Journal of Psychoeducational Assessment
2022, Vol. 0(0) 1–7
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/07342829211072399
journals.sagepub.com/home/jpa
SAGE

Test Review: MEZURE

Stefan C. Dombrowski[1], Shiri Engel[1], and James Lennon[2]

Abstract

This article reviews the administrative and psychometric properties of the MEZURE, an online-only (remote or in-person) measure of cognitive ability used to evaluate school-aged populations and adults.

Keywords

Tests, intelligence/cognition, MEZURE, test review, online assessment, remote assessment

As a consequence of the COVID-19 pandemic, test publishing companies are pivoting and adapting their existing cognitive assessment technology to accommodate remote and virtual administration. The equivalence in examinee remote performance on many of these instruments relative to traditional in-person administration remains largely unknown (Farmer et al., 2020). Users seeking a cognitive ability assessment with remote or online assessment capabilities should consider the MEZURE (Assessment Technologies, Inc, 1995; https://www.mezure.com/). The MEZURE is an online, computer-based standardized measure of cognitive ability designed for individuals ages six through adult that appears to have been in existence for approximately 20 years. It contains a standard battery of seven subtests, a screening battery of four subtests, and offers a separate, stand-alone supplemental battery containing five subtests. The current cost of the MEZURE is US $1250 per 25 test credit (i.e., US $50 per administration as of the writing of this review). Credits may be accessed by multiple people within a purchasing district or agency (level C psychologist qualification is required). The MEZURE may be administered either in-person or remotely and requires a computer (desktop or laptop) and examinee facility with a mouse. Although the Clinical Manual contains a 2020 copyright, it is unclear when the actual instrument was developed. One may infer from a review of the Clinical Manual (Assessment Technologies, Inc, 2020a; e.g., the reference list whose citations are older than 2001; the validity studies using the WISC-III; and the biographies of the lead consultants on the development of the MEZURE) that the instrument was published prior to 2001. It is also unknown whether and when the instrument was updated in any way. According to the Clinical Manual, Assessment Technologies,

[1]Department of Graduate Education, Leadership & Counseling, Rider University, Lawrenceville, NJ, USA
[2]Graduate Psychology Program, New Jersey City University, Jersey City, NJ, USA

Corresponding Author:
Stefan C. Dombrowski, Department of Graduate Education, Leadership & Counseling, Rider University, 2083 Lawrenceville Road, Lawrenceville, NJ 08684, USA.
Email: sdombrowski@rider.edu

Inc., Inc is the present copyright holder and advertises the MEZURE for use with school-aged populations (https://www.mezureschools.com/), police, corporate, and military personnel (https://www.mezuremilitary.com/). It is unknown whether the instrument was standardized on law enforcement, corporate, or military personnel. Further, the MEZURE can be found under a variety of additional names including ATI (Assessment Technologies), COMIT (Computer-Optimized Multimedia-Integrated Test), and SAAVI Recruiter (the law enforcement form of ATI). Nonetheless, each of these references source back to the instrument presented and discussed in the Clinical Manual (Assessment Technologies, Inc, 2020a) and User Manual (Assessment Technologies, Inc., 2020b).

The Clinical Manual reports that the MEZURE was developed to align with Cattel–Horn's Gf–Gc theory of cognitive abilities (Cattell, 1963; Horn, 1965) as well as the Cattell–Horn–Carroll model of cognitive abilities (McGrew, 2005). The Screening Battery contains two Fluid Reasoning (Gf) subtests (Visual Closure and Analogies) and two Crystallized Ability (Gc) subtests (Information and Categorization) and requires approximately 15–20 min for administration. The results of the screening battery provide a brief measure of cognitive functioning. The Standard Battery is composed of seven subtests that produce a composite IQ score and two index scores (Fluid Reasoning and Crystalized Ability): four Fluid Reasoning subtests (Visual Closure, Visual Analogies, Visual Memory, and Auditory Memory) and three Crystallized Ability subtests (Categorization, Information, and Vocabulary), requiring approximately 25–30 min in total for administration. The total score from the seven subtests is aggregated to provide a composite IQ score. Composite and index scores are scaled such that the mean is equal to 100 and the *SD* is 15. Subtest scores are scaled such that the mean is 10 and the *SD* is 3. Five stand-alone supplemental subtests are offered: Processing Speed, Social Apperception, Auditory Memory with Visual Distractions, Auditory Memory with Auditory Distractions, and Visual Memory with Auditory Distractions. These subtests do not contribute to the calculation of the main battery's composite or index scores although the two supplemental memory subtests contribute to a Distraction Resistance Index score

Administration and Scoring

The full battery of tests requires anywhere from 45–60 minutes and can be administered in a single sitting, or over multiple sessions. The standard battery requires approximately 25–30 minutes for administration while the supplemental battery requires an additional 15–20 minutes. The entire test is administered and scored online eliminating the need for a physical manual, scoring sheets, or manipulatives. The Clinical Manual indicates that the MEZURE was adapted into Spanish and Russian. However, only US-based English language norms and psychometric data are presented in the Clinical Manual. The instrument should not be administered to visually impaired, color blind, or hearing-impaired individuals due to the visual and auditory nature of the tasks.

The MEZURE is Federal Rights to Privacy Act (FERPA) and Health Insurance Portability and Accountability Act (HIPAA) compliant and claims to offer military-level encryption to keep it secure. Although the Clinical Manual indicates that a computer (desktop or laptop) with working internet connection is required, a website describing the MEZURE indicates that a tablet or iPad may be used (https://www.mezureschools.com/product-info).

When administered remotely, additional requirements are that both the examiner and examinee have a separate computer (with a mouse), access to the internet, and a working webcam for interactions and observations to take place. No further equipment, manual, scoring books, or manipulatives are necessary for administration as everything from administration to scoring is conducted virtually. The examiner's role in administrating the MEZURE is minimal. All necessary instructions for each subtest are delivered directly by the testing program.

The MEZURE provides two training items prior to the start of the test that allow the examinee to practice mouse skills and become familiar with the question–answer format. After screening, entry points as well as basal and ceiling points are all automatically calculated using Dynamic Adaptive Routing Technology (DART) to make determinations.

Upon completion, the program will calculate scores, provide a subtest scatter screen, and produce a visual representation of scores in the form of graphs. The composite IQ score will be available if the entire brief or standard battery have been completed. For the standard battery, a Fluid and Crystalized cluster score will be available. Examiners have the option of printing out scores and graphs.

Technical Adequacy

Norming

Prior to engaging in the standardization of the MEZURE, the test publisher conducted two pilot studies for the purpose of refining the instrument and determining which items to retain using classical item analyses and expert review. Following the pilot studies, the MEZURE was administered to 4184 participants matched to the US Census (year not provided) in terms of parental education level, geographic region, ethnicity, and sex. Table 5.1 in the Clinical Manual indicates that participants 6 through 17, stratified across each year of age, were administered the MEZURE followed by administration to an adult group. The Clinical Manual also reports that the final item pool was not determined until all the language adaptations were completed to include only those items that were culture-fair and adaptable into Spanish and Russian. Although the MEZURE is advertised as a tool to determine diagnostic categories (e.g., gifted; intellectual disability; and specific learning disability) the sample size of the clinical/diagnostic group is insufficient for analyses beyond descriptive statistics for all clinical groups except specific learning disability.

Item Development

Items on the MEZURE were analyzed using classical and item response theory methods to determine item start points and sequencing for each age group. The item discrimination index was also calculated to determine whether items functioned differently across comparison groups with respect ethnicity (e.g., African American/Caucasian, Asian/Caucasian, Hispanic/Caucasian, and Another Ethnicity), geographic region (e.g., Northeast, North Central, South, and West), and residence location (Urban, Rural, and Suburban). Results of all analyses demonstrated no evidence of bias. The MEZURE was adapted to provide Spanish and Russian language options. Thorough checks were carried out with multiple interpreters representing regional dialects and cultures to ensure items with social, cultural, and linguistic differences were eliminated.

Reliability

Test–retest. Test–retest reliability for the MEZURE was calculated by administering the test to 40 participants at two-time intervals ranging from 3 weeks to 3 months. The reliability coefficients for the Vocabulary and Visual Memory subtests were high (.92 and .90, respectively). Acceptable coefficient levels were established for the Auditory Memory, Information, and Visual Analogies subtests at .88, .85, and .82, respectively. Categorization and Visual Closure subtests were below the acceptable level at .70 and .64, respectively. The test–retest reliability coefficients for the Social Apperception and Processing Speed subtests were acceptable at .85 and .81, respectively.

Inter-scorer. Inter-scorer error for the MEZURE was eliminated as the test is administered using a computer. Consequently, the interscorer reliability was not calculated.

Internal Consistency. Cronbach's alpha and split-half correlations were calculated across four of the seven cognitive subtests: Visual Analogies, Information, Vocabulary, and Categorization. The Clinical Manual indicated that Internal Consistency was not calculated for Visual Closure, Visual Memory, and Auditory Memory noting in a footnote that those subtests did not lend themselves to calculation of coefficient alpha. Internal Consistency was not calculated for Social Apperception and Processing Speed due to the speeded nature of tasks within these subtests. Table 6.1 in the Clinical Manual indicated that the alpha and split-half scores for Visual Analogies and Information subtests were consistently high (>.90) across the entire age range. Scores were in the acceptable range for the Vocabulary and Categorization subtests although the internal consistency level for Vocabulary below age 12 trended lower (<.80) but was still in the acceptable range. The standard error of measurement (SEM) was calculated for the same four of seven MEZURE standard battery subtests for which internal consistency was calculated, and the two supplemental subtests (Social Apperception and Processing Speed) for which test–retest coefficients were calculated. The SEM was not reported for Visual Closure, Visual Memory, and Auditory Memory.

Validity

The test publisher utilized three approaches to establish the validity of the MEZURE: content, criterion, and construct validity.

Content Validity. The Clinical Manual reports that content validity of the MEZURE was managed through modification and elimination of test items following each pilot study. The Clinical Manual reports that teams of psychologists and educators reviewed items and the resulting item analyses to ensure appropriateness of test content.

Criterion Validity. To establish criterion validity correlations between the MEZURE, subtest scores and achievement test scores were derived with two studies using the Iowa Test of Basic Skills (ITBS). Table 5 in the Clinical Manual demonstrates that the relationship between the MEZURE, and the ITBS demonstrates that the relationship was moderate to strong, ranging from .54 to .74 and generally consistent with the correlations of other major tests of achievement and cognitive ability.

Construct Validity. Table 6.4 in the Clinical Manual presents the correlations between the MEZURE standard battery and the WISC-III. The correlations between the MEZURE total score and the WISC-III FSIQ, Verbal IQ, and Performance IQ were .79, .70, and .72, respectively, indicating strong evidence of construct validity. The MEZURE's Processing Speed subtest was also correlated with WISC-III's Processing Speed Index producing a moderate correlation (r = .63). The MEZURE did not evaluate the relationship between its Fluid Reasoning and Crystalized Ability indices and the WISC-III Full Scale, Verbal, and Performance IQs.

Several additional analyses were carried out to establish construct validity of the standard battery of the MEZURE. First, the scores among various groups (i.e., gifted [n = 15], learning disabilities [LD; n = 44], and mental retardation [*sic*][n = 9]) were reviewed by inspecting the mean scores of each group. The Clinical Manual reports that the pattern of scores was consistent with expectations where individuals classified as gifted scored higher than the mean, while individuals with LD and intellectual disability scored slightly and significantly lower, respectively.

Only the LD clinical group contained a sufficient sample size for statistical analyses (e.g., matched samples t-test) beyond simple descriptive statistics yet these analyses were not undertaken.

To establish internal or structural validity, in addition to conducting a visual inspection of the correlations among the subtests and the MEZURE's total score, the Clinical Manual reports the results of a factor analysis of the seven-subtest total MEZURE battery. The Clinical Manual reported that an oblique rotation was used to determine the alignment of subtests with respective factors. Evidence from the analysis reported in the Clinical Manual suggests that the MEZURE measures a composite IQ and two distinct index scores (Gf and Gc).

Commentary and Recommendations

The MEZURE may be a viable option for the administration of an online only, remotely administered or in-person test of cognitive ability. The MEZURE has been on the marketplace for at least two decades. In an era of COVID-19, where test publishers are scrambling to offer remote options, the MEZURE may be considered ahead of its time. The instrument is easy to administer so long as the examinee has access to a computer and facility with a mouse. The scoring is completed automatically which eliminates scoring error. The MEZURE's approach to standardization (i.e., norming, reliability, and validity) appears to meet many of the Standards for Educational and Psychological Testing (American Education Research Association, 2014). The test publisher engaged in a series of two pilot studies prior to actual standardization to refine and further develop the instrument. This is a laudable and noteworthy exploratory practice that helps the instrument avoid costly errors of commission and omission. The User Manual (Assessment Technologies, Inc., 2020b) is thorough, well-written, and very easy to follow. The MEZURE's items are engaging, well-constructed, and data in the Clinical Manual suggest bias is minimized. The instrument appears linked to theory and offers many advantages over the traditional paper-pencil approach to cognitive ability testing. Because administration is automated, the occasional sources of examiner error are eliminated.

The MEZURE automatically saves information entered by an examinee. However, in the case of technology malfunction (such as WI-FI disconnection), or if an examinee accidentally closes the window in the middle of a subtest, the score may be lost, and the examinee will be required to retake the subtest which may pose problems with practice effects.

From a technical perspective although the instrument appears to be well-constructed, the actual date of development is unknown. The Clinical Manual contains a copyright date of 2020 which may be misleading. One may easily infer from the Clinical Manual that the instrument was developed in the early 2000s. For instance, one study furnished information on the relationship of the MEZURE with the WISC-III (i.e., the WISC-IV was published in 2003). Additionally, the biography of the director of the clinical development team listed a reference to a forthcoming book to be published in 2001 and all references in the Clinical Manual are 2000 or earlier. Finally, the Clinical Manual reports that the high number of standardization participants (a laudable aspect of the MEZURE) was matched on numerous demographics with the U.S. Census. However, the actual year of census utilized for matching was not furnished. This suggests that either the MEZURE, the information in the Clinical Manual, or both need significant updating. It is also unknown whether the standardization occurred remotely or in-person. And various websites indicate that the MEZURE may be administered via tablet and iPad; however, these modalities were not available at the time of standardization and equivalence studies would be required prior to their use especially as it is unlikely that examinees will have a mouse linked to an iPad or tablet.

The Clinical Manual reports that the standard battery's structural validity was assessed via "exploratory factor analysis with oblique rotation" (p10, Table 2.2). However, neither the category of factor extraction (i.e., principal components; principal axis, and maximum likelihood) nor the type of rotation (i.e., promax and oblimin) was disclosed. Additional commonly used factor analytic statistics were also not presented in the Clinical Manual (e.g., eigenvalues, communality values, explained common and total variance). The Clinical Manual reports that performance on various groups (ethnicity, age and classification) was compared by an inspection of means; however, the more appropriate approach would have been to use multigroup confirmatory factor analysis (i.e., invariance analyses) to determine whether the instrument could be compared across groups. The MEZURE was adapted into Spanish and Russian. However, the Clinical Manual does not report whether the instrument was separately normed on Spanish or Russian speaking populations. Further, the presentation of additional reliability analyses beyond test–retest for Visual Closure and the memory subtests would have been useful as would have the subsequent calculation of their SEM. Given this omission interpretation of the memory subtests, the Distraction Resistance Index, and Visual Closure should only be cautiously undertaken, if not avoided entirely. Beyond structural validity evidence the Clinical Manual did not furnish reliability, SEM, and other information for the Gc and Gf indices suggesting that those indices should not be interpreted unless that information can be provided. Finally, the MEZURE's Clinical Manual contains some interpretive guidance that is anachronistic and should be eschewed (see Dombrowski et al., 2021; McGill et al., 2018). This includes reliance on interpretation of subtests, subtest scatter, and the disavowal of the composite IQ in the presence of a large split between the Gf and Gc indices.

Overall, the MEZURE has significant strengths that potentially make it an appealing option for remote cognitive ability assessment. It is at the vanguard of remote IQ testing and offered an alternative to the traditional paper-and-pencil approach long before the marketplace sought out this approach to cognitive ability testing. It also appears well-developed, easy to use and score, offers an online remote option, and provides a degree of structural validity support for its proposed theoretical structure. However, the strengths and unique features of the MEZURE must be considered against the backdrop of several concerns. This includes outdated norms that appear over 20 years old, which may produce scores influenced by the Flynn effect (Flynn, 1984). The Clinical Manual also omitted critically important reliability information (e.g., SEM and confidence intervals) for several subtests (e.g., all memory subtests and Visual Closure) and index scales (e.g., Gc and Gf, and Distraction Resistance Scale) suggesting that subtest and composite scores from those subtests/scales should only be cautiously interpreted, if at all, until this information becomes available. Sample size data for key clinical groups (gifted, ID) is also insufficient. When clinical data was of sufficient size (e.g., LD), analyses beyond simple descriptive statistics (e.g., matched sample t-tests) were not furnished, making it difficult to determine whether the MEZURE is appropriate for use with these groups. With an updating of the norms and an addressing of the concerns presented in this review the MEZURE may well become a viable option for remote or in-person virtual cognitive ability assessment.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Stefan C. Dombrowski https://orcid.org/0000-0002-8057-3751

References

American Educational Research Association (2014). *American psychological association, and national council on measurement in education standards for educational and psychological testing*. American Educational Research Association.

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

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

Assessment Technologies, Inc (2020b). *MEZURE user manual*. Author.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, *54*(1), 1–22. https://doi.org/10.1037/h0046743

Dombrowski, S. C., McGill, R. J., Farmer, R. L., Kranzler, J. H., & Canivez, G. L. (2021). Beyond the rhetoric of evidence-based assessment: A framework for critical thinking in clinical practice. *School Psychology Review*. Advance online publication. https://doi.org/10.1080/2372966X.2021.1960126

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

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*(1), 29–51. https://doi.org/10.1037/0033-2909.95.1.29

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

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. https://doi.org/10.1016/j.jsp.2018.10.007.

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). Guilford Press.