Title
Validation of a matrix reasoning task for mobile devices.

Journal
Behavior research methods, 51(5)

ISSN
1554-351X

Authors
Pahor, Anja
Stavropoulos, Trevor
Jaeggi, Susanne M
et al.

Publication Date
2019-10-01

DOI
10.3758/s13428-018-1152-2

Peer reviewed

Published in final edited form as:

Behav Res Methods. 2019 October ; 51(5): 2256–2267. doi:10.3758/s13428-018-1152-2.

Validation of a Matrix Reasoning Task for Mobile Devices

Anja Pahor[1,*], **Trevor Stavropoulos**[1], **Susanne M. Jaeggi**[2], **Aaron R. Seitz**[1]

[1]University of California, Riverside, United States

[2]University of California, Irvine, United States

Abstract

Many cognitive tasks have been adapted for tablet-based testing, but tests to assess nonverbal reasoning ability as measured by matrix-type problems that are suited for repeated testing have yet to be adapted for, and validated on mobile platforms. Drawing on previous research, we developed the University of California Matrix Reasoning Task (UCMRT) - a short, user-friendly measure of abstract problem solving with 3 alternate forms that works on tablets and other mobile devices and that is targeted towards high-ability populations frequently used in the literature (i.e. college students). To test the psychometric properties of UCMRT, a large sample of healthy young adults completed parallel forms of the test and a subsample also completed Raven's Advanced Progressive Matrices, as well as a math test, and furthermore, college records of academic ability and achievement were collected. These data show that UCMRT is reliable and has adequate convergent and external validity. UCMRT is self-administrable, freely available for researchers, facilitates repeated testing of fluid intelligence, and resolves numerous limitations of existing matrix tests.

Keywords

UCMRT; reasoning; fluid intelligence; matrix problems; validity; mobile

Introduction

Fluid intelligence refers to the ability to solve problems without relying on explicit knowledge derived from previous experience (Cattell, 1963). Raven's Advanced Progressive Matrices (APM) is one of the most widely used standardized tests that is used as proxy for higher order cognitive ability (Arthur & Day, 1994), and as such, it is often administered to undergraduate university students. One of the drawbacks of APM is its protracted test time, which ranges from 40 to 60 minutes. Although this may be appropriate for a single testing session, researchers often choose to administer multiple cognitive tests extending the testing time to several hours, which can lead to fatigue and decreased participant engagement (Ackerman & Kanfer, 2009). Moreover, the lack of parallel forms of APM has led researchers to split the test, but due to the limited set of items, this approach only results in two versions, which does not allow for a third assessment, for example in a longitudinal

[*]*Corresponding Author:* UCR Brain Game Center, University Village, 1201 University Ave #204, Riverside, CA 92507, anjap@ucr.edu, (951) 827-2054.

design. It is not always evident that the resulting versions are similar in difficulty (Jaeggi, Buschkuehl, Shah, & Jonides, 2014), and most importantly, the reduced amount of items diminishes test reliability (Sefcek, Miller, & Figueredo, 2016). Matzen and colleagues (2010) addressed those issues by creating a software for systematically generating large numbers of matrix problems, which were validated against Raven's Standard Progressive Matrices (SPM). Even though this presents a solution in terms of quantity of test items, the quality of the drawings presents certain limitations, especially if presented on a small screen.

In general, online tests of reasoning ability can be expensive for researchers and offer little control over the content of the tasks. Recently, there has been a shift towards developing products that are in the public domain, such as the International Cognitive Ability Resource (ICAR), which includes 11 matrix reasoning problems (Condon & Revelle, 2014), as well as 27 progressive matrices. ICAR collaborators have contributed more items that could allow for repeated testing, but no psychometric data is available for these item sets (ICAR Catalogue). A growing number of researchers and healthcare services are adopting cognitive testing through touch screen devices yet to our knowledge, a validated tablet-based measure of analytical intelligence with multiple versions does not exist. To address limitations of current approaches in the field, we developed a modified matrix reasoning test (UCMRT) largely based on matrix problems generated by Sandia Laboratories (Matzen et al., 2010), which (1) only takes 12–15 minutes to complete, (2) has 3 parallel forms, (3) can be administered on tablets and mobile devices, (4) does not require the presence of an administrator, and (5) is designed for people of above average aptitude.

The goals of the present study were to evaluate the psychometric properties of UCMRT, and to examine whether the difficulty level is appropriate for healthy young adults. Specifically, we tested whether the parallel forms are comparable, and we also examined convergent and external validity by comparing performance on the new test with Raven's APM, as well as Math scores, and we further examined the relationship between UCMRT performance and standardized proficiency tests and GPA.

Methods

Participants

A total of 713 participants (Mean age = 20.02 years, SD = 2.74; N_{Female} = 494, N_{Male} = 201, $N_{Other/Unknown}$ = 18) conducted at least one session of UCMRT over a period of 6 months. Participants were a sample of ethnically diverse university students (cf. Table S1, Supplemental Materials) recruited at University of California, Riverside (N = 353) or University of California, Irvine (N = 360) with normal or corrected-to-normal vision. All participants provided informed consent and received either course credit or monetary compensation for participation. Most participants (N = 676) were asked for permission to obtain Educational Records, which was granted by 416 students. A subset of the participants also completed a newly developed tablet-based Math Test (N = 483), and some of these (N = 238) also completed Raven's Advanced Progressive Matrices, allowing for analysis of convergent and external validity (see section "Convergent Validity Tasks" for more information). Another subset of participants (N = 213) performed alternate versions of

UCMRT on two occasions which allowed us to calculate test-retest reliability. All study procedures were approved by the UCR and UCI Institutional Review Boards.

University of California Matrix Reasoning Task (UCMRT)

UCMRT consists of Raven-like matrix problems, most of which are based on the matrices produced by Sandia National Laboratories[1]. The software developed by this laboratory can systematically generate a large number of matrices, which provide a good match to Raven's Standard Progressive Matrices (Matzen et al., 2010). However, we believe that the Sandia matrix problems are limited in terms of graphics: overlapping items are difficult to discern, and some of the details are very small, making it difficult to judge shape and shading properties crucial to solving the problem (cf. Figure 1). Hence, we redesigned the three versions of the task using larger, non-overlapping stimuli that are not limited to gray-scale, while keeping the same number and types of rules and the basic structure of the task (3×3 matrix with 8 answer alternatives). Because our target population was healthy young adults, we excluded the easy, one-relation Sandia matrix problems, in which only one rule governs the patterns of changes in the matrix, and focused instead on problems that contain at least two relations or are logic-based. While the answer options are presented horizontally in Sandia Matrices (2 rows, 4 columns), the answer options in UCMRT are presented vertically to maximize the space available on mobile devices (4 rows, 2 columns; cf. Figure 2).

Each alternate form of UCMRT consists of 2 example problems, 6 practice problems, and 23 test problems. During practice, participants receive feedback (correct/incorrect) along with an explicit explanation of the rules that must be combined to solve the problem (Hossiep, Turck, & Hasella, 1999). If a participant fails to pass the practice criteria (at least 4 correct out of 6), the practice is repeated once with a different set of problems. The test portion consists of 2 two-relation problems at the beginning, followed by a mixed order of 15 three-relation problems and 6 logic problems with a time limit of 10 minutes. No feedback is provided, but participants can change their answer, skip problems, and navigate back and forth between the problems[2], akin to paper and pencil versions of Raven's Matrices (Raven, Raven, & Court, 1998). The problem number (e.g. 5/23) is shown in the top right corner of the screen and a countdown timer is presented in the upper left corner (cf. Figure 2). The task ends when the participants submits the answers or when the 10-minute limit is reached.

Matrix Reasoning problems can be broken down by type as defined by Matzen et al. (2010). UCMRT contains two types of object transformation problems: 2-relation problems (2 rules govern the pattern of changes within a matrix; cf. Figure 2.2) and 3-relation problems (3 rules make up the pattern of changes; cf. Figures 2.4 – 2.6). The rules featured in these problems are shape, orientation, size, number, as well as shading/color[3],[4]. The 3-relation problems can be further distinguished based on the number of diagonal or outward

[1]L. E. Matzen, Sandia National Laboratories, P.O. Box 5800, Mail Stop 1188, Albuquerque, NM 87185–1188

[2]The log files record each action the participant makes on a given problem, which enables researchers to examine problem solving patterns and time spent per each problem, a feature that is not available in paper and pencil versions of matrix problems.

[3]Different shades and colors are characterized by different levels of luminance and should be distinguishable to participants with color vision deficiencies.

[4]All subjects completed a color naming test. Two participants with potential color vision deficiencies (green named as red; green named as yellow) performed within normal range (15/23 and 13/23 correct).

transformations (1–3) in the matrix. UCMRT also contains three types of logic problems (cf. Figure 2.3): addition/conjunction (AND), disjunction (OR), and exclusive disjunction (XOR). Each alternate version consists of two 2-relation problems, three 3-relation problems with 1 transformation, six 3-relation problems with 2 transformations, six 3-relation problems with 3 transformations, and six logic problems. To control for context effects, problems are not ordered by problem type (with the exception of 2-relation problems) nor are they specifically ordered in terms of difficulty apart from the overall structure in which 2-relation and 3-relation problems with fewer transformations are distributed at the beginning, while those that require more transformations are distributed towards the end. In all three versions, the order is the same with respect to problem type, however, the rules that govern the problem type can differ[5].

Implementation

UCMRT runs on Unity, which supports multiple platforms and can be utilized to make the test more available over time. Presently, UCRMT is supported by iOS and Android, and can be released on other platforms per request. The app supports creation of usernames that can be used to hide the identity of the participant. Researchers who wish to use UCMRT are encouraged to contact one of the authors. Data files are logged locally (text files are stored on the device itself), in addition, they are logged on an Amazon-housed server as long as a wireless internet connection is maintained. Server-stored data can be made available to researchers upon request.

Convergent Validity Tasks

Two tasks were used to assess convergent validity of UCMRT. The first task was Raven's Advanced Progressive Matrices Set II (Raven et al., 1998), a valid and widely used measure of fluid intelligence (Arthur & Woehr, 1993). APM Set II consists of 36 problems in ascending difficulty. Each problem comprises a 3×3 matrix with the lower right entry missing, and the participant has to select one out of 8 answer options that best completes the matrix. The participants solved a paper and pencil version of the task, using Set I as practice (12 problems) and Set II as the test, with a time limit of 25 minutes. This time restriction was applied to approximate the conditions in UCMRT and to avoid ceiling performance often observed in our lab with longer or no time restrictions[6]. The second task was a tablet-based Math Test consisting of 21 word-based math problems with a time limit of 15 minutes (cf. Math Test, Supplemental Materials) administered via Qualtrics Software (Qualtrics, Provo, UT).

Educational Data

College grade point average (GPA), high school GPA, as well as SAT and ACT scores were obtained from university records. GPA reflects overall performance during the entire academic year and is on a 4.0 scale. SAT and ACT are standardized tests used for college

[5]For example, the rules that govern a given 3-relation problem with 3 transformations are Shape, Shading, and Orientation in Version A, Shape, Size and Number in Version B, and Shape, Orientation and Size in Version C.

[6]Timed performance on APM is predictive of untimed performance on the same test (Frearson & Eysenck, 1986; Hamel & Schmittmann, 2006; Salthouse, 1993; Unsworth & Engle, 2005).

admission in the United States. The tests are somewhat different, but are universally accepted by colleges and universities, thus students can decide which test they want to take. Since it was unclear which scores were affected by recent major changes in SAT sections, particularly on verbal measures (furthermore, some scores were not scaled), and given that the maximum SAT score changed from of 2400 to 1600 in 2016, we decided to exclude SAT Verbal/Evidence-based Reading and Writing and SAT Total measures from the analyses. While certain changes were implemented on SAT Math, the scale has remained the same (200–800) and therefore, it was not excluded. For ACT, scores on Reading, Math, Writing, and Science sections were obtained, all of which were included in the analyses.

Procedure

Performance on UCMRT was obtained in the context of three studies. In one study, participants completed a battery of tablet-based cognitive tests, including the Math Test and UCMRT, which were both presented at the beginning of a session (in that order). In another study, we aimed to estimate improvement on the battery of cognitive tests (excluding the Math Test) and assess testretest reliability, thus the participants were asked to attend a second session 10–14 days later. Both sessions took place at the same time of day, and UCMRT was completed at the beginning of each session. Participants were randomly divided into 3 groups: one group completed Version A in Session 1 and Version B in Session 2, the other completed Version B in Session 1 and Version C in Session 2, and the third completed Version C in Session 1 and Version A in Session 2. In a third study, participants were randomly assigned to complete one of the three versions of UCMRT (A, B, and C), one of the two versions of the Math Test (I and II), as well as the APM test. The order of the two matrix reasoning tests (APM and UCMRT) was counterbalanced and separated by the Math Test.

Statistical Analyses

SPSS Version 24 and JASP Version 0.9.0.1 (JASP Team, 2018) were used to analyze the data. The results of Frequentist analyses are supplemented with Bayes Factors, specifically BF_{10}, which grades the intensity of the evidence that the data provide for H_1 versus H_0. BF_{10} between 1 and 3 is considered to be only anecdotal evidence for H_1; 3–10: moderate evidence; 10–30: strong evidence; 30–100: very strong evidence; >100 extreme evidence (Lee & Wagenmakers, 2013; Wagenmakers et al., 2018).

Results

UCMRT Performance and Reliability

Ten outliers (1.4%) were removed from the sample: six individuals based on the number of responses (i.e. skipping at least 48% of the problems; z >= 4), and four based on the number of correctly solved problems (|z| >= 2.5). Descriptive statistics are presented in Table 1. There was no statistical difference in performance comparing the three versions of UCMRT as determined by a one-way ANOVA ($F(2,700) = 2.70$, $p = 0.07$, $\eta^2 = 0.01$), which was further confirmed with a JZS Bayes Factor ANOVA with default prior scales ($BF_{10} = 0.21$). While Bonferroni-corrected post-hoc tests showed no significant differences between the pairs of tests (A-B: $p = 0.63$, B-C: $p = 0.82$, and A-C: $p = 0.06$), Bayesian post-hoc tests

revealed anecdotal evidence in favor of H1 for the A-C comparison (A-C: $BF_{10,U} = 1.78$), but this was not observed for the A-B ($BF_{10,U} = 0.21$) or B-C ($BF_{10,U} = 0.18$) comparisons.

The internal consistency of the 23 problems, based on Cronbach's α, was .66, .76 and .72 in versions A, B and C, respectively (combined versions: α = .71; cf. Table 1). Similarly, Cronbach's α for sets of Sandia matrices consisting of 42 problems each was .76 (Matzen et al., 2010). For the 36 items of Raven's APM, Cronbach's α was 0.82 in our sample and 0.84 in other data sets (Arthur & Day, 1994; Bors & Stokes, 1998). For a short version consisting of 18 problems, α of .64 (Unsworth, Redick, Lakey, & Young, 2010) and .79 (Sefcek et al., 2016) was reported. Considering the relatively low number of problems and short testing time of UCMRT, the internal consistency of the three versions seems adequate.

We also report descriptive statistics for a subsample that was subject to more rigorous selection criteria (Table 1). Only participants who correctly solved the first two problems of UCMRT (i.e. easy, 2-relation problems) were included in the analysis. Performance on the three versions of UCMRT was not statistically different as determined by a one-way ANOVA ($F(2,384) = 0.22$, $p = 0.80$, $\eta^2 = 0.001$; $BF_{10} = 0.04$). Bonferroni-corrected post-hoc tests showed no significant differences between the pairs of tests (A-B: $p = 1$; $BF_{10,U} = 0.16$, B-C: $p = 1$; $BF_{10,U} = 0.13$, and A-C: $p = 1$; $BF_{10,U} = 0.17$). By demonstrating an understanding of the rules that apply to the 2relation problems, this subsample may be particularly suited for UCMRT testing. Nevertheless, as demonstrated above, the entire sample also showed adequate performance.

Problem Type

Overall, accuracy decreased as the number of relations/transformations increased ($M_{2\text{-REL}}$ = 73.47%, SE2-REL = 1.21%; M3-REL-1 = 68.94%, SE3-REL-1 = 1.08%; M3-REL-2 = 61.93%, SE3-REL- = 0.89%; $M_{3\text{-REL-3}} = 38.48\%$, $SE_{3\text{-REL-3}} = 0.88\%$)[7], which is consistent with the findings reported by Matzen et al. (2010) in a sample of undergraduate university students. For logic-based problems, average accuracy across all versions was 50.88% (SE = 1.00%), which is higher than that reported for Sandia Logic Problems (M = 37.9%, SE = 3%). One of the reasons for this may be improved visibility of details and shapes, which are non-overlapping in UCMRT (cf. Figure 2), and the inclusion of the practice problems in our version. Figure 3 shows average accuracy based on problem type, presented separately for participants that completed different versions of UCRMT and correctly solved the first two problems (whole sample data presented in Figure S1, Supplemental Materials). One-way ANOVAs were used to determine whether performance on the three versions differed on each problem type. For 3-relation problems with 1 transformation, performance on the three versions of the test was not statistically different ($F(2,384) = 2.21$, $p = 0.11$; $BF_{10} = 0.23$). For 3-relation problems with 2 transformations, both types of analyses indicated that there was a statistically significant difference in performance between the three versions ($F(2,384) = 7.32$, $p < 0.05$; $BF_{10} = 19.48$). Bonferroni-corrected post-hoc tests showed that performance on Version A was significantly lower than performance on Version B ($p = 0.02$; $BF_{10,U} = 10.31$) and performance on Version C ($p < 0.001$; $BF_{10,U} = 142.92$), but there was

[7] N = 703; collapsed across versions

no significant difference between versions B and C ($p = 1$; $BF_{10,U} = 0.15$). For 3-relation problems with 3 transformations ($F(2,384) = 1.23$, $p = 0.29$; $BF_{10,U} = 0.09$) and for Logic problems ($F(2,384) = 0.16$, $p = 0.85$; $BF_{10,U} = 0.03$), no significant difference in performance was observed (cf. Figure 3). Overall, performance on the subtypes of problems is well matched with the exception of 3-relation problems with 2 transformations. This may not be problematic since, as demonstrated in the previous section, the final scores on the three versions are adequately balanced, and participants using A also scored worse on the APM (see below).

Test-retest

Eight outliers (3.8%) were removed ($z_{NoResponse} >= 4$ and $z_{Hits} >= 2.5$) from the dataset in which participants completed two sessions (N = 213). Descriptive statistics for performance on different versions of UCMRT at two time points are presented in Table 2. Test-retest reliability as measured by Pearson's correlation coefficient was 0.62 ($p < 0.001$; $BF_{10} > 100$). A paired samples t-test collapsed across versions showed that performance in the two sessions was not significantly different ($t(204) = -1.82$, $p = 0.07$, Cohen's d[8]$= 0.13$; $BF_{10} = 0.40$). These results may be of use to future interventions studies trying to estimate improvement on alternate versions of UCMRT in the absence of an intervention. Repeated testing on APM divided into parallel forms (18 items per test) in no-contact control groups has shown similar results, with Cohen's d effect sizes ranging from 0.10 (Jaeggi et al., 2010) to 0.38 (Hogrefe, Studer-Luethi, Kodzhabashev, & Perrig, 2017) and some even reporting worse performance in the second session relative to the first (Clark, Lawlor-Savage, & Goghari, 2017; Colom et al., 2013; Redick et al., 2013; Stough et al., 2011).

On the other hand, significant changes in performance in the two sessions were observed as a function of group (Table 2). Paired samples t-tests showed that in Group 1, accuracy increased in the second session relative to the first ($t(67) = 3.91$, $p < 0.001$, Cohen's $d = 0.47$; $BF_{10} = 102.3$) and a similar trend was observed for Group 2 ($t(67) = 1.83$, $p = 0.07$, Cohen's $d = 0.22$; $BF_{10} = 0.64$). In contrast, Group 3 showed a decrease in performance in the second session relative to the first ($t(68) = -2.54$, $p = 0.01$, Cohen's $d = - 0.31$; $BF_{10} = 2.57$).

While we were not able to test all possible pairs of alternate forms, the results shown in Figure 4 suggest that different groups of participants show similar performance on the same version of the test in different sessions. Independent-samples t-tests showed that there were no significant differences in performance on the three versions of UCMRT at different points in time (Version A: $t(135) = -.45$, $p = 0.66$, Cohen's $d = -0.08$, $BF_{10} = 0.20$; Version B: $t(134) = -1.77$, $p = 0.08$, Cohen's $d = -0.30$, $BF_{10} = 0.76$; Version C: $t(135) = 0.21$, $p = 0.83$, Cohen's $d = 0.04$, $BF_{10} = 0.19$).

[8]Accounts for the correlation between pre- and post-test measures:
$(Mean_{Post} - Mean_{Pre}) / \sqrt{(SD^2_{Pre} + SD^2_{Post} - 2r_{PrePost} * SD_{Pre} * SD_{Post})}$

Convergent and External Validity

Raven's APM

Two additional outliers were removed based on performance on Raven's APM Set II (z >= 2.5). Performance on UCMRT significantly correlated with performance on APM both in the total sample ($N = 233$, $r = .58$, $p < 0.001$; $BF_{10} > 100$) and in the three subgroups of participants (Group A: $N = 79$, $r = .44$, $p < 0.001$, $BF_{10} > 100$; Group B: $N = 80$, $r = .67$, $p < 0.001$, $BF_{10} > 100$; Group C: $N = 74$, $r = 0.58$, $p < 0.001$, $BF_{10} > 100$) (cf. Figure 5). Matzen et al. (2010) reported a correlation of .69 between accuracy for Sandia matrices and for Raven's Standard Progressive Matrices (SPM). However, the authors only included a subset of SPMs in the analysis that had structures that were comparable to those of the Sandia problems. Instead, we report correlations with all problems in APM Set II, many of which are governed by noncomparable rules and transformations.

The difficulty level of the two matrix reasoning tests is comparable; in the sample of 233 participants, average accuracy was 51.9% (SD = 17.2) on UCMRT and 53.5% (SD = 14.9) on APM. A paired samples *t*-test showed that performance on the two tests was not significantly different ($t(232) = -1.59$, $p = 0.11$, Cohen's $d = 0.10$; $BF_{10} = 0.25$). As can be seen in Figure 6, APM accuracy for the three groups follows a pattern similar to UCMRT accuracy, with group A performing slightly worse than groups B and C, which may indicate inherent group differences in reasoning ability, suggesting that small differences between the versions may reflect cohort effects rather than differences in difficulty between the A, B and C measures.

Math Test

We developed two alternate versions of the Math Test (cf. Supplemental Materials): 255 participants completed version I (M = 8.47, SD = 2.86) and 232 participants completed version II (M = 8.06, SD = 3.23). Performance on the two versions was not statistically different as determined by an independent samples t-test ($t(485) = -1.49$, $p = 0.14$; $BF_{10} = 0.29$) hence the data was collapsed. As expected, performance on the Math test showed strong correlations with SAT Math ($r = .55$, $p < 0.001$; $BF_{10} > 100$) and ACT Math ($r = .61$, $p < 0.001$; $BF_{10} > 100$). Further, performance on the Math Test correlated significantly with performance on UCMRT ($r = .36$; $p < 0.001$; $BF_{10} > 100$) and APM ($r = .34$, $p < 0.001$; $BF_{10} > 100$) providing further evidence for the validity of UCMRT (see Table 3).

Tests of Academic Proficiency

Accuracy on UCMRT and APM was correlated with academic proficiency, where available (see Table 3). Performance on both UCMRT and APM showed small, but significant correlations with current college GPA ($r_{UCMRT} = .13$, $p = 0.007$, $BF_{10} = 2.19$; $r_{APM} = .16$, $p = 0.045$, $BF_{10} = 0.75$). This is line with previous research in which performance on APM was somewhat weakly related to first-term ($r = .19$, $p < 0.05$) and second-term college GPA ($r = .18$, $p < 0.05$) (Coyle & Pillow, 2008). On a side note, the only measure that correlated significantly with high school GPA was college GPA ($r = .16$, $p = 0.003$; $BF_{10} = 5.03$)[9]. While high school GPA and performance on college-admission tests is typically highly correlated (Koretz et al., 2016; Westrick, Le, Robbins, Radunzel, & Schmidt, 2015), this was

not observed in the present sample, at least not in ACT and SAT subtests included in the analysis.

Performance on both UCMRT and APM also correlated with SAT Math (r_{UCMRT} = .45, p < 0.001, BF_{10} > 100; r_{APM} = .44, p < 0.001, BF_{10} > 100) and ACT Math (r_{UCMRT} = .35, p < 0.001, BF_{10} > 100; r_{APM} = .35, p = 0.02, BF_{10} = 3.26), which is consistent with the literature (Koenig, Frey, & Detterman, 2008; Rohde & Thompson, 2007). On the other hand, we did find differential correlations as a function of matrix reasoning test in various academic measures: Performance on UCMRT, but not on APM, predicted ACT Reading (r_{UCMRT} = .27, p < 0.001, BF_{10} = 52.53; r_{APM} = .04, p = 0.790, BF_{10} = 0.19), ACT Writing (r_{UCMRT} = .29, p < 0.001, BF_{10} > 100; r_{APM} = .03, p = 0.862, BF_{10} = 0.18), and ACT Science (r_{UCMRT} = .35, p < 0.001, BF_{10} > 100; r_{APM} = .07, p = 0.639, BF_{10} = 0.20), suggesting that UCMRT is predictive of global cognitive function. It should be noted that the sample size for APM was smaller; however, based on the correlation coefficients (0.03 – 0.07) and visual inspection of scatter plots (cf. Figure S2, Supplemental Materials), a stronger relation with ACT subtests is not expected even if the sample size was larger. There is limited research indicating that SAT and ACT composite scores predict performance on tests of general ability (Coyle & Pillow, 2008; Koenig et al., 2008) and furthermore, ACT subtests show significant correlations with reasoning ability (Goff & Ackerman, 1992); hence the validity of UCMRT is line with previous work.

Individual versions of UCMRT showed similar relations, albeit the evidence is limited by a smaller sample size in certain cases (cf. Tables S4 – S6, Supplemental Materials). Version A showed significant correlations with SAT Math (r =.48, p < 0.001, BF_{10} > 100), ACT Writing (r =.30, p = 0.028, BF_{10} = 1.80), and ACT Science (r =.40, p = 0.003, BF_{10} = 12.51). Version B significantly correlated with SAT Math (r =.46, p < 0.001, BF_{10} > 100), ACT Math (r =.40, p = 0.003, BF_{10} = 13.18), and ACT Science (r =.28, p = 0.044, BF_{10} = 1.22). Version C was related to SAT Math (r =.40, p < 0.001, BF_{10} > 100), ACT Reading (r =.37, p = 0.002, BF_{10} = 14.46), ACT Math (r =.40, p < 0.001, BF_{10} = 31.03), ACT Writing (r =.33, p = 0.007, BF_{10} = 5.37), and ACT Science (r =.34, p = 0.006, BF_{10} = 6.17). In summary, all three versions correlated with APM (Figure 5), the Math Test, as well as SAT Math and ACT Science, and all but version A correlated with ACT Math.

Discussion

UCMRT is a tablet-based matrix reasoning test with three parallel versions that can serve as proxy of fluid intelligence. The A, B and C versions of the test were validated in a large sample of college students at two different sites. Overall, performance was similar across the three versions and showed adequate internal consistency. Alternate form reliability was comparable to the literature (Colom et al., 2010; Freund & Holling, 2011; Unsworth, Heitz, Schrock, & Engle, 2005), suggesting that the participants did not show significant increases in performance over time and that versions were comparable. Convergent validity was established by comparing UCMRT to Raven's APM. In addition, performance on UCMRT

[9]Note that the correlation in the present study may be lower because we obtained records of current GPA, rather than just 1st-year GPA, resulting in a mixture of first year and higher year GPA.

correlated with a Math Test, College GPA, as well as with scores obtained on college admissions tests, showing similar correlations as APM, thereby demonstrating external validity. In fact, performance on UCMRT correlated significantly with all ACT subtest scores (Math, Science, Reading and Writing) whereas performance on APM only showed significant correlations with ACT Math.

Matrix reasoning problems have been the hallmark of tests estimating non-verbal fluid intelligence for over 8 decades (Raven, 1938/1956). Matrices adapted to different ability levels were developed over time, yet these remained remarkably unchanged in the past two decades (Raven et al., 1998). The core aspects of these types of problems, such as ease of administration and relative independence from language, are worth retaining; however, faced with a limited set of test items, many of which can be found online, it is time to expand the set and to adapt it for modern technology. While there have been certain efforts in this direction[10], UCMRT is unique in that it consists of multiple validated versions that will be freely available to researchers. In doing this, we hope to collect large data sets, and attempt to unify studies using a wide variety of fluid intelligence tasks. The parallel forms of UCMRT are especially appropriate for longitudinal studies, for example those investigating the effects of an intervention at multiple time points. While alternate versions may show small differences in terms of accuracy, in our case, they could be driven by inherent group differences, and such issues can be addressed with counterbalancing where test repetition is needed.

Perhaps the greatest advantage of UCMRT is its short administration time and that it is self-administrable, which allows for remote testing. The log files instantly provide the number of problems that were solved correctly, incorrectly, or skipped, which is easily understandable for researchers, clinicians, and users alike. More detailed log files are also available that provide insight into problem-solving patters and reaction times, features that are not available in standard paper and pencil tests.

Since the rules that make up the problems are clearly defined, there is no limit to the number of items that can be produced. In the future we plan to develop the items procedurally, leading to even more parallel versions of UCMRT. In addition, we plan to release sets of items that are better customized for different age- and ability-groups.

Conclusions

Overall, our data suggest that UCMRT is a reliable and valid measure of non-verbal problem solving that is predictive of academic proficiency and could serve as a proxy of fluid intelligence. Moreover, it can be used to differentiate among people at the high end of intellectual ability, akin to Raven's Advanced Progressive Matrices (APM). Compared to APM, the UCMRT is shorter (10 minutes excluding practice), offers three parallel-test versions, and can be used on iOS and Android devices.

[10] https://www.cambridgebrainsciences.com; https://icar-project.com

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work has been supported by the National Institute of Mental Health (Grant No. 1R01MH111742–01), and in addition, SMJ is supported by the National Institute on Aging (Grant No. 1K02AG054665–01). SMJ has an indirect financial interest with the MIND Research Institute, whose interest is related to this work.

We thank Laura E. Matzen for providing access to the matrices generated by Sandia National Laboratories.

Bibliography

Ackerman PL, & Kanfer R (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. Journal of Experimental Psychology. Applied, 15(2), 163–181. doi:10.1037/a0015719 [PubMed: 19586255]

Arthur W, & Day DV (1994). Development of a Short form for the Raven Advanced Progressive Matrices Test. Educational and Psychological Measurement, 54(2), 394–403. doi:10.1177/0013164494054002013

Arthur W, & Woehr DJ (1993). A confirmatory factor analytic study examining the dimensionality of the raven's advanced progressive matrices. Educational and Psychological Measurement, 53(2), 471–478. doi:10.1177/0013164493053002016

Bors DA, & Stokes TL (1998). Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. Educational and Psychological Measurement, 58(3), 382–398. doi:10.1177/0013164498058003002

Cattell RB (1963). Theory of fluid and crystallized intelligence: A critical experiment. Journal of Educational Psychology, 54(1), 1–22. doi:10.1037/h0046743

Clark CM, Lawlor-Savage L, & Goghari VM (2017). Working memory training in healthy young adults: Support for the null from a randomized comparison to active and passive control groups. Plos One, 12(5), e0177707. doi:10.1371/journal.pone.0177707 [PubMed: 28558000]

Colom R, Quiroga MÁ, Shih PC, Martínez K, Burgaleta M, Martínez-Molina A, … Ramírez I (2010). Improvement in working memory is not related to increased intelligence scores. Intelligence, 38(5), 497–505. doi:10.1016/j.intell.2010.06.008

Colom R, Román FJ, Abad FJ, Shih PC, Privado J, Froufe M, … Jaeggi SM (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. Intelligence, 41(5), 712–727. doi:10.1016/j.intell.2013.09.002

Condon DM, & Revelle W (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. Intelligence, 43, 52–64. doi:10.1016/j.intell.2014.01.004

Coyle TR, & Pillow DR (2008). SAT and ACT predict college GPA after removing g. Intelligence, 36(6), 719–729. doi:10.1016/j.intell.2008.05.001

Frearson W, & Eysenck HJ (1986). Intelligence, reaction time (RT) and a new 'odd-man-out' RT paradigm. Personality and Individual Differences, 7(6), 807–817. doi:10.1016/0191-8869(86)90079-6

Freund PA, & Holling H (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. Intelligence, 39(4), 233–243. doi:10.1016/j.intell.2011.02.009

Goff M, & Ackerman PL (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. Journal of Educational Psychology, 84(4), 537–552. doi:10.1037/0022-0663.84.4.537

Hamel R, & Schmittmann VD (2006). The 20-Minute Version as a Predictor of the Raven Advanced Progressive Matrices Test. Educational and Psychological Measurement, 66(6), 1039–1046. doi:10.1177/0013164406288169

Hogrefe AB, Studer-Luethi B, Kodzhabashev S, & Perrig WJ (2017). Mechanisms Underlying N-back Training: Response Consistency During Training Influences Training Outcome. Journal of Cognitive Enhancement, 1(4), 406–418. doi:10.1007/s41465-0170042-3

Hossiep R, Turck D, & Hasella M (1999). Bochumer Matrizentest. BOMAT advanced.

Jaeggi SM, Buschkuehl M, Shah P, & Jonides J (2014). The role of individual differences in cognitive training and transfer. Memory & Cognition, 42(3), 464–480. doi:10.3758/s13421-013-0364-z [PubMed: 24081919]

ICAR Catalogue. Version 1.0, 06 I 17. Retrieved from https://icarproject.com/ICAR_Catalogue.pdf on 08/19/2018.

Jaeggi SM, Studer-Luethi B, Buschkuehl M, Su Y-F, Jonides J, & Perrig WJ (2010). The relationship between n-back performance and matrix reasoning — implications for training and transfer. Intelligence, 38(6), 625–635. doi:10.1016/j.intell.2010.09.001

JASP Team (2018). JASP (Version 0.9.0.1) [Computer software]

Koenig KA, Frey MC, & Detterman DK (2008). ACT and general cognitive ability. Intelligence, 36(2), 153–160. doi:10.1016/j.intell.2007.03.005

Koretz D, Yu C, Mbekeani PP, Langi M, Dhaliwal T, & Braslow D (2016). Predicting freshman grade point average from college admissions test scores and state high school test scores. AERA Open, 2(4), 233285841667060. doi:10.1177/2332858416670601

Lee MD, & Wagenmakers E-J (2013). Bayesian cognitive modeling: A practical course. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139087759

Matzen LE, Benz ZO, Dixon KR, Posey J, Kroger JK, & Speed AE (2010). Recreating Raven's: software for systematically generating large numbers of Raven-like matrix problems with normed properties. Behavior Research Methods, 42(2), 525–541. doi:10.3758/BRM.42.2.525 [PubMed: 20479184]

Qualtrics [Computer software]. Copyright © 2018 Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA. https://www.qualtrics.com

Raven JC (1938/1956). Guide to Progressive Matrices. (rev. ed.). London, H.K. Lewis.

Raven J, Raven JC, & Court JH (1998). Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4. Advanced Progressive Matrices Sets I & II. Oxford, UK: Oxford Psychologists Press.

Redick TS, Shipstead Z, Harrison TL, Hicks KL, Fried DE, Hambrick DZ, … Engle RW (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. Journal of Experimental Psychology. General, 142(2), 359–379. doi:10.1037/a0029082 [PubMed: 22708717]

Rohde TE, & Thompson LA (2007). Predicting academic achievement with cognitive ability. Intelligence, 35(1), 83–92. doi:10.1016/j.intell.2006.05.004

Salthouse TA (1993). Influence of working memory on adult age differences in matrix reasoning. British Journal of Psychology, 84 (Pt 2), 171–199. [PubMed: 8319054]

Sefcek JA, Miller GF, & Figueredo AJ (2016). Development and Validation of an 18Item Medium Form of the Ravens Advanced Progressive Matrices. SAGE Open, 6(2), 215824401665191. doi:10.1177/2158244016651915

Stough C, Camfield D, Kure C, Tarasuik J, Downey L, Lloyd J, … Reynolds J (2011). Improving general intelligence with a nutrient-based pharmacological intervention. Intelligence, 39(2–3), 100–107. doi:10.1016/j.intell.2011.01.003

Unsworth N, & Engle R (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. Intelligence, 33(1), 67–81. doi:10.1016/j.intell.2004.08.003

Unsworth N, Heitz RP, Schrock JC, & Engle RW (2005). An automated version of the operation span task. Behavior Research Methods, 37(3), 498–505. doi:10.3758/BF03192720 [PubMed: 16405146]

Unsworth N, Redick TS, Lakey CE, & Young DL (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. Intelligence, 38(1), 111–122. doi:10.1016/j.intell.2009.08.002

Wagenmakers E-J, Love J, Marsman M, Jamil T, Ly A, Verhagen J, … Morey RD (2018). Bayesian inference for psychology. Part II: Example applications with JASP. Psychonomic Bulletin & Review, 25(1), 58–76. doi:10.3758/s13423-017-1323-7 [PubMed: 28685272]

Westrick PA, Le H, Robbins SB, Radunzel JMR, & Schmidt FL (2015). College Performance and Retention: A Meta-Analysis of the Predictive Validities of ACT® Scores, High School Grades, and SES. Educational Assessment, 20(1), 23–45. doi:10.1080/10627197.2015.997614

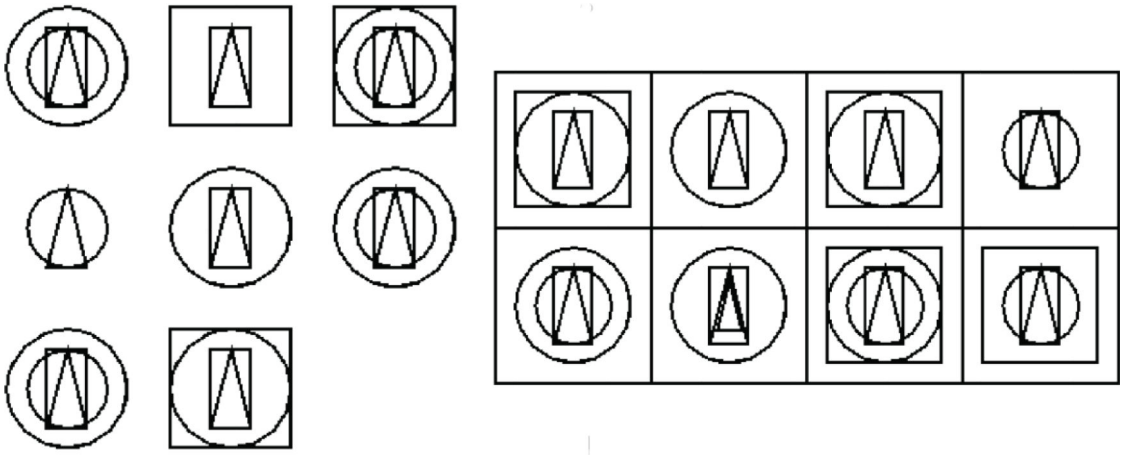

Figure 1:
Example of a logic problem generated by the Sandia Software. The 3×3 matrix on the left represents the problem with the lower right entry missing, and the 8 answer options are presented on the right.

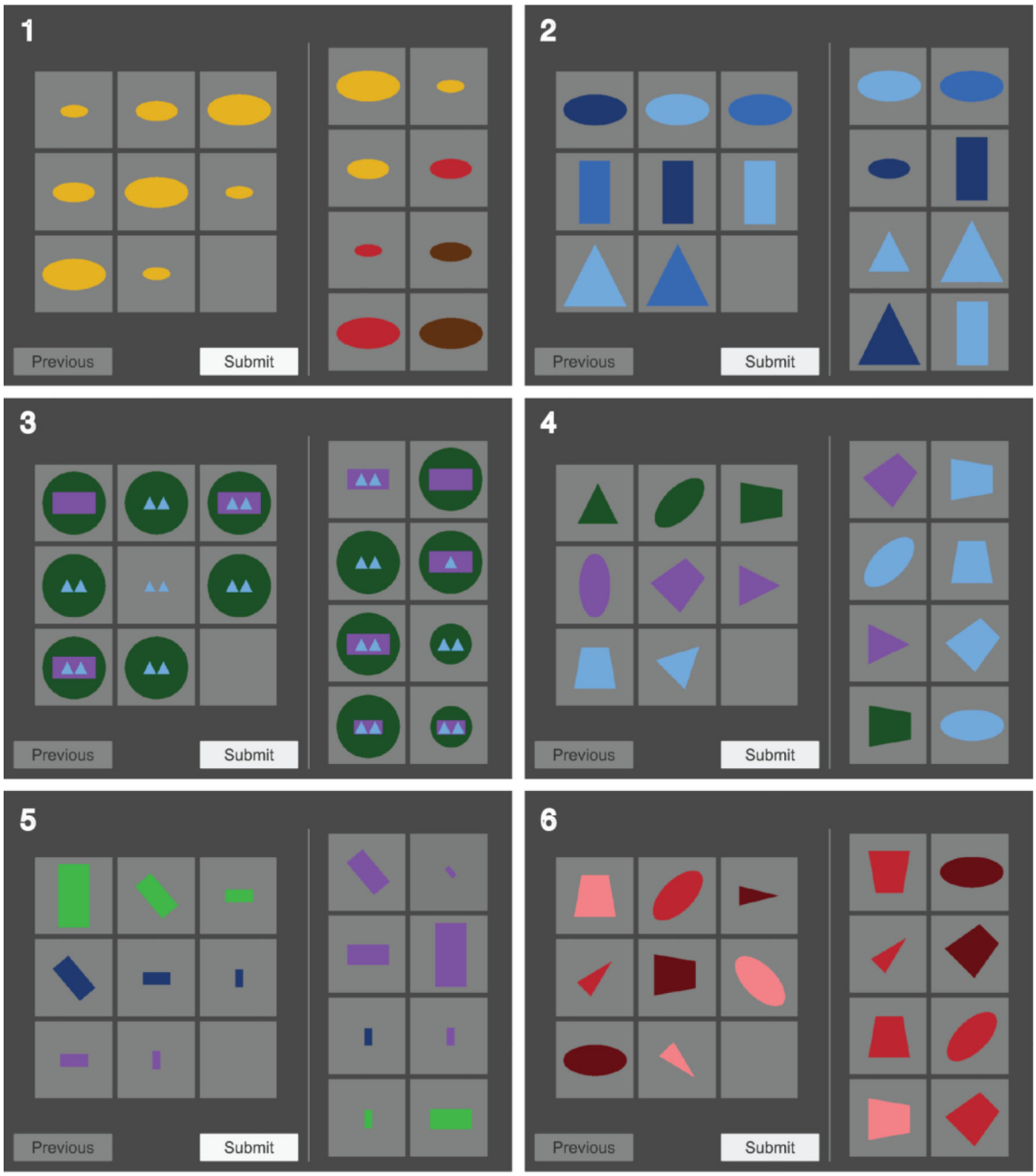

Figure 2:
Examples of UCMRT problems using the same structure as Sandia matrices, apart from the answer options which are presented vertically. All types of problems are shown in the practice section: (1) 1-relation problem, (2) 2-relation problem, (3) logic, (4) 3-relation with one transformation, (5) 3-relation with two transformations, and (6) 3-relation with three transformations.

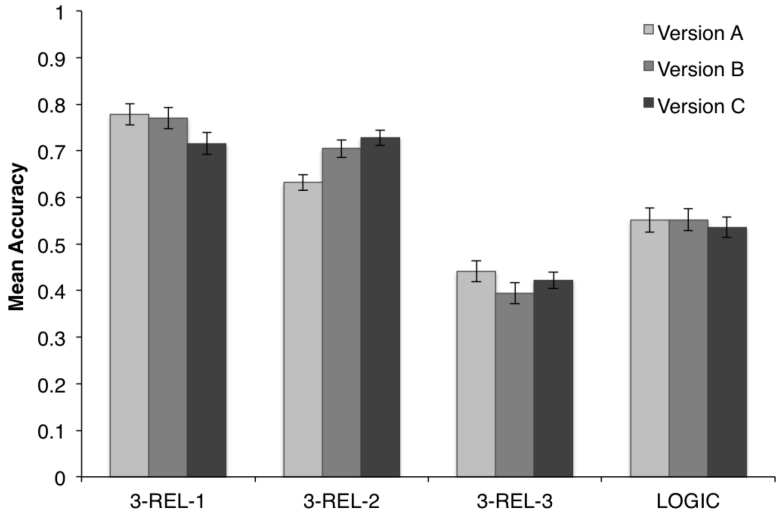

Figure 3:

Mean Accuracy based on problem type for A, B and C versions in participants who correctly solved 2-relation problems. 3-REL-1 = 3-relation with 1 transformation, 3-REL-2 = 3relation with 2 transformations, 3-REL-3 = 3-relation with 3 transformations.

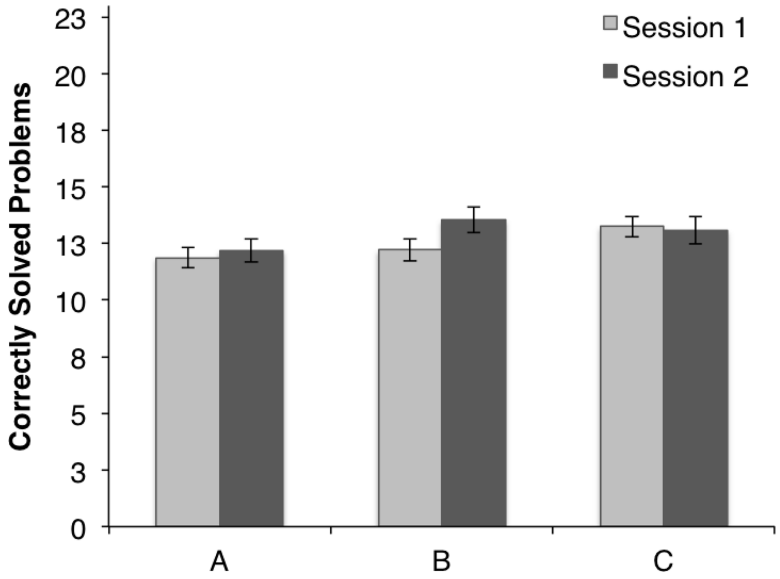

Figure 4:
Average number of correctly solved problems on alternate versions of UCMRT at two time points. Each participant completed 2 out of 3 alternate versions. Error bars = SEM.

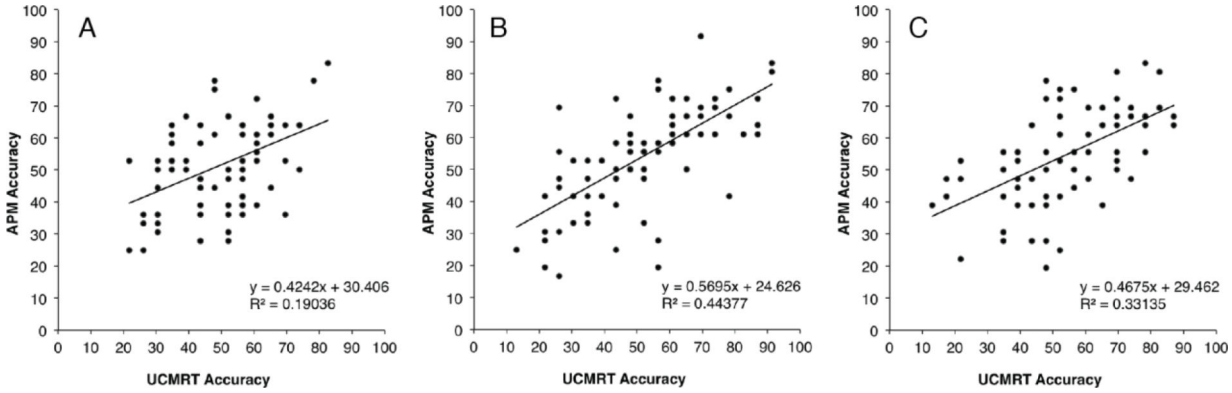

Figure 5:

Scatter plots illustrating the correlation between UCMRT and APM accuracy for the three groups of participants that solved A, B and C versions of UCMRT.

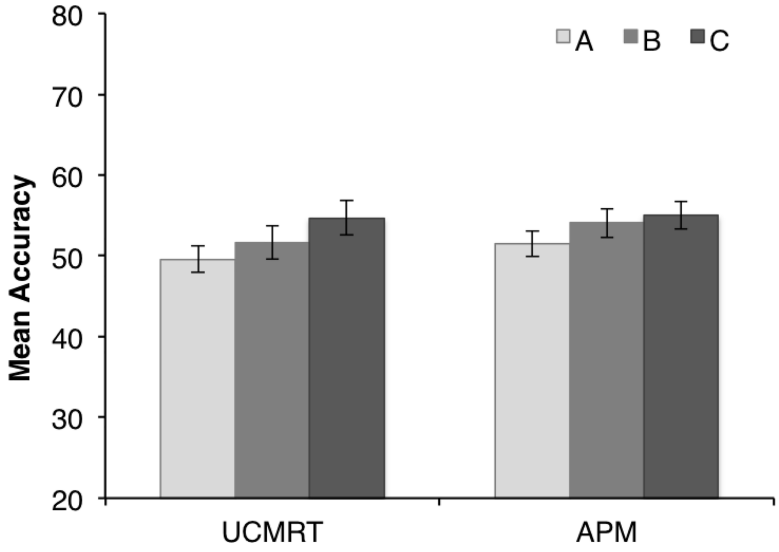

Figure 6:
Mean accuracy of UCMRT and APM in the three groups of participants that solved the alternate versions of UCMRT. Error bars = SEM.

Table 1:

Descriptive statistics for UCMRT scores (maximum = 23).

	Entire Sample				Subsample*		
	All Versions	**Version A**	**Version B**	**Version C**	**Version A**	**Version B**	**Version C**
Minimum	3	4	3	3	7	5	4
Maximum	22	20	22	22	20	22	22
Mean	12.55	12.12	12.58	12.98	13.92	14.17	14.22
Std. Error	.15	.24	.28	.26	.33	.34	.29
St. Deviation	4.02	3.72	4.32	3.94	3.39	3.89	3.49
Median	12	12	13	13	14	13	14.5
Variance	16.14	13.80	18.74	15.52	11.49	15.15	12.15
Skewness	.04	.15	.04	−.09	.01	.09	−.13
Kurtosis	−.52	−.67	−.47	−.48	−.88	−.52	−.42
Cronbach's α	.71	.66	.76	.72	.62	.74	.67
N	703	234	243	226	105	132	150

*Only including participants who correctly solved the first two problems of UCRMT.

Table 2:

Descriptive statistics for repeated testing on different versions of UCMRT.

	All Versions		Group 1		Group 2		Group 3	
	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2	Session 1	Session 2
Version	/	/	**A**	**B**	**B**	**C**	**C**	**A**
Minimum	3	2	4	2	3	3	4	4
Maximum	21	23	20	22	20	23	21	21
Mean	12.45	12.94	11.88	13.54	12.22	13.09	13.25	12.19
Std. Error	.27	.32	.47	.57	.49	.60	.45	.50
St. Deviation	3.89	4.63	3.83	4.70	4.00	4.97	3.76	4.17
Median	12	13	12	14	13	13	13	12
Variance	15.14	21.46	14.70	22.04	16.00	24.71	14.16	17.36
Skewness	−.06	−.22	.05	−.52	−.27	−.22	.11	.03
Kurtosis	−.31	−.47	−.63	−.28	.01	−.50	−.43	−.28
Cronbach's α	.69	.79	.68	.81	.73	.82	.70	.73
Correlation	$r = 0.62$, $p < 0.001$ $BF_{10} > 100$		$r = 0.68$, $p < 0.001$ $BF_{10} > 100$		$r = 0.64$, $p < 0.001$ $BF_{10} > 100$		$r = 0.62$, $p < 0.001$ $BF_{10} > 100$	
N	205		68		68		69	

Table 3:

Pearson correlation coefficients for performance on UCMRT (all versions), APM, Educational Records and Math.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. UCMRT	1 N = 703									
2. APM	.58** N = 233	1 N = 233								
3. College GPA	.13** N = 410	.16* N = 150	1 N = 410							
4. HS GPA	0.03 N = 330	−0.09 N = 113	.16** N = 330	1 N = 330						
5. SAT Math	.45** N = 314	.44** N = 110	.22** N = 313	0.10 N = 300	1 N = 314					
6. ACT Reading	.27** N = 172	0.04 N = 48	.18* N = 172	−0.02 N = 168	.53** N = 141	1 N = 172				
7. ACT Math	.35** N = 172	.35* N = 48	.29** N = 172	0.09 N = 168	.84** N = 141	.43** N = 172	1 N = 172			
8. ACT Writing	.29** N = 171	0.03 N =48	.26** N = 171	−0.02 N = 167	.62** N = 140	.82** N = 171	.57** N = 171	1 N = 171		
9. ACT Science	.35** N = 171	0.07 N = 48	.23** N = 171	0.08 N = 167	.68** N = 140	.60** N = 171	.67** N = 171	.64** N = 171	1 N = 171	
10. Math	.36** N = 483	.34** N = 224	.18** N = 273	0.08 N = 216	.55** N = 206	.34** N = 111	.61** N = 111	.47** N = 110	.47** N = 110	1 N = 483

* p < 0.05

** p <0.01

APM = Raven's Advanced Progressive Matrices, HS = High