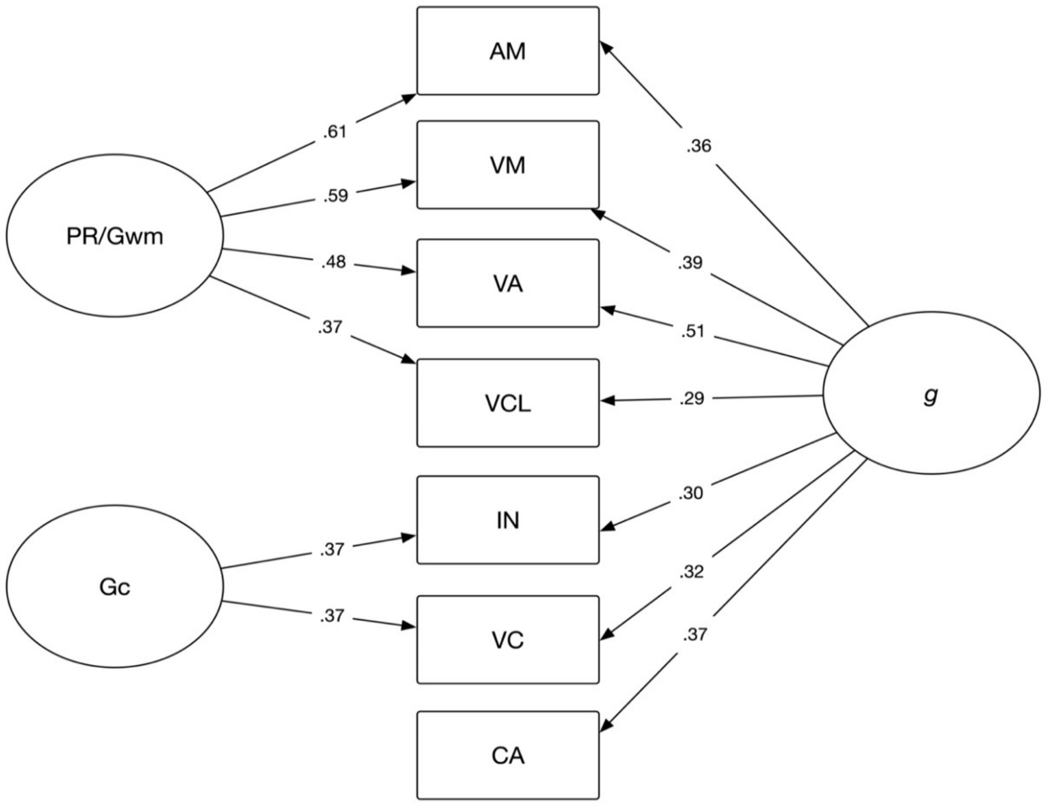
MEZURE Sources of Variance: Confirmatory Bifactor Analysis (g + 2 Factors with Categorization on g only)

	General		PR/Gwm		Gc			
	b	S^2	b	S^2	b	S^2	h^2	u^2
Auditory Memory	.36	.13	**.61**	.38	-.05	.00	.51	.49
Visual Memory	.39	.16	**.59**	.35	-.03	.00	.50	.50
Visual Analogies	.51	.26	**.48**	.23	.06	.00	.49	.51
Visual Closure	.29	.09	**.37**	.14	-.02	.00	.23	.78
Information	.30	.09			**.37**	**.13**	.22	.78
Vocabulary	.32	.10			**.37**	**.13**	.24	.76
Categorization	.37	.14					.14	.86
Explained Total Variance		.14		.16		.04	.33	.67
Explained Common Variance		.42		.47		.12	1.00	
ω_{h}/ω_{hs}		.41		.47		.14		
H		.53		.61		.24		
FDI		.73		.78		.49		
PUC		.57						

Note. *b* = standardized loading of subtest on factor, S^2 = variance explained, h^2 = communality, u^2 = uniqueness, ω_H = Omega-hierarchical (general factor), ω_{HS} = Omega-hierarchical subscale (group factors), H = construct reliability or replicability index, FDI = factor determinancy index, PUC = percentage of uncontaminated correlations. PR = Perceptual Reasoning (combination of Gf/Gv), Gwm = Working Memory, Gf=Fluid Reasoning, Gv=Visual-Spatial. Gc=Crystallized Ability.

Figure 1

Direct hierarchical (bifactor), with standardized coefficients, for the MEAZURE



Note. AM = Auditory Memory, VM = Visual Memory, VA = Visual Analogies, VCL = Visual Closure, IN = Information, VC = Vocabulary, CA = Categorization, *g* = general intelligence, PR/GWM = Perceptual Reasoning/Working Memory, Gc = Crystallized Ability.