Validation of a matrix reasoning task for mobile devices

Anja Pahor¹ · Trevor Stavropoulos¹ · Susanne M. Jaeggi² · Aaron R. Seitz¹

Published online: 26 October 2018

© Psychonomic Society, Inc. 2018

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 to 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 three alternate forms that works on tablets and other mobile devices and that is targeted at a high-ability population 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 and a math test; furthermore, we collected college records of academic ability and achievement. 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

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 a 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 min. 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 design. It is also 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 the test's reliability (Sefcek, Miller, & Figueredo, 2016). Matzen et al. (2010) addressed those issues by creating 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 the quantity of test items, the quality of the drawings presents certain limitations, especially if the items are 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 toward 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 are available for these item sets (ICAR Catalogue, 2017). 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 the limitations of current approaches in the field, we developed a modified matrix reasoning test (the University of California Matrix Reasoning Task [UCMRT]) based largely on matrix problems generated by Sandia Laboratories (Matzen et al., 2010), which

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-018-1152-2>) contains supplementary material, which is available to authorized users.

✉ Anja Pahor
anjap@ucr.edu

¹ University of California, Riverside, CA, USA

² University of California, Irvine, CA, USA

(1) only takes 12–15 min to complete, (2) has three 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 that on Raven's APM, as well as math scores, and furthermore, we examined the relationship between UCMRT performance and both scores from standardized proficiency tests and GPA.

Method

Participants

A total of 713 participants (mean age = 20.02 years, $SD = 2.74$; $N_{\text{Female}} = 494$, $N_{\text{Male}} = 201$, $N_{\text{Other}} = 18$) took part in at least one session of UCMRT over a period of 6 months. The participants were a sample of ethnically diverse university students (see Table S1 of the supplemental materials) recruited at the University of California, Riverside (UCR; $N = 353$), or the University of California, Irvine (UCI; $N = 360$), with normal or corrected-to-normal vision. All participants provided informed consent and received either course credit or monetary compensation for their participation. Most participants ($N = 676$) were asked for permission to obtain their educational records, which was granted by 416 of the 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 APM, allowing for analyses of convergent and external validity (see the Convergent Validity Tasks section for more information). Another subset of the participants ($N = 213$) performed alternate versions of the 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)

The 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 SPM (Matzen et al., 2010). However, we believe that the Sandia matrix problems are

limited in terms of their graphics: Overlapping items are difficult to discern, and some of the details are very small, making it difficult to judge shape and shading properties that are crucial to solving the problem (see Fig. 1). Hence, we redesigned the three versions of the task using larger, nonoverlapping stimuli that are not limited to grayscale, while keeping the same number and types of rules and the basic structure of the task (3×3 matrix with eight 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. Whereas the answer options are presented horizontally in the Sandia matrices (two rows, four columns), the answer options in UCMRT are presented vertically, to maximize the space available on mobile devices (four rows, two columns; see Fig. 2).

Each alternate form of UCMRT consists of two example problems, six 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 four out of six correct), the practice is repeated once with a different set of problems. The test portion begins with 2 two-relation problems, followed by a mixed order of 15 three-relation problems and 6 logic problems, with a time limit of 10 min. 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. The task ends when the participants submit the answers or when the 10-min limit is reached.

Matrix reasoning problems can be broken down by type, as defined by Matzen et al. (2010). The UCMRT contains two types of object transformation problems: two-relation problems (i.e., two rules govern the pattern of changes within a matrix; cf. Fig. 2.2) and three-relation problems (i.e., three rules make up the pattern of changes; cf. Figs. 2.4–2.6). The rules featured in these problems relate to shape, orientation, size, and number, as well as shading/color.³ The three-relation problems can be further distinguished on the basis of the number of diagonal or outward transformations (1–3) in the

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 the time spent on each problem, a functionality that is not available with paper-and-pencil versions of matrix problems.

³ Different shades and colors are characterized by different levels of luminance and should thus be distinguishable to participants with color vision deficiencies. All participants completed a color-naming test. Two participants with potential color vision deficiencies (green named as red or green named as yellow) performed within the normal range (15/23 and 13/23 correct).

Fig. 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 cell missing, and the eight answer options are presented on the right

Fig. 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) one-relation

problem, (2) two-relation problem, (3) logic, (4) three-relation with one transformation, (5) three-relation with two transformations, and (6) three-relation with three transformations

matrix. UCMRT also contains three types of logic problems (see Fig. 2.3): addition/conjunction (AND), disjunction (OR), and exclusive disjunction (XOR). Each alternate version consists of 2 two-relation problems, 3 three-relation problems with one transformation, 6 three-relation problems with two transformations, 6 three-relation problems with three transformations, and 6 logic problems. To control for context effects, the problems are not ordered by problem type (with the exception of the two-relation problems), nor are they specifically ordered in terms of difficulty—apart from the overall structure, in which two-relation and three-relation problems with fewer transformations are distributed at the beginning, whereas those that require more transformations are distributed toward 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.⁴

Implementation

UCMRT runs on Unity, which supports multiple platforms and can be utilized to make the test more available over time. Presently, UCMRT is supported by iOS and Android, and it can be released on other platforms per request. The app supports the 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 (i.e., text files are stored on the device itself), and in addition, they are logged on an Amazon-hosted server as long as a wireless Internet connection is maintained. The server-stored data can be made available to researchers upon request.

Convergent validity tasks

Two tasks were used to assess the convergent validity of UCMRT. The first task was Raven's APM 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 cell missing, and the participant has to select one out of eight 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 min. This time restriction was applied in order to approximate the conditions for the UCMRT and to avoid the ceiling performance often observed in our lab with longer or no time restrictions.⁵ The

second task was a tablet-based Math Test consisting of 21 word-based math problems with a time limit of 15 min (see Math Test in the supplemental materials), administered via Qualtrics software (Qualtrics, Provo, UT, <https://www.qualtrics.com>).

Educational data

For those participants who gave consent, college grade point average (GPA), high school GPA, and SAT and ACT scores were obtained from university records. GPA reflects overall performance during the academic year and is on a 4.0 scale. The SAT and ACT are standardized tests used for college 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 would be affected by recent major changes in the SAT sections, particularly for verbal measures (and furthermore, some scores were not scaled), and also given that the maximum SAT score changed from of 2,400 to 1,600 in 2016, we decided to exclude the SAT Verbal/Evidence-Based Reading and Writing and SAT Total measures from the analyses. Although certain changes were implemented on the SAT Math measure, its scale has remained the same (200–800), and therefore it was not excluded. For ACT, scores on the 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 presented together 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 to assess test–retest 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. These participants were divided randomly into three groups: One group completed Version A in Session 1 and Version B in Session 2; another 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, or C), one of the two versions of the Math Test (I or II), as well as the APM. The order of the two matrix reasoning tests (APM and UCMRT) was counterbalanced, and they were always separated by the Math Test.

⁴ For example, the rules that govern a given three-relation problem with three transformations are shape, shading, and orientation in Version A; shape, size, and number in Version B; and shape, orientation, and size in Version C.

⁵ Timed performance on APM is predictive of untimed performance on the same test (Frearson & Eysenck, 1986; Hamel & Schmittmann, 2006; Salthouse, 1993; Unsworth & Engle, 2005).

Table 1 Descriptive statistics for UCMRT scores (maximum = 23)

	Entire Sample				Subsample ^a		
	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
Std. 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

^a Only includes participants who correctly solved the first two problems of UCMRT.

Statistical analyses

SPSS Version 24 and JASP Version 0.9.0.1 (JASP Team, 2018) were used to analyze the data. Along with conventional null-hypothesis statistical testing, Bayes factors (BFs) were calculated using the default priors in JASP. BFs calculate the probability of the null (H_0) or alternative (H_1) hypothesis being true, given the present data. Thus, the results of our frequentist analyses were supplemented with BFs—specifically, BF_{10} , which grades the intensity of the evidence that the data provide for H_1 versus H_0 . A BF_{10} value between 1 and 3 is considered to be only anecdotal evidence for H_1 ; 3–10 indicate moderate evidence; 10–30, strong evidence; 30–100, very strong evidence; and > 100, extreme evidence (Lee & Wagenmakers, 2013; Wagenmakers et al., 2018).

Bonferroni correction was applied to post-hoc computed Student's t tests; note that the reported effect size, Cohen's d , does not correct for multiple comparisons.

Results

UCMRT performance and reliability

Ten outliers (1.4%) were removed from the sample: six individuals on the basis of number of responses (i.e., skipping at least 48% of the problems; $z_s = 4$), and four on the basis of number of correctly solved problems ($|z| \geq 2.5$). Descriptive statistics are presented in Table 1. There was no statistical difference in performance among the three versions of the UCMRT, as determined by a one-way analysis of variance (ANOVA) [$F(2, 700) = 2.70, p = .07, \eta^2 = .01$], which was further confirmed by a JZS Bayes factor ANOVA with default prior scales ($BF_{10} = 0.21$).

Although Bonferroni-corrected post-hoc tests showed no significant differences between the pairs of tests (A–B, $p = .63$; B–C, $p = .82$; and A–C, $p = .06$), Bayesian post-hoc tests revealed anecdotal evidence in favor of H_1 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: $\alpha = .71$; cf. Table 1). Similarly, Cronbach's α for the 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 .82 in our sample and .84 in other data sets (Arthur & Day, 1994; Bors & Stokes, 1998). For a short version consisting of 18 problems, α s of .64 (Unsworth, Redick, Lakey, & Young, 2010) and .79 (Sefcek et al., 2016) have been reported. Considering the relatively low number of problems and short testing time of the 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., the easy, two-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 [$F(2, 384) = 0.22, p = .80, \eta^2 = .001, BF_{10} = 0.04$]. 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: $p = 1, BF_{10,U} = 0.17$). Because they demonstrated an understanding of the rules that apply to the two-relation problems, this subsample may have been particularly suited to UCMRT testing. Nevertheless, as we demonstrated above, the entire sample also showed adequate performance.

Fig. 3 Mean accuracy based on problem type for the A, B, and C versions of UCMRT, in participants who correctly solved two-relation problems. 3-REL-1 = three-relation with one transformation, 3-REL-2 = three-relation with two transformations, 3-REL-3 = three-relation with three transformations

Problem type

Overall, accuracy decreased as the number of relations/transformations increased ($M_{2\text{-REL}} = 73.47\%$, $SE_{2\text{-REL}} = 1.21\%$; $M_{3\text{-REL-1}} = 68.94\%$, $SE_{3\text{-REL-1}} = 1.08\%$; $M_{3\text{-REL-2}} = 61.93\%$, $SE_{3\text{-REL-2}} = 0.89\%$; $M_{3\text{-REL-3}} = 38.48\%$, $SE_{3\text{-REL-3}} = 0.88\%$),⁶ which is consistent with the findings reported by Matzen et al. (2010) in a sample of undergraduate university students. For logic-based problems, the average accuracy across all versions was 50.88% ($SE = 1.00\%$), which is higher than that reported for the Sandia logic problems ($M = 37.9\%$, $SE = 3\%$). Some reasons for this may have been the improved visibility of details and shapes, which are nonoverlapping in UCMRT (see Fig. 2), as well as the inclusion of practice problems in our version. Figure 3 shows average accuracy based on problem type, presented separately for participants who completed the different versions of UCMRT and correctly solved the first two problems (whole-sample data are presented in Fig. S1, supplemental materials). One-way ANOVAs were used to determine whether performance on the three versions differed for each problem type. For three-relation problems with one transformation, performance on the three versions of the test was not statistically different [$F(2, 384) = 2.21$, $p = .11$, $BF_{10} = 0.23$]. For three-relation problems with two transformations, both types of analyses indicated a statistically significant difference in performance between the three versions [$F(2, 384) = 7.32$, $p < .05$, $BF_{10} = 19.48$]. Bonferroni-corrected post-hoc tests showed that performance on Version A was significantly lower than performance on Version B ($p = .02$, $BF_{10,U} = 10.31$) and on Version C ($p < .001$, $BF_{10,U} = 142.92$), but there was no significant difference between Versions B and C ($p = 1$, $BF_{10,U}$

$= 0.15$). For three-relation problems with three transformations [$F(2, 384) = 1.23$, $p = .29$, $BF_{10,U} = 0.09$] and for logic problems [$F(2, 384) = 0.16$, $p = .85$, $BF_{10,U} = 0.03$], no significant difference in performance was observed (see Fig. 3). Overall, performance on the subtypes of problems was well matched, with the exception of the three-relation problems with two transformations. This may not be problematic, since, as we demonstrated in the previous section, the final scores on the three versions were adequately balanced, and the participants using Version A also scored worse on the APM (see below).

Test-retest

Eight outliers (3.8%) were removed ($z_{\text{NoResp}} \geq 4$ and $z_{\text{Hits}} \geq 2.5$) from the data set, in which participants completed two sessions ($N = 213$). Descriptive statistics for the 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 .62 ($p < .001$, $BF_{10} > 100$). A paired-samples t test collapsed across versions showed that performance was not significantly different in the two sessions [$t(204) = -1.82$, $p = .07$, Cohen's $d^7 = 0.13$, $BF_{10} = 0.40$]. These results may be of use to future intervention studies trying to estimate improvements on the 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,

⁶ $N = 703$, collapsed across versions

⁷ Accounts for the correlation between pre- and posttest measures: $(\text{Mean}_{\text{Post}} - \text{Mean}_{\text{Pre}}) / \sqrt{(SD_{\text{Pre}}^2 + SD_{\text{Post}}^2 - 2r_{\text{PrePost}} * SD_{\text{Pre}} * SD_{\text{Post}})}$.

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
Std. 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	$r = .62, p < .001, BF_{10} > 100$		$r = .68, p < .001, BF_{10} > 100$		$r = .64, p < .001, BF_{10} > 100$		$r = .62, p < .001, BF_{10} > 100$	
<i>N</i>	205		68		68		69	

& Perrig, 2017), and some even reporting worse performance in the second session than in 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 [$t(67) = 3.91, p < .001$, Cohen's $d = 0.47$, $BF_{10} = 102.3$], and a similar trend was observed for Group 2 [$t(67) = 1.83, p = .07$, 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 [$t(68) = -2.54, p = .01$, Cohen's $d = -0.31$, $BF_{10} = 2.57$].

Although we were not able to test all possible pairs of alternate forms, the results shown in Fig. 4 suggest that different groups of participants showed similar performance on the same version of the test in different sessions. Independent-samples *t* tests showed no significant differences in performance on the three versions of UCMRT at different points in time [Version A: $t(135) = -0.45, p = .66$, Cohen's $d = -0.08$, $BF_{10} = 0.20$; Version B: $t(134) = -1.77, p = .08$, Cohen's $d = -0.30$, $BF_{10} = 0.76$; Version C: $t(135) = 0.21, p = .83$, Cohen's $d = 0.04$, $BF_{10} = 0.19$].

Convergent and external validity

Raven's APM Two additional outliers were removed on the basis of performance on Raven's APM Set II ($z \geq 2.5$). Performance on UCMRT was significantly correlated with performance on APM, both in the total sample ($N = 233, r = .58, p < .001, BF_{10} > 100$) and in the three subgroups of participants (Group A: $N = 79, r = .44, p < .001, BF_{10} >$

100; Group B: $N = 80, r = .67, p < .001, BF_{10} > 100$; Group C: $N = 74, r = .58, p < .001, BF_{10} > 100$) (see Fig. 5). Matzen et al. (2010) reported a correlation of .69 between accuracy for the Sandia matrices and Raven's SPM. However, the authors only included in the analysis a subset of the SPMs 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 levels of the two matrix reasoning tests are 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 [$t(232) = -1.59, p = .11$, Cohen's $d = 0.10$, $BF_{10} = 0.25$]. As can be seen in Fig. 6, the APM accuracy for the three groups follows a pattern similar to that of UCMRT accuracy, with Group A performing slightly worse than Groups B and C, which may indicate inherent group differences in reasoning ability, suggesting that the 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 (see the 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 was determined by an independent-samples *t* test [$t(485) = -1.49, p = .14, BF_{10} = 0.29$]; hence, the data were collapsed. As expected, performance on the Math Test showed strong correlations with SAT Math ($r = .55, p < .001, BF_{10} > 100$) and ACT Math ($r = .61, p < .001, BF_{10} > 100$). Furthermore, performance on

Fig. 4 Average numbers of correctly solved problems on the alternate versions of UCMRT at two time points. Each participant completed two out of three alternate versions. Error bars = SEMs

the Math Test correlated significantly with performance on both UCMRT ($r = .36, p < .001, BF_{10} > 100$) and APM ($r = .34, p < .001, BF_{10} > 100$), providing further evidence for the validity of UCMRT (see Table 3).

Tests of academic proficiency We correlated accuracy on UCMRT and APM with academic proficiency, where those data were available (see Table 3). Performance on both UCMRT and APM showed small but significant correlations with current college GPA ($r_{\text{UCMRT}} = .13, p = .007, BF_{10} = 2.19$; $r_{\text{APM}} = .16, p = .045, BF_{10} = 0.75$). This is line with previous research in which performance on APM was somewhat weakly related to both first-term ($r = .19, p < .05$) and second-term ($r = .18, p < .05$) college GPA (Coyle & Pillow, 2008). As a side note, the only measure that correlated significantly with high school GPA was college GPA ($r = .16, p = .003, BF_{10} = 5.03$).⁸ Although 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 the ACT and SAT subtests included in the analysis.

Performance on both UCMRT and APM also correlated with SAT Math ($r_{\text{UCMRT}} = .45, p < .001, BF_{10} > 100$; $r_{\text{APM}} = .44, p < .001, BF_{10} > 100$) and ACT Math ($r_{\text{UCMRT}} = .35, p < .001, BF_{10} > 100$; $r_{\text{APM}} = .35, p = .02, BF_{10} = 3.26$), 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 ($r_{\text{UCMRT}} = .27, p < .001, BF_{10} = 52.53$; $r_{\text{APM}} = .04, p = .790, BF_{10} = 0.19$), ACT Writing ($r_{\text{UCMRT}} = .29, p < .001, BF_{10} > 100$; $r_{\text{APM}} = .03, p = .862, BF_{10} = 0.18$), and ACT Science ($r_{\text{UCMRT}} = .35, p < .001, BF_{10} > 100$; $r_{\text{APM}} = .07, p = .639, BF_{10} = 0.20$), suggesting that UCMRT is predictive of global cognitive function. It should be noted that the sample size for APM was smaller; however, on the basis of the correlation coefficients (.03–.07) and visual inspection of scatterplots (cf. Fig. S2, supplemental materials), a stronger relation with ACT subtests would not be expected even if the sample size were 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 in line with previous work.

Individual versions of UCMRT showed similar relations, albeit the evidence was limited by a smaller sample size in certain cases (cf. Tables S4–S6, supplemental materials). Version A showed significant correlations with SAT Math ($r = .48, p < .001, BF_{10} > 100$), ACT Writing ($r = .30, p = .028, BF_{10} = 1.80$), and ACT Science ($r = .40, p = .003, BF_{10} = 12.51$). Version B significantly correlated with SAT Math ($r = .46, p < .001, BF_{10} > 100$), ACT Math ($r = .40, p = .003, BF_{10} = 13.18$), and ACT Science ($r =$

⁸ Note that the correlation in the present study may have been lower because we obtained records of current GPA, rather than just first-year GPA, resulting in a mixture of first-year and higher-year GPAs.

Fig. 5 Scatterplots illustrating the correlations between UCMRT and APM accuracy for the three groups of participants who solved the A, B, and C versions of UCMRT

.28, $p = .044$, $BF_{10} = 1.22$). Version C was related to SAT Math ($r = .40$, $p < .001$, $BF_{10} > 100$), ACT Reading ($r = .37$, $p = .002$, $BF_{10} = 14.46$), ACT Math ($r = .40$, $p < .001$, $BF_{10} = 31.03$), ACT Writing ($r = .33$, $p = .007$, $BF_{10} = 5.37$), and ACT Science ($r = .34$, $p = .006$, $BF_{10} = 6.17$). In summary, all three versions correlated with APM (Fig. 5), the Math Test, and 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 a proxy for 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 that in 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 the versions were comparable. Convergent validity was established by comparing UCMRT to Raven's APM. In addition, performance on UCMRT correlated with a newly developed Math Test, college GPA, and scores obtained on college admission tests, showing correlations similar to those for 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 showed significant correlations only with ACT Math.

Matrix reasoning problems have been the hallmark of tests estimating nonverbal fluid intelligence for over eight decades

Fig. 6 Mean accuracy on the UCMRT and APM in the three groups of participants who solved the alternate versions of UCMRT. Error bars = SEMs

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									
	<i>N</i> = 703									
2. APM	.58**	1								
	<i>N</i> = 233	<i>N</i> = 233								
3. College GPA	.13**	.16*	1							
	<i>N</i> = 410	<i>N</i> = 150	<i>N</i> = 410							
4. HS GPA	.03	–.09	.16**	1						
	<i>N</i> = 330	<i>N</i> = 113	<i>N</i> = 330	<i>N</i> = 330						
5. SAT Math	.45**	.44**	.22**	.10	1					
	<i>N</i> = 314	<i>N</i> = 110	<i>N</i> = 313	<i>N</i> = 300	<i>N</i> = 314					
6. ACT Reading	.27**	.04	.18*	–.02	.53**	1				
	<i>N</i> = 172	<i>N</i> = 48	<i>N</i> = 172	<i>N</i> = 168	<i>N</i> = 141	<i>N</i> = 172				
7. ACT Math	.35**	.35**	.29**	.09	.84**	.43**	1			
	<i>N</i> = 172	<i>N</i> = 48	<i>N</i> = 172	<i>N</i> = 168	<i>N</i> = 141	<i>N</i> = 172	<i>N</i> = 172			
8. ACT Writing	.29**	.03	.26**	–.02	.62**	.82**	.57**	1		
	<i>N</i> = 171	<i>N</i> = 48	<i>N</i> = 171	<i>N</i> = 167	<i>N</i> = 140	<i>N</i> = 171	<i>N</i> = 171	<i>N</i> = 171		
9. ACT Science	.35**	.07	.23**	.08	.68**	.60**	.67**	.64**	1	
	<i>N</i> = 171	<i>N</i> = 48	<i>N</i> = 171	<i>N</i> = 167	<i>N</i> = 140	<i>N</i> = 171	<i>N</i> = 171	<i>N</i> = 171	<i>N</i> = 171	
10. Math	.36**	.34**	.18**	.08	.55**	.34**	.61**	.47**	.47**	1
	<i>N</i> = 483	<i>N</i> = 224	<i>N</i> = 273	<i>N</i> = 216	<i>N</i> = 206	<i>N</i> = 111	<i>N</i> = 111	<i>N</i> = 110	<i>N</i> = 110	<i>N</i> = 483

* $p < .05$, ** $p < .01$, APM = Raven's Advanced Progressive Matrices, HS = high school

(Raven, 1938). Matrices adapted to different ability levels have been developed over time, yet the Raven matrices have 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, when researchers are faced with a limited set of test items, many of which can be found online, it is time to expand the set and adapt it for modern technology. Although there have been certain efforts in this direction,⁹ UCMRT is unique in that it consists of multiple validated versions that will be freely available to researchers. By making these available, we hope to be able to collect large data sets and 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. Although alternate versions may show small differences in terms of accuracy, in our case such differences could have been driven by inherent group differences, such issues can be addressed with counterbalancing when test repetition is needed.

Perhaps the greatest advantages of UCMRT are its short administration time and the fact that it is self-administrable, which allows for remote testing. Log files instantly provide the numbers of problems that were solved correctly, solved incorrectly, or skipped, data that will be easily understandable for researchers, clinicians, and users alike. More detailed log

files are also available that provide insight into problem-solving patterns and reaction times, functionality that is not available with 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 nonverbal problem solving that is predictive of academic proficiency and could serve as a proxy for fluid intelligence. Moreover, it can be used to differentiate among people at the high end of intellectual ability, akin to Raven's APM. As compared to APM, the UCMRT is shorter (10 min, excluding practice), offers three parallel test versions, and can be used on iOS and Android devices.

Author note This work has been supported by the National Institute of Mental Health (Grant No. 1R01MH111742) to A.R.S. and S.M.J.; in addition, S.M.J. is supported by the National Institute on Aging (Grant No. 1K02AG054665). S.M.J. has an indirect financial interest with the MIND Research Institute, whose interest is related to this work. We thank David Arellano, Austin Moon, and research assistants at UCR and UCI for data collection. We also thank Laura E.

⁹ e.g., <https://www.cambridgebrainsciences.com>, <https://icar-project.com>

Matzen for providing access to the matrices generated by Sandia National Laboratories and Gabriella Del Greco for helping with the design of the interface.

References

- Ackerman, P. L., & 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, 163–181. <https://doi.org/10.1037/a0015719>
- Arthur, W., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, 54, 394–403. <https://doi.org/10.1177/0013164494054002013>
- Arthur, W., & Woehr, D. J. (1993). A confirmatory factor analytic study examining the dimensionality of the raven's advanced progressive matrices. *Educational and Psychological Measurement*, 53, 471–478. <https://doi.org/10.1177/0013164493053002016>
- Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement*, 58, 382–398. <https://doi.org/10.1177/0013164498058003002>
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, 1–22. <https://doi.org/10.1037/h0046743>
- Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (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, e0177707. <https://doi.org/10.1371/journal.pone.0177707>
- Colom, R., Quiroga, M. Á., Shih, P. C., Martínez, K., Burgaleta, M., Martínez-Molina, A., . . . Ramírez, I. (2010). Improvement in working memory is not related to increased intelligence scores. *Intelligence*, 38, 497–505. <https://doi.org/10.1016/j.intell.2010.06.008>
- Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., . . . Jaeggi, S. M. (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, 712–727. <https://doi.org/10.1016/j.intell.2013.09.002>
- Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, 43, 52–64. <https://doi.org/10.1016/j.intell.2014.01.004>
- Coyte, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing *g*. *Intelligence*, 36, 719–729. <https://doi.org/10.1016/j.intell.2008.05.001>
- Frearson, W., & Eysenck, H. J. (1986). Intelligence, reaction time (RT) and a new “odd-man-out” RT paradigm. *Personality and Individual Differences*, 7, 807–817. [https://doi.org/10.1016/0191-8869\(86\)90079-6](https://doi.org/10.1016/0191-8869(86)90079-6)
- Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, 39, 233–243. <https://doi.org/10.1016/j.intell.2011.02.009>
- Goff, M., & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology*, 84, 537–552. <https://doi.org/10.1037/0022-0663.84.4.537>
- Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, 66, 1039–1046. <https://doi.org/10.1177/0013164406288169>
- Hogrefe, A. B., Studer-Luethi, B., Kodzhabashev, S., & Perrig, W. J. (2017). Mechanisms underlying *N*-back training: Response consistency during training influences training outcome. *Journal of Cognitive Enhancement*, 1, 406–418. <https://doi.org/10.1007/s41465-017-0042-3>
- Hossiep, R., Turck, D., & Hasella, M. (1999). Bochumer Matrizentest. BOMAT advanced.
- ICAR Catalogue. (2017). Version 1.0, 06 I 17. Retrieved August 19, 2018, from https://icar-project.com/ICAR_Catalogue.pdf
- Jaeggi, S. M., Buschkuhl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition*, 42, 464–480. <https://doi.org/10.3758/s13421-013-0364-z>
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between *n*-back performance and matrix reasoning—Implications for training and transfer. *Intelligence*, 38, 625–635. <https://doi.org/10.1016/j.intell.2010.09.001>
- JASP Team. (2018). JASP (Version 0.9.0.1) [Computer software]. Retrieved from <https://jasp-stats.org/download/>
- Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, 36, 153–160. <https://doi.org/10.1016/j.intell.2007.03.005>
- Koretz, D., Yu, C., Mbekeani, P. P., 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, 233285841667060. <https://doi.org/10.1177/2332858416670601>
- Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian cognitive modeling: A practical course. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9781139087759>
- Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, 42, 525–541. <https://doi.org/10.3758/BRM.42.2.525>
- Raven, J. C. (1938). Progressive matrices: A perceptual test of intelligence. London: H.K. Lewis.
- Raven, J. C., Court, J. H., & Raven, J. (1998). Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 4. Advanced Progressive Matrices, Sets I & II. Oxford, UK: Oxford Psychologists Press.
- Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., . . . Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, 142, 359–379. <https://doi.org/10.1037/a0029082>
- Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35, 83–92. <https://doi.org/10.1016/j.intell.2006.05.004>
- Salthouse, T. A. (1993). Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology*, 84(Part 2), 171–199.
- Sefceck, J. A., Miller, G. F., & Figueredo, A. J. (2016). Development and validation of an 18-item medium form of the Ravens Advanced Progressive Matrices. *SAGE Open*, 6, 215824401665191. <https://doi.org/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, 100–107. <https://doi.org/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, 67–81. <https://doi.org/10.1016/j.intell.2004.08.003>
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37, 498–505. <https://doi.org/10.3758/BF03192720>
- Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence*, 38, 111–122. <https://doi.org/10.1016/j.intell.2009.08.002>
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Morey, R. D. (2018). Bayesian inference for psychology: Part II. Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76. <https://doi.org/10.3758/s13423-017-1323-7>
- Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College Performance and retention: A meta-analysis of the predictive validities of ACT® scores, high school grades, and SES. *Educational Assessment*, 20, 23–45. <https://doi.org/10.1080/10627197.2015.997614>