Development of a New Measure of Cognitive Ability Using Automatic Item Generation and Its Psychometric Properties

Ji Hoon Ryoo1,2*, Sunhee Park3*, Hongwook Suh4, Jaehwa Choi5, and Jongkyum Kwon6

Abstract
In the development of cognitive science understanding human intelligence and mind, measurement of cognitive ability has played a key role. To address the development in data scientific point of views related to cognitive neuroscience, there has been a demand of creating a measurement to capture cognition in short and repeated time periods. This paper introduces an innovative measure of cognitive ability based on automatic item generation approach, which can efficiently and effectively measure cognitive ability over time. We also examine its psychometric properties. Content validity of the assessment was considered based on the Cattell-Horn-Carroll theory, and construct validity via convergent and divergent validities was examined by confirmatory factor analysis. The reliability of the measure examined by internal consistencies as well as test-retest reliabilities of each subdomain of cognitive ability were satisfactory. The psychometric properties found clearly support its potential utilities in both educational and clinical settings, especially in a field requiring repeated measures of cognitive ability.

Keywords
automatic item generation, longitudinal study, measure of cognitive ability, reliability, validity

Introduction
Study of the mind in the human brain is one of main research topics in cognitive science (Friedenberg & Silverman, 2006), and the purpose of cognitive-scientific study of the mind is to learn mechanisms that underlie cognitive performance (Sternberg, 1986). In pursuit of understanding the brain mechanisms, a comprehensive and efficient assessment tool measuring cognitive abilities is required (Gur et al., 2010) along with identifying and understanding a structure of human cognition (Wasserman, 2019). On the other hand, it is still a challenge to develop a practical assessment tool even if the domains of cognitive ability are assumed to be identified, due to a lack of consensus between theoretical aspects of intelligence and practical aspects of intelligence (Canivez & Youngstrom, 2019). In spite of such a challenge, many tools and scales have been developed on the strength of psychometric theories (Beaumier, 2015; Bruijnen et al., 2020; Caemmerera et al., 2020; Dombrowski et al., 2021; Geisinger, 2019; McDicken et al., 2019). Based on the Cattell-Horn-Carroll (CHC: Bryan & Mayer, 2020; Flanagan et al., 2013; McGrew, 2009; Schneider & McGrew, 2012, 2018) theory of cognitive abilities, this study uniquely demonstrates how an innovative psychometric method, known as an automatic item generation (AIG; Drasgow et al., 2006; Embretson & Yang, 2006; Gierl & Haladyna, 2013; Gierl et al., 2021; Irvine & Kylonen, 2002), helps to develop a new practical assessment tool of cognitive ability.

Cognitive ability tests are often used as repeated measures in longitudinal study although there is inherent limitation in serial tests, which is called as the “practice effect” (Kaufman, 2010).
1990; Temkin et al., 1999). For example, in numerous clinical settings, serial cognitive tests are administered to investigate the effect of treatments on changes of cognitive abilities (Cerull et al., 2019; Elman et al., 2018; D. M. Jacobs et al., 2017) and to make decisions on disease progress or recovery (Beglinger et al., 2005). Serial cognitive assessments are also essential when the new pharmaceutical treatments for conditions influencing cognition is developed (Beglinger et al., 2005). The similar demand exists in educational settings, particularly, when longitudinal studies are conducted. On the other hand, repeated tests may also result in the practice effect in the educational settings, defined as changes in a person’s test performance on re-testing (Hausknecht et al., 2007), due to examinees’ memory and familiarity of items.

Although individually administered tests in the cognitive assessment tools provide numerous information on examinees, group-based tests have received considerable attention in both clinical and educational settings, especially in diagnostic tests for screening purposes. Clinicians continue to consider the administration of group tests because of short clinical test intervals to screen patient’s current ability (Sparrow & Davis, 2000). In educational settings, individually administered ability tests have some shortcomings for talent identification even though they are very helpful for diagnostic purposes (Lohman & Gambrell, 2012). In other words, the issue of equity is raised because the expensive cost of individualized assessments limits the opportunities of children who can be assessed. Only those who can afford to pay for tests can be provided the opportunities of testing or retesting (Renzulli, 2005). The nonverbal battery of the Cognitive Abilities Test (CogAT; Lohman, 2011) and the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 2008) administered as a group test have been developed for the identification of gifted students in educational settings, but the CogAT nonverbal test has an issue on practice effects (Lohman & Gambrell, 2012). Thus, there has been a demand on developing a group-based measure of nonverbal cognitive ability to reduce the practice effects.

To address the issues of practice effects and group administration in a measure of cognitive ability, we have developed a new measure of cognitive ability using automatic item generation (AIG; Drasgow et al., 2006; Embretson & Yang, 2006; Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002). AIG is a method of producing computerized item models to generate item instances which have similar psychometric properties such as item difficulty. Therefore, the approach is greatly beneficial to a removal of the practice effect. Compared to traditional (manual and individual) item writing process, the AIG-based method has been known as an effective and efficient assessment development process (Wainer, 2002). Furthermore, AIG also provides other possible benefits such as increment of test security by minimizing item exposure, reduction of the unit cost of assessment item, and increment of test-retest reliability in a situation of repeated measures because it can efficiently generate isomorphic items (e.g., at the similar level of difficulty) using item models (Choi, 2018; Wainer, 2002).

In addition, we also applied the Cattell-Horn-Carroll theory of cognitive abilities which is well-known as the comprehensive and integrated model of cognitive abilities by greatly specifying the range of cognitive abilities. Accordingly, two-fold of the current study objectives is (1) to demonstrate how the newly developed nonverbal cognitive ability test is developed using AIG and (2) to examine its psychometric properties. To meet the objectives, we hypothesize that the new measure of cognitive ability based on the CHC theory holds validity and reliability via empirical data analysis.

Method

Cognitive Domains in the Measure of Cognitive Ability

To address the demand of group-administrated nonverbal cognitive measures, we have developed a new measure of cognitive ability (called as MOCA) including item models to measure fluid reasoning (Gf) and visual processing (Gv) in the CHC theory. The notion of an item model in AIG restructures the process of establishing guidelines and standards in traditional item writing by using computer coding. AIG allows item/test writers to develop constructs to be measured a priori utilizing the item/test design principles (Bormuth, 1970; Mislevy, 2018; Thorndike, 1971), which is important because cognitive domains have been identified over the century and its structure/atlas of domains have been understandable and acceptable in the CHC theory.

In the MOCA, the fluid reasoning (Gf) refers to the ability to solve unfamiliar problems without relying on previous knowledge, and the Gf of the MOCA was also specified with three sub-domains: (1) inductive reasoning (IR), (2) general sequential reasoning (SR) known as deductive reasoning, and (3) quantitative reasoning (QR). The visual processing (Gv) refers to the ability to use visual imagery to solve certain problem, and the Gv of the MOCA was measured by examining the capabilities to use simulated mental imagery. For example, the MOCA measures Gv by asking examinees to simulate how the movement of one figure affects another, or how figures might look from a different angle, which will be discussed in detail in the Results section.

Item Models Created in the Measure of Cognitive Ability (MOCA)

Initially, we developed 100 item models covering the two broad abilities, Gf and Gv, which generate item instances measuring four sub-domains: three sub-domains, inductive reasoning (IR), sequential reasoning (SR), and quantitative reasoning (QR) within fluid reasoning (Gf) and one sub-domain, visualization (Vz) within visual processing (Gv). Then, we made two forms of the MOCA test consisting of 36
is the sum of the estimated unstandardized factor loadings among indicators of the same factor, \( \hat{\phi} \) is the estimated factor variance, and \( \sum \hat{\theta}_i \) is the sum of the unstandardized error variances of those indicators. In CFA, factor loadings, error variances, and error covariances are estimated, which influences true and total variance. Thus, to measure factor reliability within CFA model, factor reliability facilitating the CFA estimates is a preferred method to computing Cronbach’s alpha with unrefined composite scores for the scale (Brown, 2006).

**Construct validity.** As another psychometric evaluation, we examined construct validity obtaining convergent validity and discriminant validity in addition to showing that factor loadings are greater than .45 (Brown, 2006, 2015). These indexes of convergent and discriminant validities

### Table 1. Form of the MOCA Based on the Item Models in the Sub-Domains.

|                      | Fluid reasoning (Gf) | Visual processing (Gv) |
|----------------------|-----------------------|------------------------|
|                      | Inductive reasoning (IR) | Sequential reasoning (SR) | Quantitative reasoning (QR) | Visualization (Vz) | Total |
| Unique item models   | 6 (2L, 2M, 2H)        | 3 (2M, 1H)             | 6 (2L, 2M, 2H) | 3 (1L, 1M, 1H) | 18 (5L, 7M, 6H) |
| Common item models   | 6 (2L, 2M, 2H)        | 3 (1L, 1M, 1H)         | 6 (2L, 2M, 2H) | 3 (1L, 1M, 1H) | 18 (6L, 6M, 6H) |
| Total                | 12 (4L, 4M, 4H)       | 6 (1L, 3M, 2H)        | 12 (4L, 4M, 4H) | 6 (2L, 2M, 4H) | 36 (11L, 13M, 12H) |

Note. The number and letter in the parenthesis indicates the number of item models in the difficulty level. L=low; M=medium; H=high.
provide how the underlying constructs are measured accurately and separately to each other, respectively. As noted in Brown (2015), the CFA results provide evidence how strongly indicators of a latent variable are interrelated (convergent validity) and how lowly latent variables are correlated (discriminant validity). Convergent validity was provided by obtaining factor reliabilities are greater than .70 (Nunnally & Bernstein, 1994) and discriminant validity was provided by obtaining factor correlations are lower than .80 (Brown, 2015).

**Results**

**Measure of Cognitive Ability (MOCA)**

**Item models in inductive reasoning (IR).** In the MOCA, as the key sub-domain of fluid reasoning (Gf), IR is defined as “the ability to discern rules and patterns in what is observed.” (Schneider & Newman, 2015). IR ability is measured in the two sub-tests named as figure classification (FC) and figure matrices (FM) depicted in Figure 1. One item instance from a FC item model was shown in Figure 2a with a set of figures as a stem in the item. To answer correctly, examinees should understand the conceptual link and pattern of sliding object to the right-hand corner from the set of figures in A1. And then examinees select the figure that matches the set of figures in the same way. The five options given from the alternatives in the item model were generated with one correct answer and four distractors. The cognitive model for solving the item was depicted via a flowchart in Figure 3.

In Figure 2b, we demonstrated an item instance from a FM item model. Examinees are asked to find the proper objects in the lower right-hand corner. To solve the problem, examinees must induce the following relations: (1) shape changes from row to row, (2) two different objects are moving toward to the center from column to column, and (3) the two objects are laid at the center. Those relations can be induced from A1 and A2 in Figure 2b. If examinees understood the three relations, they should select the proper image for A3 from the options listed that are generated from the alternatives in the item model. As a category of cognition of figural relations, P. I. Jacobs and Vandeventer (1972) pointed out that this kind of item models for IR ability is able to be generated using a principle of cognitive test items and also allows to reduce practice effects by changing patterns and objects. The procedure can be found via a flowchart in Figure 4.

**Item models in quantitative reasoning (QR).** QR referring to “the ability to reason with quantities, mathematical relations, and operators” (Schneider & McGrew, 2018) is assessed in the two other sub-tests called Number Analogies (NA) and Number Puzzles (NP) in the MOCA. For example, as shown in Figure 5a for the NA sub-test, examinees should find out the relation between the two given pairs in A1 and A2 and then they are asked to select the most appropriate number for the question mark (?) in A from the options given below. In the NP item presented in Figure 5b, system of equations with variables denoted by the question mark (?) and the diamond (◊) is provided. Different from items discussed before, examinees are asked to solve an equation in A1 first, by making both sides of the equation equal. After identifying ◊ (=16), examinees apply to the equation A that is 48 ÷ ? = 16. Next, they are required to select the number from the possible options to make both sides of the equation A equal. The item within the NP sub-test is different from the other items in NA, FM, and FC in terms of the procedure that they need to apply backward approach instead of forward approach.

**Item models in sequential reasoning (SR).** SR defined as “the ability to reason logically using known premises and principles” (Schneider & McGrew, 2012) is assessed in the Number Series (NS) in the MOCA where a series of numbers is presented. To solve problems in this sub-test of NS, examinees are asked to select which number from the
Figure 2. Two types of item instances for the inductive reasoning with the fluid reasoning (Gf): (a) an item in figure classification and (b) an item in figure matrices.

Figure 3. Cognitive model of an item (Figure 2a) in figure classification.
1. Encode figures (squares/triangles) and elements (sizes/border lines/moving directions) from A1 and A2.

2. Determine corresponding figures and elements from A1 and A2.

3. Infer a rule from instances given at A1 and A2 (e.g., Smaller figures with thick border lines and dots move in larger figures with dotted border lines)

4. Encode figures (circles) and elements (sizes/moving directions/border lines) of A3.

5. Compare corresponding figures and elements of A3 with figures and elements of A1 and A2.

6. Confirm a rule instance and imagine an answer in your mind.

7. Compare what you imagine in your mind with each response option.

8. Please go back to the stage 1 if there is no correct answer.
Or, please choose the right answer if there is a correct answer.

Figure 4. Cognitive model of an item (Figure 2b) in figure matrices.

Figure 5. Two types of item instances for the quantitative reasoning with the fluid reasoning (Gf): (a) an item in number analogies and (b) an item in number puzzle.
options below should follow. In this test, examinees should find out the relation between the numbers in the series of number. The example of the NS is shown in Figure 6a.

**Item models in visualization (Vz).** Vz within the visual processing (VP) is defined as “the ability to perceive complex patterns and mentally simulate how they might look when transformed (e.g., rotated, twisted, inverted, changed in size, partially obscured).” (Schneider & McGrew, 2018, p. 126). This is the key ability of Gv, and Vz is measured in the sub-test, Paper Folding (PF), where a series of figures in A is presented as shown in Figure 6b. The series of figures are given in a paper of a square or △-shape of a paper that will be further folded with holes cut in circle, triangle, or clover shapes, etc. The one-headed arrow indicates how the paper will be folded. In these situations, examinees must visualize what the figure A looks like in their mind when the paper A is unfolded. Based on the paper folding, examinees should select the most proper figure from the possible answers.

**Reliabilities and Validity Based on CFA of MOCA**

**Model evaluation.** We fit the second order factor model depicted in Figure 7 (adding items in Figure 1) into two datasets collected from two forms of the MOCA at Time 1. The hierarchical structure was considered to distinguish two levels: the sub-domain levels of IR, QR, SR, and Vz nested within Gf (of IR, QR, and SR) and Gv (of Vz) where Gv serves as a phantom construct of Vz. We obtain good fit indexes for both datasets such as RMSEA=0.032 and 0.026, CFI=0.969 and 0.977, and SRMR=0.072 and 0.062 for Form A and Form B, respectively.

**Factor reliability and validity.** We examined the five factor reliabilities for IR (.938 and .877), SR (.899 and .934), QR (.962 and .960), VZ (.916 and .923), and Gf (.808 and .917) for Form A and Form B, respectively, where Gv is same as VZ. These factor reliabilities also provided convergent validities with higher factor loadings on all of four sub-domains that were greater than .45 except item 2 of QR in Form B (.429), item 27 of IR in Form A (.317), and IR of Gf in Form A (.350). Factor correlations between Gf and Gv were .685 for Form A and .642 for Form B that are lower than .80, which provides discriminant validities. Thus, based on CFA results of From A and Form B, construct validity of the MOCA was confirmed.

**Item Analyses Based on Two Parameter Logistic (2PL) Model for the MOCA**

**Item parameter estimates of Form A and Form B.** The hierarchical factor structure of the MOCA was confirmed via CFA. Now, we examined conformity of difficulty level between a priori difficulty levels of item models from AIG and difficulty levels of item instances estimated via item response theory (IRT). When selecting 54 item models for two MOCA tests,
Form A and Form B, we considered three classes in terms of difficulty: low, medium, and high. Due to the complexity of procedure for answering cognitive questions shown at Figures 3 and 4, the three classes may not give any criteria or cut-off in terms of estimated difficulties via IRT. Based on the comparison among three classes, we found that items in the low class have significantly lower estimated difficulties than items in the medium or high classes ($t_{13} = -1.963, -3.687, -2.199,$ and $-2.256$ and $p = .063, .001, .039,$ and .032 on Form A at Time 1, Form A at Time 2, Form B at Time 1, and Form B at Time 2, respectively). However, the estimated difficulties in the medium and high classes were not statistically different from each other. On the other hand, the discrimination parameters were not statistically different in terms of the a priori difficulty levels. The results were summarized in Tables 2 to 4.

Test-retest reliability based on two time points. Using the sample of examinees who took both MOCA at Time 1 and Time 2, we examined the test-retest reliabilities using Pearson correlation coefficients based on the narrow ability factor scores of IR, QR, SR, and Vz. All of correlations of narrow abilities between Time 1 and Time 2 were significant ($p = .000$). In addition, all of correlations were higher than cross-correlations with other narrow abilities except SR (Table 4). However, the correlations were slightly lower ranged from .415 of SR to .606 of QR. As expected, correlations between Vz and any of IR, QR, and SR were lower than other correlations (Figure 8).

Practice Effects
Based on the narrow abilities of IR, QR, SR, and Vz, we examined the mean differences over time using dependent sample $t$-tests. Both Form A and Form B indicated that IR and Vz do not show any change over time while QR and SR do show development at Time 2. The results were listed in Table 5. In terms of sub-tests, we observed that Figure Classification and Figure Matrices of IR, and Paper Folding of Vz do not show any practice effect. On the other hand, Number Series of SR and Number Analogies and Number Puzzles of QR are improved at the second time point.

Discussion
The MOCA is not an exhaustive measure of all cognitive domains as depicted in Figure 1 but it measures the four narrow abilities, inductive reasoning (IR), sequential reasoning (SR), quantitative reasoning (QR), and visualization (Vz), nested within two broad abilities, fluid reasoning (Gf) and visual processing (Gv). However, these four domains are essential in educational setting, which is the reason that many other group-administered measures also include of these domains. On the other hand, MOCA was uniquely created by using AIG, which is beneficial in educational setting because it can be used as repeated measures.

MOCA has not only been used in educational settings to understand students’ strengths and weaknesses in terms of cognition but it will also be launched in clinical settings including Atopic Dermatitis school program (Jang et al., 2015). Although Elbin et al. (2019) recently showed that the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT), the computerized neurocognitive battery, for short-term serial assessment of neurocognitive functioning was suitable in repeated administration, the use of the battery is limited to clinical settings. MOCA also covers broad age groups due to its characteristics of a nonverbal cognitive test.
In this study, MOCA has been validated via psychometrics including item analysis and examining practice effects. Its reliabilities of cognitive domains and the hierarchical structure of six subtests were also validated via a CFA. Its reliability and validity mean more than its verbatim because the properties were drawn from item models of automatic item generation (AIG) instead of item instances. In other words, MOCA developed using AIG was psychometrically sound at the level of item models and also is applicable in intensive longitudinal data analysis utilizing massive item production that is possible due to the isomorphism of item instances generated from item models. Thus, all MOCA assessments for each individual will be different but conform within the same item model set in terms of measuring cognitive ability.

With the psychometric properties, MOCA can be utilized to broader age ranges of students and its capability as a group-administered measure contribute educational equity of taking cognitive ability test. As an AIG-implemented measure, MOCA does not only hold item security so that it can prevent impacts of construct irrelevant variables from practice effects but also eventually provides more economically beneficial tests to students by generating many item instances from item model. As we go through COVID-19 pandemic, educators and parents concern the disparity of student’s academic achievement which is associated with their cognitive ability.

| Narrow domain | Item | A priori | Form A at time 1 | Form A at time 2 | Form B at time 1 | Form B at time 2 |
|---------------|------|----------|-----------------|-----------------|-----------------|-----------------|
|               |      | difficulty | difficulty SE | difficulty SE | difficulty SE | difficulty SE |
| IR Low        | I19  | Low       | 0.976 0.179    | 0.438 0.133    | -0.885 0.102   | -0.149 0.146   |
| Low           | I20  | Medium    | 1.959 0.320    | 0.769 0.283    | 0.904 0.168    | 0.572 0.218    |
| Medium        | I21  | Medium    | 0.825 0.166    | 1.453 0.389    | -0.831 0.143   | -0.427 0.139   |
| Low           | I22  | Low       | 0.499 0.110    | 0.606 0.224    | 0.717 0.124    | 0.819 0.282    |
| Medium        | I23  | Medium    | -0.358 0.125   | 1.745 0.380    | 0.083 0.134    | 1.916 0.507    |
| Low           | I24  | High      | -0.138 0.089   | 0.130 0.160    | 0.295 0.128    | -0.197 0.137   |
| Low           | I25  | Low       | -0.982 0.086   | -0.725 0.085   | -1.072 0.104   | -0.876 0.159   |
| Medium        | I26  | Medium    | 0.181 0.134    | -0.155 0.172   | 0.080 0.101    | -0.546 0.131   |
| Medium        | I27  | Medium    | 3.511 0.769    | 2.047 0.527    | 0.385 0.102    | 0.638 0.262    |
| Low           | I28  | Low       | -0.798 0.105   | -0.458 0.115   | -0.477 0.099   | -0.073 0.150   |
| Medium        | I29  | Medium    | 0.004 0.111    | -0.037 0.131   | -0.041 0.113   | -0.426 0.177   |
| High          | I30  | High      | 1.162 0.228    | 1.194 0.449    | 1.150 0.205    | 0.624 0.296    |
| Medium        | I31  | Medium    | -0.914 0.069   | -0.702 0.131   | -0.877 0.067   | -0.359 0.101   |
| Medium        | I32  | Medium    | -1.026 0.067   | -0.305 0.106   | -0.754 0.053   | -0.342 0.097   |
| High          | I33  | High      | 0.727 0.135    | 0.851 0.228    | -0.210 0.072   | 0.248 0.157    |
| Low           | I34  | Low       | -0.967 0.056   | -0.657 0.075   | -0.885 0.050   | -0.666 0.086   |
| Medium        | I35  | Medium    | -1.033 0.072   | -0.391 0.091   | -0.877 0.059   | -0.398 0.090   |
| High          | I36  | High      | -0.489 0.128   | -0.659 0.168   | -0.387 0.092   | -0.452 0.120   |
| Medium        | I37  | Medium    | 0.495 0.185    | 2.712 0.767    | 1.754 0.327    | 3.563 1.049    |
| High          | I38  | High      | -0.444 0.098   | 0.537 0.240    | 0.510 0.114    | 0.428 0.269    |
| Low           | I39  | Low       | -1.339 0.110   | -0.600 0.157   | -1.351 0.099   | -0.153 0.104   |
| Medium        | I40  | Medium    | -0.379 0.066   | 1.580 0.417    | -0.526 0.077   | 1.172 0.383    |
| High          | I41  | High      | 0.063 0.198    | 0.879 0.307    | -0.043 0.159   | 1.240 0.421    |
| Low           | I42  | Low       | -0.907 0.058   | -0.642 0.068   | -0.485 0.059   | 0.146 0.139    |
| Medium        | I43  | Medium    | -0.842 0.043   | -0.365 0.085   | -0.710 0.059   | -0.366 0.092   |
| High          | I44  | High      | -0.707 0.047   | -0.071 0.108   | -0.903 0.062   | -0.383 0.089   |
| Low           | I45  | Low       | -0.829 0.040   | -0.393 0.087   | -0.734 0.052   | -0.226 0.107   |
| Medium        | I46  | Medium    | -0.872 0.042   | -0.342 0.094   | -0.859 0.053   | -0.240 0.090   |
| High          | I47  | High      | -0.842 0.047   | -0.319 0.086   | -0.833 0.047   | -0.165 0.109   |
| Low           | I48  | Low       | 0.024 0.075    | -0.543 0.112   | -0.078 0.103   | -0.220 0.128   |
| Medium        | I49  | Medium    | 0.153 0.072    | -0.371 0.128   | 0.934 0.164    | 0.863 0.285    |
| Medium        | I50  | Medium    | 0.117 0.090    | -0.182 0.177   | 0.728 0.119    | 1.238 0.370    |
| Low           | I51  | Low       | -0.077 0.083   | 0.155 0.144    | 0.349 0.106    | 0.364 0.240    |
| Medium        | I52  | Medium    | -0.212 0.083   | -0.352 0.158   | -0.120 0.096   | -0.091 0.209   |
| High          | I53  | High      | 0.598 0.128    | 1.224 0.359    | 0.759 0.159    | 1.226 0.315   |
### Table 3. Discriminant Parameter Estimates of the MOCA Items Using 2PL IRT Model.

| Narrow domain | Item | A priori difficulty | Form A at time 1 | Form A at time 2 | Form B at time 1 | Form B at time 2 |
|---------------|------|---------------------|------------------|------------------|-----------------|-----------------|
|               |      |                     | Est. disc | SE  | Est. disc | SE  | Est. disc | SE  | Est. disc | SE  |
| IR            | 119  | Low                 | 0.657   | 0.089 | 1.286   | 0.249 | 1.230   | 0.141 | 1.192   | 0.205 |
|               | 120  | Medium              | 0.563   | 0.088 | 0.620   | 0.147 | 0.706   | 0.098 | 0.860   | 0.175 |
|               | 121  | Medium              | 0.686   | 0.093 | 0.739   | 0.159 | 0.789   | 0.110 | 1.329   | 0.230 |
|               | 122  | Low                 | 0.944   | 0.106 | 0.893   | 0.183 | 0.954   | 0.112 | 0.760   | 0.169 |
|               | 123  | Medium              | 0.736   | 0.098 | 0.788   | 0.162 | 0.652   | 0.099 | 0.600   | 0.144 |
|               | 124  | High                | 1.015   | 0.115 | 1.001   | 0.202 | 0.714   | 0.098 | 1.287   | 0.240 |
|               | 125  | Low                 | 1.583   | 0.189 | 2.453   | 0.578 | 1.397   | 0.159 | 1.426   | 0.289 |
|               | 126  | Medium              | 0.653   | 0.090 | 0.920   | 0.184 | 0.897   | 0.109 | 1.465   | 0.301 |
|               | 127  | Medium              | 0.351   | 0.075 | 0.590   | 0.135 | 1.015   | 0.118 | 0.757   | 0.170 |
|               | 128  | Low                 | 1.088   | 0.134 | 1.525   | 0.312 | 1.001   | 0.113 | 1.162   | 0.234 |
|               | 129  | Medium              | 0.789   | 0.102 | 1.324   | 0.256 | 0.782   | 0.103 | 1.002   | 0.202 |
|               | 130  | High                | 0.545   | 0.084 | 0.437   | 0.111 | 0.630   | 0.094 | 0.630   | 0.152 |
| SR            | 113  | Medium              | 1.932   | 0.175 | 1.459   | 0.317 | 2.131   | 0.209 | 1.990   | 0.335 |
|               | 114  | Medium              | 2.192   | 0.235 | 1.702   | 0.292 | 2.726   | 0.255 | 2.140   | 0.353 |
|               | 115  | High                | 0.810   | 0.099 | 1.107   | 0.224 | 1.375   | 0.135 | 1.201   | 0.230 |
|               | 116  | Low                 | 2.603   | 0.264 | 2.763   | 0.729 | 3.021   | 0.271 | 2.542   | 0.537 |
|               | 117  | Medium              | 2.114   | 0.239 | 2.126   | 0.390 | 2.400   | 0.237 | 2.418   | 0.419 |
|               | 118  | High                | 0.741   | 0.099 | 1.091   | 0.211 | 1.079   | 0.111 | 1.542   | 0.305 |
| QR            | 101  | Low                 | 1.526   | 0.244 | 1.540   | 0.457 | 1.280   | 0.194 | 1.542   | 0.385 |
|               | 102  | Medium              | 0.494   | 0.082 | 0.447   | 0.116 | 0.487   | 0.082 | 0.401   | 0.106 |
|               | 103  | High                | 0.977   | 0.126 | 0.758   | 0.163 | 0.874   | 0.101 | 0.618   | 0.140 |
|               | 104  | Low                 | 1.684   | 0.194 | 1.125   | 0.247 | 1.930   | 0.217 | 2.005   | 0.342 |
|               | 105  | Medium              | 1.599   | 0.155 | 0.603   | 0.135 | 1.333   | 0.146 | 0.577   | 0.137 |
|               | 106  | High                | 0.419   | 0.076 | 0.684   | 0.152 | 0.532   | 0.084 | 0.490   | 0.122 |
|               | 107  | Low                 | 2.484   | 0.201 | 3.871   | 1.309 | 1.864   | 0.183 | 1.368   | 0.239 |
|               | 108  | Medium              | 3.564   | 0.381 | 2.502   | 0.466 | 2.170   | 0.186 | 2.527   | 0.486 |
|               | 109  | High                | 2.967   | 0.267 | 1.732   | 0.255 | 2.277   | 0.212 | 2.546   | 0.485 |
|               | 110  | Low                 | 4.144   | 0.403 | 2.346   | 0.431 | 2.668   | 0.217 | 1.881   | 0.269 |
|               | 111  | Medium              | 3.670   | 0.390 | 2.063   | 0.354 | 2.669   | 0.230 | 2.932   | 0.518 |
|               | 112  | High                | 3.047   | 0.322 | 2.489   | 0.431 | 3.095   | 0.332 | 1.810   | 0.305 |
| Vz            | 131  | Low                 | 1.335   | 0.145 | 1.647   | 0.320 | 0.864   | 0.115 | 1.406   | 0.261 |
|               | 132  | Medium              | 1.527   | 0.165 | 1.325   | 0.269 | 0.798   | 0.105 | 0.792   | 0.179 |
|               | 133  | Medium              | 1.070   | 0.120 | 0.887   | 0.178 | 1.013   | 0.117 | 0.675   | 0.158 |
|               | 134  | Low                 | 1.125   | 0.131 | 1.203   | 0.206 | 0.969   | 0.118 | 0.738   | 0.178 |
|               | 135  | Medium              | 1.108   | 0.135 | 1.036   | 0.211 | 0.936   | 0.113 | 0.767   | 0.173 |
|               | 136  | High                | 0.874   | 0.108 | 0.631   | 0.139 | 0.754   | 0.104 | 0.840   | 0.176 |

### Table 4. Test-Retest Reliabilities of Narrow Domains at Form A and Form B.

| Form   | Form A | Form B |
|--------|--------|--------|
|        | Time 2 | Time 2 |
| Time 1 |        |        |
| IR     | 0.524  | 0.578  |
| QR     | 0.371  | 0.606  |
| SR     | 0.373  | 0.428  |
| Vz     | 0.440  | 0.524  |

Note. P-values of all correlations are less than 0.01; Bold indicates test-retest reliabilities on the diagonal and largest correlations on each row.
ability. MOCA would be a solution to examine the development of student’s cognitive ability.

**Limitations**

In this study, the number of assessments was limited to two time points. It is needed to examine the practice effects in intensive longitudinal data. In addition, linking in item analysis was ignored because this study mainly focused on its psychometrics of reliabilities and validities instead of scale development. As future research, we will examine the status of cognition by considering the anchor items. Although we considered two forms of the MOCA, it has not been considered as age- and grade-based assessment tools. In educational settings, we will consider tailored MOCA tests in terms of grade level and gifted/special education status. In clinical settings, we will consider other tailored MOCA tests consisting of different sets of cognitive domains. As a new measure of cognitive ability, it is also important to explore the concurrent validity with other cognitive measures. However, MOCA was unique in terms of implementing AIG and thus, the concurrent validity study is not a simple way to measure by examining association with the other cognitive measures but requires adjustments between AIG and the traditional item generation. We plan to study the concurrent validity.

**Concluding Remarks**

The MOCA, based on CHC theory and AIG method, was developed as a new measure of cognitive ability that fits studies requiring repeated measures of cognition. This psychometrically sound measure will contribute to cognitive science that researchers need to measure one’s cognition by using the MOCA.

**Acknowledgments**

The authors thank the CAFA AIG team and Hyun Suk Ryoo for technical assistant to implement and administer the MOCA online. We also thank teachers who helped students access the MOCA online.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.
Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by a faculty support grant from the Saban Research Institute of the Children’s Hospital Los Angeles.

Institutional Review Board (IRB)
Data used in this study were approved by the Gyeongsang National University IRB (GIRB-A19-Y-0019)

ORCID ID
Ji Hoon Ryoo https://orcid.org/0000-0001-7546-6911

Supplemental Material
Supplemental material for this article is available online.

References
Beaujean, A. A. (2015). John Carroll’s views on intelligence: Bi-factor vs. higher-order models. *Journal of Intelligence, 3*(4), 121–136. https://doi.org/10.3390/jintelligence3040121

Beglinger, L., Gaydos, B., Tangphadaniels, O., Duff, K., Kareken, D., Crawford, J., Fastenau, P., & Siemens, E. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology, 20*(4), 517–529. https://doi.org/10.1016/j.acn.2004.12.003

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bormuth, J. R. (1970). *On the theory of achievement test items*. University of Chicago Press.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (pp. xiii, 475). Guilford Press.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed., pp. xvii, 462). Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). SAGE.

Bruin, C. J. W. H., Dijkstra, B. A. G., Valvoort, S. J. W., Budy, M. J. J., Beurmanjer, H., De Jong, C. A. J., & Kessels, R. P. C. (2020). Psychometric properties of the Montreal cognitive assessment (MoCA) in healthy participants aged 18–70. *International Journal of Psychiatry in Clinical Practice, 24*(3), 293–300. https://doi.org/10.1016/j.ijpcp.2020.1746348

Bryan, V. M., & Mayer, J. D. (2020). A meta-analysis of the correlations among broad intelligences: Understanding their relations. *Intelligence, 81*(1), 101469. https://doi.org/10.1016/j.intell.2020.101469

Caemmerer, J. M., Keith, T. Z., & Reynolds, M. R. (2020). Beyond individual intelligence tests: Application of Cattell-Horn-Cardell theory. *Intelligence, 79*, 101433. https://doi.org/10.1016/j.intell.2020.101433

Canivez, G. L., & Youngstrom, E. A. (2019). Challenges to the Cattell-Horn-Cardell theory: Empirical, clinical, and policy implications. *Applied Measurement in Education, 32*(3), 232–248. https://doi.org/10.1080/08957347.2019.1619562

Cerulla, N., Arcusa, A., Navarro, J-B., de la Osa, N., Garolera, M., Ener, C., Chico, G., & Fernández-Morales, L. (2019). Cognitive impairment following chemotherapy for breast cancer: The impact of practice effect on results. *Journal of Clinical and Experimental Neuropsychology, 41*(3), 290–299. https://doi.org/10.1080/13803395.2018.1546381

Choi, J. (2018). Roles and impacts of automatic item generation on assessment research, practice, and policy. In S. Swayze & V. Ford (Eds.), *Innovative applications of knowledge discovery and information resources management* (pp. 143–158). IGI Global.

Choi, J., & Zhang, X. (2019). Computerized item modeling practices using computer adaptive formative assessment automatic item generation system: A tutorial. *The Quantitative Methods for Psychology, 15*(3), 214–225.

Dombrowski, S. C., McGill, R. J., & Morgan, G. B. (2021). Monte Carlo modeling of contemporary intelligence test (IQ) factor structure: Implications for IQ assessment, interpretation, and theory. *Assessment, 28*(3), 977–993.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and Testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). American Council on Education/Praeger Publishers.

Elbin, R. J., Fazio-Sumrov, K., Anderson, M. N., D’Amico, N. R., Said, A., Grossel, A., Schatz, P., Lipinski, D., & Womble, M. (2019). Evaluating the suitability of the immediate post-concussion assessment and cognitive testing (ImPACT) computerized neurocognitive battery for short-term, serial assessment of neurocognitive functioning. *Journal of Clinical Neuroscience, 62*, 138–141. https://doi.org/10.1016/j.jocn.2018.11.041

Elman, J. A., Jak, A. J., Panizzon, M. S., Tu, X. M., Chen, T., Reynolds, C. A., Gustavson, D. E., Franz, C. E., Hatton, S. N., Jacobson, K. C., Toomey, R., McKenzie, R., Xian, H., Lyons, M. J., & Kremen, W. S. (2018). Underdiagnosis of mild cognitive impairment: A consequence of ignoring practice effects. *Alzheimer’s & Dementia, 14*(10), 372–381. https://doi.org/10.1016/j.jalz.2018.04.003

Embroton, S., & Yang, X. (2006). 23 Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26, pp. 747–768)*. Elsevier. https://doi.org/10.1016/S0169-7161(06)26023-1

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). John Wiley & Sons, Inc.

Friedenberg, J., & Silverman, G. (2006). *Cognitive science: An introduction to the study of mind*. Sage Publications.

Geisinger, K. F. (2019). Empirical considerations on intelligence testing and models of intelligence: Updates for educational measurement professionals. *Applied Measurement in Education, 32*(3), 193–197. https://doi.org/10.1080/08957347.2019.1619564

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation* (1st ed.). Routledge. https://doi.org/10.4324/9781003025634

Gur, R. C., Richard, J., Hughett, P., Calkins, M. E., Macy, L., Bilker, W. B., Brensingher, C., & Gur, R. E. (2010). A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. *Journal of Neuroscience Methods, 187*(2), 254–262. https://doi.org/10.1016/j.jneumeth.2009.11.017

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty, Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability.
Raykov, T. (1997). Estimation of composite reliability for congeneric measures. Applied Psychological Measurement, 21(2), 173–184. https://doi.org/10.1177/01466216970212006
Raykov, T. (2004). Behavioral scale reliability and measurement invariance evaluation using latent variable modeling. Behavior Therapy, 35(2), 299–331. https://doi.org/10.1016/S0005-7994(04)80041-8
Renzi, M. (2005). Equity, excellence, and economy in a system for identifying students in gifted education: A guidebook. National Research Center on the Gifted and Talented.
Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 99–144). Guilford Publications.
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 Press.
Schneider, W. J., & Newman, D. A. (2015). Intelligence is multidimensional: Theoretical review and implications of specific cognitive abilities. Human Resource Management Review, 25(1), 12–27. https://doi.org/10.1016/j.hrmanr.2014.09.004
Sparrow, S. S., & Davis, S. M. (2000). Recent advances in the assessment of intelligence and cognition. Journal of Child Psychology and Psychiatry, 41(1), 117–131. https://doi.org/10.1017/S1359102199994049
Stemberg, R. J. (1986). Inside intelligence: Cognitive science enables us to go beyond intelligence tests and understand how the human mind solves problems. American Scientist, 74(2), 137–143.
Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. Journal of the International Neuropsychological Society, 5(4), 357–369. https://doi.org/10.1017/S1355617799544068
Thorndike, R. L. (1971). Educational measurement (2nd ed.). American Council on Education.
Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. In S. H. Irvine & P. C. Kyllonen (Eds.), Item generation for test development (pp. 287–305). Erlbaum Associates.
Wasserman, J. D. (2019). Deconstructing CHC. Applied Measurement in Education, 32(3), 249–268. https://doi.org/10.1080/08957347.2019.1619563