Evaluating untimed and timed abridged versions of Raven's Advanced Progressive Matrices

Antoinette Poulton[1], Kathleen Rutherford[1], Sarah Boothe[1], Madeleine Brygel[1], Alice Crole[1], Loren Richard Bruns Jr.[2], Richard O. Sinnott[2], Robert Hester[1]

[1]Melbourne School of Psychological Sciences, University of Melbourne, Parkville 3010, VIC, Australia

[2]Computing and Information Systems, University of Melbourne, Parkville 3010, VIC, Australia

*Corresponding author: Melbourne School of Psychological Sciences, University of Melbourne, Parkville 3010, VIC, Australia. Tel.: +61 3 8344 6377. Fax: +61 3 9347 6618. ORCiD: 0000-0001-9291-6128. Twitter: https://twitter.com/antwo

Email: antoinette.poulton@unimelb.edu.au; poultonantoinette@gmail.com

Abstract

Introduction: Raven's Advanced Progressive Matrices (APM) are frequently utilised in clinical and experimental settings to index intellectual capacity. As the APM is a relatively long assessment, abridged versions of the test have been proposed. The psychometric properties of an untimed 12-item APM have received some consideration in the literature, but validity explorations have been limited. Moreover, both reliability and validity of a timed 12-item APM have not previously been examined.

Method: We considered the psychometric properties of untimed (Study 1; $N = 608$; $M_{age} = 27.89$, $SD = 11.68$) and timed (Study 2; $N = 479$; $M_{age} = 20.93$, $SD = 3.12$) versions of a brief online 12-item form of the APM.

Results: Confirmatory factor analyses established both versions of the tests are unidimensional. Item response theory analyses revealed that, in each case, the 12 items are characterised by distinct differences in difficulty, discrimination, and guessing. Differential item functioning showed few male/female or native English/non-native English performance differences. Test-retest reliability was .65 (Study 1) to .69 (Study 2). Both tests had medium to large correlations with the Wechsler Abbreviated Scale of Intelligence (2nd ed.) Perceptual Reasoning Index ($r = .50$, Study 1; $r = .56$, Study 2) and Full-Scale IQ ($r = .34$, Study 1; $r = .41$, Study 2).

Conclusion: In sum, results suggest both untimed and timed online versions of the brief APM are psychometrically sound. As test duration was found to be highly variable for the untimed version, the timed form might be a more suitable choice when it likely to form part of a longer battery of tests. Nonetheless, classical test and item response theory analyses, plus validity considerations, suggest the untimed version might be the superior abridged form.

Keywords: Raven's Advanced Progressive Matrices, online administration, Wechsler Intelligence Scales, validity, classical test theory, item response theory

Introduction

Raven's Advanced Progressive Matrices (APM: Raven, Raven, & Court, 1998) are widely used in clinical and experimental contexts to assess intellectual capacity (Evers et al., 2012). While these matrices offer advantages, especially in terms of validity and reliability, the APM is a relatively long assessment. This potentially impacts participant burden, particularly when it is incorporated into a longer battery of tests (Tatel et al., 2022). Abridged forms of the APM have been proposed, though the psychometric properties of these versions – especially when offered online – have not been extensively examined. With growing numbers of studies in psychology reliant – wholly or partially – on online presentation (Gosling & Mason, 2015; Sargis et al., 2013; Sassenberg & Ditrich, 2019), finding concise measures that facilitate remote but valid assessment of intellectual ability are increasingly important. With the advent of COVID-19, in-person testing is frequently unavailable due to lockdowns and associated restrictions, or it is hampered by the need to implement social distancing, cleaning/sanitising, and personal protective equipment protocols. This too has highlighted the need for succinct yet valid online assessment. In this paper, we thus aim to explore the psychometric properties of two brief online versions of the APM.

When undertaking the APM, individuals are presented with a series of increasingly complex abstract patterns, each with a missing element, and they are required to identify the missing component from eight choices (Raven et al., 1998). The test is unidimensional and has both high internal consistency (>.81) and split-half reliability (>.90; Alderton & Larson, 1990; Arthur & Woehr, 1993; Bors & Stokes, 1998; Bors & Vigneau, 2003; Rushton, Skuy, & Bons, 2004; Waschl, Nettelbeck, Jackson, & Burns, 2016). The APM has been found to correlate significantly with Wechsler Adult Intelligence Scale full scale (.55-.69), verbal (.42), and performance IQ (.55; McLaurin, Jenkins, Farrar, & Rumore, 1973; Paul, 1986). Norms are available for both timed and untimed administration (Raven et al., 1998). Test-retest reliability over a 45-day interval has been reported at .83 (Bors & Forrin, 1995). Research has found the matrices to be particularly useful when assessing individuals with

limited English proficiency (Mills & Tissot, 1995). Additionally, no significant differences have been found between scores derived from paper-and-pencil or computer-based versions of the test (Arce-Ferrer & Martinez Guzman, 2009; Williams & McCord, 2006).

A relatively long test, individuals generally take upward of 40 minutes to complete the full set of 36 APM matrices. This limits its utility in many situations where the test is being used as part of a larger time-consuming battery (Tatel et al., 2022). Several abridged versions of the APM have thus been proposed (e.g., Ablard & Mills, 1996; Bors & Stokes, 1998; Chiesi, Ciancaleoni, Galli, & Primi, 2012). The most frequently cited of these is one suggested by Arthur and Day (1994). They divided the 36 matrices from the original measure into 12 groups of three items – each triad contained items of similar difficulty – and then selected the item in each group with the highest item-total correlation; this yielded a brief 12-item form of the APM (APM-12) that generally preserved progressive item difficulty (conceived as percent correct or percent error in the literature) and reduced administration time to approximately 15 minutes (Arthur et al., 1999; Arthur & Day, 1994; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). Correlations between the APM and APM-12 have been reported in the range .66-.72 (Arthur et al., 1999; Arthur & Day, 1994).

Psychometric investigations of the (pen and paper) APM-12 have determined it has acceptable internal consistency (.61-.69), while confirmatory factor analysis (CFA) suggests a single factor best describes its structure (Ablard & Mills, 1996; Arthur et al., 1999; Arthur & Day, 1994; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). Item response theory (IRT) analyses indicate the 3-parameter (difficulty, discrimination, and guessing) model has good fit (Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). The APM-12 has also been found to be metrically equivalent across age (≥14 years), gender, and country in IRT analyses (Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). Test-retest reliability (7-10 day intervals) has been reported as .75-.76 (Arthur et al., 1999; Arthur & Day, 1994). While the APM-12 has been validated against several tests of intellectual capacity (e.g., Wesman Personnel

Classification Test, Wonderlic Personnel Test, and Bennett Mechanical Comprehension Test; Arthur & Day, 1994; Arthur et al., 1999; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012), concurrent validity with the more widely utilised Wechsler tests has not been established.

The aim of the current project was to assess the psychometric properties of an untimed online-administered APM-12 (henceforth referred to as APM-12U) and validate it against the Wechsler Abbreviated Scale of Intelligence (2nd ed.; WASI-II). The WASI-II was chosen as it is one of the most recently published of the Wechsler suite of tests and, unlike the WAIS-IV or Wechsler Intelligence Scale for Children, can be used to assess the IQ of individuals across a wide range of ages (6 to 60; Wechsler, 2011). We additionally sought to investigate both reliability and validity of a timed online version of this test, the APM-12T. While the APM instruction manual allows for untimed or timed administration (Raven et al., 1998), the psychometrics of the APM-12T have, to our knowledge, never been investigated. We also anticipated time to complete the APM-12U might be highly variable when administered online and in the absence of the researcher. The APM-12T might thus be preferrable when it forms part of an extensive battery of assessments. Limiting the amount of time participants can spend on any one component of a battery of tests reduces participant burden and thus increases the likelihood all sections of the battery will be completed. The results of two separate studies are presented below.

Materials and Method

Participants

Study 1 consisted of 608 individuals who completed the APM-12U. A number of these participants (n = 90) accepted an invitation to take part in the test-retest reliability component of this study. A further sub-sample also completed the WASI-II (n = 87). Study 2 comprised 479 individuals who undertook the APM-12T. A number of these (n = 144) accepted an invitation to take part in the test-retest reliability component of this study. A

further sub-sample also completed the WASI-II (n = 42). Data pertaining to each study is available via the Open Science Framework (Poulton, 2021).

Participants in both studies were recruited through adverts posted in and around the University of Melbourne as well as via researcher networks and social media posts. Participants under the age of 13 years were excluded. An ability to answer demographic questions was considered evidence of sufficient English language fluency. The University of Melbourne Human Ethics Committee approved the study in accordance with the standards for ethical research of the National Health and Medical Research Council. All demographic details are listed in Table 1. **[Table 1 near here]**

Procedure

After reading a plain language statement and providing informed consent, participants answered an online researcher-devised demographic survey. They then attempted an online 12-item form of the APM. At the end of the test, participants were provided with their raw score. Individuals who accepted the emailed invitation to take part in the test-retest component of each study were required to activate a link so they could attempt to solve each matrix a second time. Participants who undertook the WASI-II were required to visit the laboratory so a trained research assistant could administer these sections of the experiment. All participants went into a draw to win a $200 gift voucher; those who completed additional study components were provided with a corresponding number of bonus draw entries.

Measures

APM-12U/APM-12T

Both the APM-12U and APM-12T are comprised of a practice item (matrix 1 of APM-Set I) plus matrices 1, 4, 8, 11, 15, 18, 21, 23, 25, 30, 31, and 35 of APM-Set II (Arthur & Day, 1994). Participants can take as much time as required to solve each matrix when completing the APM-12U, whereas they are given 60 seconds to solve each matrix when undertaking the APM-12T. In determining timing parameters for the APM-12T, we sought to

give participants less time than the average time used to complete the APM-12U (15 minutes). We expected this would challenge participants and thus reduce the number reaching ceiling. This is also in keeping with recent research suggesting, that for timed administration, anything between 60-79 seconds per item can be considered "standard" (Tatel et al., 2022). As per the APM's instructional manual, participants are informed the APM-12U/APM-12T assesses perception and clear thinking (Raven et al., 1998). They received correct/incorrect plus explanatory feedback on the practice matrix before they were permitted to attempt the 12 items. APM-12U/APM-12T test scores were operationalized as the sum of correct items.

WASI-II

The WASI-II is an individually administered test of intelligence (Wechsler, 2011). It comprises four subtests: Vocabulary, Block Design, Similarities, and Matrix Reasoning. Administration time is around 30 minutes. A Verbal Comprehension Index is derived from the Vocabulary and Similarities subtests, while a Perceptual Reasoning Index is calculated from the Block Design and Matrix Reasoning subtests. All four subtests contribute to Full-Scale IQ (FSIQ-4). The WASI-II has been found to correlate .91 with the Wechsler Adult Intelligence Scale-IV (Wechsler, 2011).

Data Analyses

Average APM-12U and APM-12T scores were computed for the full samples and for sub-samples undertaking test-retest and validity components of each study. Regarding Study 1 APM-12U data, standardised values for skewness (-0.45, *SE* = .10) and kurtosis (-0.45, *SE* = .20) were $Z_{skewness} = 4.50$ and $Z_{kurtosis} = 2.25$, indicating significant skew and kurtosis. However, this is common when the sample size is large, as such samples give rise to small standard errors. Sample sizes greater than 200 warrant examination of the probability-probability (P-P) plot to determine normality (Tabachnick & Fidell, 2013). These plots revealed no obvious systemic deviation from the straight line, suggesting the data was

normally distributed. Regarding Study 2 APM-12T data, standardised values for skewness (-0.15, *SE* = .11) and kurtosis (-0.27, *SE* = .22) were $Z_{skewness}$ = 1.36 and $Z_{kurtosis}$ = 1.22, indicating no significant skew or kurtosis (at $p < .05$). Differences in APM-12U and APM-12T score as a function of male/female and native English/non-native English speaking were examined using independent *t*-tests, while correlations were utilised to investigate the relationship between score and age, years of education, and time taken to complete the test. Progression item difficulty was conceived as percent error. Reliability – specifically, internal consistency and test-retest – was examined. Prior to commencing validity analyses, differences between the WASI-II verbal comprehension, perceptual reasoning, and FSIQ-4 scores of male/female and native English/non-native English speaking were examined using independent *t*-tests. Validity was assessed using correlational analyses. Effect sizes for *t*-tests were computed using Cohen's *d*; they were interpreted according to Cohen's guidelines: 0.20 = small, 0.50 = moderate, and 0.80 = large effect (Cohen, 1988). Confidence intervals (95% CIs) were calculated for *t*-test and correlational results.

After preliminary exploratory analyses (see supplementary material), dimensionality was assessed via CFA (Mplus 8.1) utilizing the weighted least squares means and variance adjusted estimation method, which is advocated when variables are categorical (Muthén & Muthén, 2010). Comparative fit indices (CFI) greater than 0.95, root mean square error of approximation (RMSEA) values not more than 0.06, and standardized root mean square residual (SRMR) indices less than 0.08 are indicative of acceptable fit (Schmitt, 2011).

IRT analyses involved first determining the most appropriate model. There has been debate surrounding the suitability of models in IRT investigations of Raven's Progressive Matrices. A three-parameter model (3-PL) – comprising difficulty (*b*), discrimination (*a*), and guessing (*c*) – is advocated by some, as the multiple-choice nature of responses can engender guessing (Gallini, 1983; Harris et al., 2020), while others suggest that, given there are eight response choices, guessing is irrelevant and opt for a two-parameter model (2-PL; Çikrikçi-

Demirtasli, 2002). Recent literature suggests a 3-PL model is most appropriate for tests that have a multiple choice format coded as correct/incorrect, though a comparison of models is still recommended (Harris et al., 2020; Tay et al., 2015). Thus, in keeping with previous studies of the APM/APM-12 (Chiesi, Ciancaleoni, Galli, & Primi, 2012; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012; Kpolovie & Emekene, 2016), we examined the fit of each model.

Both 2-PL and 3-PL IRT analyses (Xcalibre 4.2.2.0) utilizing marginal maximum likelihood estimation were performed (Wright & Linacre, 1994). Model fit was assessed by inspecting item fit statistics (*z* residuals) and associated *p*-values, upper/lower *b*-, *a*-, and *c*-parameter bounds, and by determining whether there was a significant difference between models. *P*-values of item fit statistics should be non-significant (Guyer & Thompson, 2014; Maydeu-Olivares, 2013). The *b*-parameter, which indicates item difficulty and corresponds to the location where the probability of endorsing a correct response equates to 50%, generally falls between -3.00 and 3.00, while the *a*-parameter, which refers to the extent to which an item discriminates between participants, is usually between 0.30 and 4.00 (Guyer & Thompson, 2014). The *c*-parameter captures the probability of a correct answer, from someone of low cognitive ability, due to guessing (Harris et al., 2020). Values less then 0.35 are suggested for the *c*-parameter (Baker, 2001); however, as the APM features eight possible responses, the *c*-parameter should ideally be less than 0.13 (1/8; Guyer & Thompson, 2014). Differences between models were examined using the log-likelihood statistic (-2Log; De Ayala, 2009; Tay et al., 2015). A significant difference suggests the model with the greater number of parameters has superior fit (Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). Test information functions (TIF) were derived for each of the APM-12U and APM-12T to show how much information the tests provide at each ability (or theta) level.

In order to detect item response differences between men and women, the Mantel–Haenszel statistic was used for differential item functioning (DIF). Considered the gold

standard for comparing the performance of two groups on dichotomously scored items, this involves splitting each group into severity levels and then comparing the probability of individuals in each group endorsing each item at each severity level (Gómez-Benito et al., 2018; Guyer & Thompson, 2014; Sireci & Rios, 2013; Tay et al., 2015). Log-transformed Mantel–Haenszel coefficients (M-H D) were computed for each item – values less than/greater than zero indicate males/females or native English/non-native English speakers were more likely to endorse the item – and tested for significance (Guyer and Thompson, 2014).

Results

Study 1

Descriptive data and internal consistency

APM-12U scores for the full sample ranged from 0 to 12 with a mean of 7.76 (*SD* = 2.72). Scores were normally distributed, though 6.9% of the sample reached ceiling. There was a small difference between the total score of male (*M* = 8.06, *SD* = 2.67) and female (*M* = 7.59, *SD* = 2.74) participants, $t(606) = 2.03$, $p = .043$, $d = 0.17$, 95% CIs [0.01, 0.34], and between the scores of individuals who spoke English as a first (*M* = 7.64, *SD* = 2.76) versus other (*M* = 8.21, *SD* = 2.51) language, $t(606) = -2.13$, $p = .034$, $d = 0.21$, 95% CIs [-0.41, -0.02]. Participants completed the 12 items in just over 15 minutes (944.44s, *SD* = 1113.20; range: 93.26-17050.53), although it should be noted timing data was available for 567 participants only. APM-12U scores were correlated with age, $r(606) = .10$, $p = .015$, 95% CIs [0.02, 0.18], years of education, $r(606) = .21$, $p < .001$, 95% CIs [0.13, 0.28], and time taken to complete the assessment, $r(565) = .20$, $p < .001$, 95% CIs [0.12, 0.28], though effect sizes were small. The correlation between any one question on the APM-12U and time to complete that question was, in all but three cases (original matrices 15, 30 and 31), non-significant. Progressive item difficulty, item-total correlations, and coefficient alphas if the item is deleted are detailed in Table 2. Cronbach's alpha and Guttman's λ_2 were both .74 and the

Spearman-Brown split-half reliability co-efficient was .77. The CFA showed a single-factor model well represents the structure of the APM-12U (CFI = 0.953; RMSEA = 0.045; SRMR = 0.067). Factor loadings were all significant (Table 2).

Given the undimensionality of the APM-12U, unidimensional IRT analyses were performed. In both the 2-PL and 3-PL models, item fit *p*-values were all non-significant. While *b*- and *a*-parameter values for both models were within acceptable bounds, there was greater (or wider) discrimination between items in the 3-PL model. The *c*-parameter was less than or equal to 0.13 for all but one item (original APM item 8), but always less than 0.35 in the 3-PL model. The difference between models was significant, $\Delta\text{-}2\text{Log}_{2\text{PL-3PL}}$ (12) = 22.79, $p < .05$, indicating the 3-PL model had better fit. Values for the *b*-parameter ranged between -2.09 and 0.76 logit in the 3-PL model. Several items appeared out of order; in particular, (original APM) items 4 and 8 might be better presented in reverse order. Values for the *a*-parameter were between 0.43 and 1.10 logit; that is, they showed moderate to high discrimination. Guessing parameters were between 0.11 and 0.27. See Table 3 for item parameters, fit indices, and maximum item information. Maximum test information was 4.07 (corresponding to an ability level of 0.25) and minimum standard error was 0.50. In the -1.0 to 1.0 ability range (one standard deviation below/above the average), the amount of test information was greater than 2.45 and the standard error less than 0.64. See Figure 1 for the TIF. Details of the DIF analyses are also provided in Table 3. No item showed significant DIF as a function of sex; item 25 showed a bias against native English speakers.

Test-retest reliability

Initial APM-12U scores for the test-retest sub-sample ranged from 2 to 12 with a mean of 8.67 (*SD* = 2.60). Participants completed the 12 items in approximately 18 minutes (1126.70s, *SD* = 1251.61; range: 219.96-10366.15). Retesting took place, on average, 22 weeks after initial testing (159.86 days, *SD* = 44.97; range: 76-303). Retest APM-12U scores ranged from 2 to 12 with a mean of 9.34 (*SD* = 2.30). Retest items were completed in just

over 17 minutes (1036.74s, *SD* = 1164.87; range: 135.43-9920.49). Test-retest scores were significantly correlated, $r(88) = .65$, $p < .001$, 95% CIs [0.52, 0.76]; this represents a large effect. There was no significant difference between average test time and retest time, $t(89) = 0.56$, $p = .577$, $d = 0.06$, 95% CIs [-0.15, 0.27].

Validity

APM-12U scores of the sub-sample that also completed the WASI-II ranged from 0 to 12 with a mean of 7.98 (*SD* = 2.67). Mean WASI-II Verbal Comprehension and Perceptual Reasoning Indices plus FSIQ-4 for this sub-sample are presented in Table 4. There were no significant differences between male and female participants in terms of Verbal Comprehension, $t(85) = 1.51$, $p = .135$, $d = 0.34$, 95% CIs [-0.11, 0.79], Perceptual Reasoning, $t(85) = 1.41$, $p = .161$, $d = 0.32$, 95% CIs [-0.13, 0.76], or FSIQ-4, $t(85) = 1.86$, $p = .067$, $d = 0.42$, 95% CIs [-0.03, 0.86], on the WASI-II. There was a large significant difference between the scores of individuals who spoke English as a first or other language with regard to Verbal Comprehension, $t(85) = 6.17$, $p < .001$, $d = 1.39$, 95% CIs [0.90, 1.88], and FSIQ-4, $t(85) = 4.52$, $p < .001$, $d = 1.02$, 95% CIs [0.55, 1.48], but not Perceptual Reasoning, $t(85) = 0.50$, $p = .620$, $d = 0.11$, 95% CIs [-0.33, 0.55]. See Table 4 for WASI-II scores as a function of native and non-native English speakers as well as correlations between the APM-12U scores and WASI-II.

Study 2

Descriptive data and internal consistency

APM-12T scores for the full sample ranged from 0 to 12 with a mean of 6.56 (*SD* = 2.37). Scores were normally distributed, with 1.25% reaching ceiling. There was a small difference between the scores of male (*M* = 6.90, *SD* = 2.43) and female (*M* = 6.42, *SD* = 2.34) participants, $t(477) = 1.99$, $p = .047$, $d = 0.20$, 95% CIs [0.003, 0.40]. There was no difference between the scores of individuals who spoke English as a first versus other language, $t(477) = -1.79$, $p = .074$, $d = -0.17$, 95% CIs [-0.37, 0.02]. Participants completed

the 12 items in less than eight minutes (440.64s, *SD* = 122.34; range: 37.16-698.51); on average, they spent 36.55 (*SD* = 10.43) seconds on each item. While participants spent less time on each of the first six matrices (32.22 seconds per item) and more time on the latter six (41.18 seconds per item), they did not appear to require the full 60 seconds to complete any one question. APM-12T scores were not correlated with age, $r(477) = -.02$, $p = .676$, 95% CIs [-0.11, 0.07], or years of education, $r(477) = .07$, $p = .121$, 95% CIs [-0.02, 0.16]. There was a medium sized correlation between APM-12T scores and time taken to complete the assessment, $r(477) = .31$, $p < .001$, 95% CIs [0.23, 0.39]. Progressive item difficulty, item-total correlations, and coefficient alphas if the item is deleted are detailed in Table 2. Cronbach's alpha and Guttman's λ_2 were .60 and .61 respectively. The Spearman-Brown split-half reliability co-efficient was .61. The CFA showed a single-factor model well represents the structure of the APM-12T (CFI = 0.981; RMSEA = 0.017; SRMR = 0.056). Factor loadings were all significant (Table 2).

Unidimensional IRT analyses were also performed on the APM-12T. In both the 2-PL and 3-PL models, item fit *p*-values were all non-significant. While *b*- and *a*-parameter values for both models were within acceptable bounds, again, there was greater (or wider) discrimination between items in the 3-PL model. The *c*-parameter was less than or equal to 0.13 for most but not all items (original APM items 8, 30, and 31) in the 3-PL model; it was always less than 0.35. The difference between models was significant, $\Delta\text{-}2\text{Log}_{2\text{PL-3PL}}$ (12) = 31.57, $p < .002$, indicating the 3-PL model had better fit. Values for the *b*-parameter ranged between -1.13 and 1.67 logit in the 3-PL model. The majority of items appeared out of order. Values for the *a*-parameter were between 0.49 to 1.11 logit, indicating moderate to high discrimination. Guessing parameters were between 0.12 and 0.28. See Table 5 for item parameters, fit indices, and maximum item information. Maximum test information was 3.28 (corresponding to an ability level of 0.85) and minimum standard error was 0.55. In the -1.0 to 1.0 ability range (one standard deviation below/above the average), the amount of test

information was greater than 1.72 and the standard error less than 0.76. See Figure 1 for the TIF. Details of the DIF analyses are also provided in Table 5. No item showed significant DIF as a function of sex; item 35 showed a bias against native English speakers.

Test-retest reliability

Initial APM-12T scores for the test-retest sub-sample ranged from 1 to 12 with a mean of 6.40 (*SD* = 2.55). Participants completed the 12 items in approximately 7 minutes (420.84s, *SD* = 130.55; range: 37.16-647.66). Retesting took place 17.20 (*SD* = 6.26; range: 6-34) days after initial testing. Retest APM-12T scores ranged from 0 to 12 with a mean of 6.52 (*SD* = 2.94). Test-retest scores were significantly correlated, $r(142) = .69$, $p < .001$, 95% CIs [0.59, 0.77]; this represents a large effect. Retest items were completed in around 5 minutes (296.97s, *SD* = 124.56; range: 36.77-581.20). There was a significant difference between average test time and retest time, $t(143) = 16.70$, $p < .001$, $d = 1.39$, 95% CIs [1.16, 1.62].

Validity

APM-12T scores of the sub-sample that also completed the WASI-II ranged from 3 to 12 with a mean of 6.88 (*SD* = 2.21). Mean WASI-II Verbal Comprehension and Perceptual Reasoning Indices plus FSIQ-4 are presented in Table 4. There were no significant differences between male and female participants in terms of Verbal Comprehension, $t(40) = -0.52$, $p = .607$, $d = -0.17$, 95% CIs [-0.80, 0.47], Perceptual Reasoning, $t(40) = -0.26$, $p = .798$, $d = -0.08$, 95% CIs [-0.71, 0.55], or FSIQ-4, $t(40) = -.55$, $p = .586$, $d = -0.18$, 95% CIs [-0.81, 0.46], on the WASI-II. There were also no significant differences between the scores of individuals who spoke English as a first or other language with regard to Verbal Comprehension, $t(40) = 1.93$, $p = .061$, $d = 0.60$, 95% CIs [-0.03, 1.22], Perceptual Reasoning, $t(40) = -1.00$, $p = .323$, $d = -0.31$, 95% CIs [-0.93, 0.31], or FSIQ-4, $t(40) = 0.71$, $p = .482$, $d = 0.22$, 95% CIs [-0.39, 0.83]. Table 4 details correlations between the APM-12T scores and WASI-II.

[Tables 2-5 & Figure 1 near here]

Discussion

The aim of this project was to assess the psychometric properties of online untimed (APM-12U; Study 1) and timed (APM-12T; Study 2) 12-item versions of the APM and validate them against the WASI-II.

While both mean APM-12U score (7.76) and administration time (944.44s) were consistent with figures reported in other studies (7.30-7.98; 15 minutes; Ablard & Mills, 1996; Arthur & Day, 1994; Arthur et al., 1999; Day, Espejo, Kowollik, Boatman, & McEntire, 2007), average APM-12T score was lower (6.44) and administration time (440.64s) less. Importantly, variability of test time – and range – was reduced in Study 2. There was a medium strength positive correlation between APM-12T score and time taken to complete the test (9.6% of variance) – as opposed to the small correlation between APM-12U score and test time (4.0% of variance) – suggesting reduced test time might have impacted APM-12T mean score. At the same time, the WASI-II scores of Study 2 participants were somewhat lower than those in Study 1; so, decreased APM-12T scores might reflect the lower overall IQ of this sample.

Classical test theory analyses revealed there were small differences between the mean scores of male and female participants on both the APM-12U and APM-12T (0.6% and 0.8% of variance respectively). There was some difference between the scores of those who spoke English as a first or other language on the APM-12U (0.8% of variance), but not the APM-12T. The significance of these differences may be a reflection of the statistical power engendered by such large samples. Similarly, correlations between APM-12U score and age (1.0% of variance) and years of education (4.4% of variance) were also reasonably small. There was no correlation between APM-12T score and years of education or age. In untimed conditions, individuals with greater years of education – who might be more invested in determining a correct response – or older persons – who might require greater processing

time than younger participants – were able to spend as long on each APM question as required; under timed conditions, this was not possible.

Classical test and IRT analyses showed progressive item difficulty was generally maintained on the APM-12U, though, like other similar investigations, several early items appeared to pose more of a challenge than subsequent items (Arthur et al., 1999; Arthur & Day, 1994). Progressive item difficulty was less well maintained on the APM-12T and percent error on all items was higher. It is possible participants found the first few items of the abridged versions of the APM especially difficult because, unlike the long form of the test, they did not have the benefit of multiple relatively straightforward items early in the test. In the case of the APM-12T, response deliberation was strictly limited. Increasing the number of practice items participants must complete could, potentially, circumvent some of these issues, though this would add to test duration.

The internal consistency of the APM-12T was in keeping with values reported previously, though that of the APM-12U was somewhat higher (Arthur et al., 1999; Arthur & Day, 1994). Regardless, it has been noted that single measures of internal consistency, such as Cronbach's alpha and Guttman's λ_2, are not well suited to measuring the APM, as the increasing progressive difficulty of items contribute to heterogeneity; abridged versions of the test may exacerbate this issue as the item difficulty gradient is steeper (Arthur et al., 1999). Test-retest reliability of the APM-12U was notably lower than reported elsewhere (Arthur et al., 1999; Arthur & Day, 1994). This may be a reflection of our longer and more variable test-retest interval. APM-12T test-retest reliability was assessed over a short period and was somewhat higher. Participants who accepted the invitation to be part of the APM-12U test-retest sub-sample may have done so because they were particularly invested in improving on their initial result. This explanation is consistent with the observation that the baseline mean score of this sub-sample (but not the APM-12T test-retest sub-sample) was substantially higher than that of the whole APM-12U sample.

CFA results suggested both the APM-12U and APM-12T are unidimensional, with fit indices indicating the one-factor models had adequate to good fit; this is consistent with previous investigations into the 12-item APM (Arthur et al., 1999; Arthur & Day, 1994; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). With the APM-12T, it should be noted (original APM) item 30 had low item-total correlation and a comparatively small factor loading. Also in keeping with previous investigations, a three-parameter IRT model fit the APM-12U and APM-12T data (Chiesi, Ciancaleoni, Galli, & Primi, 2012; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012). In both cases, the 3-PL models provided greater discrimination between items than the 2-PL models. Discrimination between items was similar across both tests in the 3-PL models. Difficulty parameters revealed there was variation across items, though in the case of the APM-12U there was a greater number of easier (< 0) versus harder (> 0) items. Nonetheless, there was a greater range in terms of difficulty in the APM-12U. Guessing parameters in the APM-12U were in all but one case less than or equal to 0.13. Original APM item 8 had a higher guessing parameter, suggesting individuals with low ability could guess the response to this item at a rate greater than expected by chance. In the APM-12T, three items had guessing parameter values greater than 0.13; persons with low capability might be expected to guess the answer to original APM items 8, 30, and 31 with a frequency greater than expected by chance. Rather than leaving an APM-12T item unanswered, participants might, as the amount of time remaining diminished, have guessed a response (Tatel et al., 2022). This should not, however, have resulted in a greater than chance correct response rate. The test information function, which considers all three parameters, was satisfactory. Other IRT studies of short forms of the APM have reported similar parameter values and test information functions (Chiesi, Ciancaleoni, Galli, & Primi, 2012; Chiesi, Ciancaleoni, Galli, Morsanyi, et al., 2012; Waschl et al., 2016). Differential item functioning utilising the Mantel-Haenszel statistic – considered the gold standard for comparing the performance of two groups on dichotomously scored items

(Gómez-Benito et al., 2018; Tay et al., 2015) – identified no item-level differences between the performance of men and women on either the APM-12U or APM-12T. On each of the two abridged versions of the APM, a single item appeared to favour non-native English speakers: (original APM) item 25 on the APM-12U and (original AMP) item 35 on the APM-12T.

In terms of validity, correlations between the APM-12U/APM-12T and WASI-II Perceptual Reasoning and FSIQ-4 were in the moderate range. In particular, the correlation with Perceptual Reasoning (.50/.56) was similar to that reported in previous studies between the 36-item APM and WAIS performance IQ (.55; McLaurin et al., 1973; Paul, 1986). The non-significant correlation between the APM-12U/APM-12T and WASI-II Verbal Reasoning was likely driven by the high numbers of non-native English speakers in each sub-sample (35%/43%). These findings are in keeping with Australian research that has found non-native speakers are disadvantaged, as compared to native English speakers, on verbal subtests of the WAIS-R (Carstairs et al., 2006). Overall, correlations with WASI-II perceptual reasoning and FSIQ-4 were stronger in the case of the APM-12T than the APM-12U. A recent meta-analysis of abbreviated and/or speeded forms of the APM found, that in the case of short (12-item) versions, the estimated population correlations between the measure and spatial visualisation, verbal ability, and IQ were, respectively, 0.41, 0.27, and 0.43 (Tatel et al., 2022). In terms of speeded (60-79 seconds per item) versions of the APM, estimated population correlations between the test and spatial visualisation, verbal ability, and IQ were, respectively, 0.52, 0.30, and 0.41 (Tatel et al., 2022). While the tests of spatial visualisation, verbal ability, and IQ represented by papers included in the meta-analysis do not necessarily directly accord with WASI-II subtests or the FSIQ-4, our validity correlations are similar in that associations related to verbal reasoning are low, while those pertaining to timed administration appear to be generally stronger. It should be noted, however, that the authors of this paper found faster administration appears to strengthen the relationship between

underlying ability correlates of APM performance – such as, spatial visualisation, perceptual speed, and working memory – and APM score (Tatel et al., 2022). Though Tatel and colleagues (2022) acknowledge their findings are not definitive, this nonetheless calls into question claims that any speeded version can still be considered a sole indicator of fluid intelligence. Certainly, the original APM authors suggest that timed administration places a higher load on intellectual efficiency than perception and clear thinking (Raven et al., 1998).

Taken together, our psychometric findings tend to indicate the APM-12T offers advantages over the APM-12U in terms of mean score and distribution of results, particularly the proportion of participants likely to reach ceiling. Regarding administration time, although participants completed the APM-12U in, on average, just over 15 minutes, test times were highly variable. In the absence of the researcher, it is not possible to know if test times reflect non-compliance, compromised task performance, preoccupation with obtaining a high score, or an interruption to test taking. Administration time and variability of time taken to complete the test were more constrained on the APM-12T; this may limit participant burden and thus make it an appealing method of assessing intellectual ability when it is to form part of a longer battery. The APM-12U appears to better maintain progressive item difficulty and has superior internal consistency, though not test-retest reliability. CFA item factor loadings are higher on the APM-12U. In addition, IRT analyses suggest, relative to the APM-12T, there is an increased range in terms of difficulty, and less chance low ability participants will guess a correct answer on the APM-12U. Regarding validity, there are strong associations between scores on both abridged versions of the APM and perceptual reasoning, regardless of the native English speaker status of the participant. While these results suggest, perhaps, the superiority of the APM-12U, this study was not one of direct comparison; that is, we employed different samples – and sample sizes – in each study. Moreover, the validity component of the APM-12T study employed a relatively small sample. Thus, future studies utilising a within-subjects design are needed to confirm and extend our findings.

In conclusion, this two-part study assessed the psychometric properties of abridged online versions of Raven's APM: the untimed 12-item APM-12U and the timed 12-item APM-12T. Though test time variability and range issues may make the APM-12T a more suitable choice when it likely to form part of a longer battery of tests, classical test and IRT analyses suggest the APM-12U might be the more superior abridged form. Both measures demonstrated sound concurrent validity against the Perceptual Reasoning and FSIQ-4 indices of the WASI-II, though caution should be applied when applying administration time limits due to possible changes in construct validity. Given progressive item difficulty inconsistencies, researchers could consider the use of several additional practice items before administering either of these abridged versions of the APM, though this would increase total test duration time. It should also be noted that, compared to the complete 36-item APM, the psychometric properties and concurrent validity of these abridged versions are less robust. As such, where the outcome of testing is likely to inform clinical diagnoses or the allocation of resources, the complete APM will always provide a more comprehensive assessment of intellectual capacity. In sum, the APM-12U and APM-12T both hold promise as tools for brief online assessment of intellectual capacity.

Acknowledgements

The authors wish to thank Cameron Patrick from the Melbourne Statistical Consulting Platform for advice conducting analyses.

Disclosure of Interest

The authors report no conflict of interest.

Data and Materials Availability

The datasets analysed for the current study are available in the Open Science Framework repository (https://osf.io/ry23m/). The study was not pre-registered.

Permissions

A Pearson Research License Agreement is in place to enable use of the Raven's Progressive Matrices for research purposes.

Funding

This research was supported by an Australian National Health and Medical Research Council grant (1050766) and an Australian Research Council fellowship (FT110100088).

Ethics approval

This study was performed in line with the principles of the Declaration of Helsinki. The University of Melbourne Human Ethics Committee approved this study (ID: 1544791).

Consent to participate/ Consent for publication

All participants provided informed consent. In doing so, they acknowledged reading a Plain Language Statement that explained aggregated group level data from this study may be published or presented at conferences.

References

Ablard, K. E., & Mills, C. J. (1996). Evaluating abridged versions of the Raven's Advanced Progressive Matrices for identifying students with academic talent. *Journal of Psychoeducational Assessment*, *14*(1), 54–64. https://doi.org/https://doi.org/10.1177/073428299601400105

Alderton, D. L., & Larson, G. E. (1990). Dimensionality of Raven's Advanced Progressive Matrices items. *Educational and Psychological Measurement*, *50*(4), 887–900. https://doi.org/10.1177/0013164490504019

Arce-Ferrer, A. J., & Martinez Guzman, E. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the Raven's Standard Progressive Matrices test. *Educational and Psychological Measurement*, *69*(5), 855–867. https://doi.org/10.1177/0013164409332219

Arthur, W., & Day, D. (1994). Development of a short form for the Raven Advanced Progressive Matrices test. *Educational and Psychological Measurement*, *54*(2), 394–403. https://doi.org/10.1177/0013164494054002013

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17*(4), 354–361. https://doi.org/10.1177/073428299901700405

Arthur, W., & Woehr, D. J. (1993). A confirmatory factor analytic study examining the dimensionalty of the Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, *53*(2), 471–478. https://doi.org/10.1177/0013164493053002016

Baker, F. B. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessent and Evaluation.

Bors, D. A., & Forrin, B. (1995). Age, speed of information processing, recall, and fluid intelligence. *Intelligence*, *20*(3), 229–248. https://doi.org/10.1016/0160-2896(95)90009-8

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*(3), 382–398. https://doi.org/10.1177/0013164498058003002

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences*, *13*(4), 291–312. https://doi.org/10.1016/S1041-6080(03)00015-3

Carstairs, J. R., Myors, B., Shores, E. A., & Fogarty, G. (2006). Influence of language background on tests of cognitive abilities: Australian data. *Australian Psychologist*, *41*(1), 48–54. https://doi.org/10.1080/00050060500391878

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the Advanced Progressive Matrices. *Learning and Individual Differences*, *22*(3), 390–396. https://doi.org/10.1016/j.lindif.2011.12.007

Chiesi, F., Ciancaleoni, M., Galli, S., & Primi, C. (2012). Using the Advanced Progressive Matrices (Set I) to assess fluid ability in a short time frame: An item response theory-based analysis. *Psychological Assessment*, *24*(4), 892–900. https://doi.org/10.1037/a0027830

Çikrikçi-Demirtasli, N. (2002). A study of Raven Standard Progressive Matrices test item measures under classic and item response models: An empirical comparison. *Journal of Faculty of Educational Sciences*, *35*(1–2), 71–79.

https://doi.org/10.1501/egifak_0000000055

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Day, E. A., Espejo, J., Kowollik, V., Boatman, P. R., & McEntire, L. E. (2007). Modeling the links between need for cognition and the acquisition of a complex skill. *Personality and Individual Differences*, *42*(2), 201–212. https://doi.org/10.1016/j.paid.2006.06.012

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., Frans, Ö., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jiménez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, H. C., & Urbánek, T. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*, *17*(4), 300–319. https://doi.org/10.1027/1016-9040/a000102

Gallini, J. K. (1983). A rasch analysis of raven item data. *Journal of Experimental Education*, *52*(1), 27–32. https://doi.org/10.1080/00220973.1983.11011869

Gómez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104–109. https://doi.org/10.7334/psicothema2017.183

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Guyer, R., & Thompson, N. A. (2014). *Xcalibre: Item response theory calibration software user manual*. Assessment System Corporation.

Harris, A., McMillan, J., Listyg, B., Matzen, L., & Carter, N. (2020). Measuring Intelligence with the Sandia Matrices: Psychometric Review and Recommendations for Free Raven-Like Item Sets. *Personnel Assessment and Decisions*, *6*(3). https://doi.org/10.25035/pad.2020.03.006

Kpolovie, P. J., & Emekene, C. (2016). Item Response Theory Validation of Advanced Progressive Matrices in Nigeria. *British Journal of Psychology Research*, *4*(1), 1–32. http://www.eajournals.org/wp-content/uploads/Item-Response-Theory-Validation-of-Advanced-Progressive-Matrices-in-Nigeria1.pdf

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

McLaurin, W. A., Jenkins, J. F., Farrar, W. E., & Rumore, M. C. (1973). Correlations of IQs on verbal and nonverbal tests of intelligence. *Psychological Reports*, *33*(3), 821–822. https://doi.org/10.2466/pr0.1973.33.3.821

Mills, C. J., & Tissot, S. L. (1995). Identifying academic potential in students from under-represented populations: Is using the Ravens Progressive Matrices a good idea? *Gifted Child Quarterly*, *39*(4), 209–217. https://doi.org/https://doi.org/10.1177/001698629503900404

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's guide* (6th ed.). Muthén & Muthén.

Paul, S. M. (1986). The Advanced Raven's Progressive Matrices. *Journal of Experimental Education*, *54*(2), 95–100. https://doi.org/10.1080/00220973.1986.10806404

Poulton, A. (2021). *Evaluating abridged versions of the Advanced Progressive Matrices*. osf.io/ry23m

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulay Scales: Section 4 Advanced Progressive Matrices Sets I & II*. Pearson.

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment*, *12*(3), 220–229. https://doi.org/10.1111/j.0965-075X.2004.00276.x

Sargis, E. G., Skitka, L. J., & McKeever, W. (2013). The internet as psychological laboratory revisited: Best practices, challenges, and solutions. In Y. Amichai-Hamburger (Ed.), *The*

social net: Understanding our online behavior. Oxford University Press.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, *2*(2), 107–114. https://doi.org/10.1177/2515245919838781

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

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2–3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. Pearson.

Tatel, C. E., Tidler, Z. R., & Ackerman, P. L. (2022). Process differences as a function of test modifications: Construct validity of Raven's Advanced Progressive Matrices under standard, abbreviated and/or speeded conditions – A meta-analysis. *Intelligence*, *90*, 101604. https://doi.org/10.1016/j.intell.2021.101604

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3–46. https://doi.org/10.1177/1094428114553062

Waschl, N. A., Nettelbeck, T., Jackson, S. A., & Burns, N. R. (2016). Dimensionality of the Raven's Advanced Progressive Matrices: Sex differences and visuospatial ability. *Personality and Individual Differences*, *100*, 157–166. https://doi.org/10.1016/j.paid.2015.12.008

Wechsler, D. (2011). *WASI-II: Wechsler Abbreviated Scale of Intelligence Manual - Second edition* (Second Edi). Pearson.

Williams, J. E., & McCord, D. M. (2006). Equivalence of standard and computerized

versions of the Raven Progressive Matrices Test. *Computers in Human Behavior*, *22*(5), 791–800. https://doi.org/10.1016/j.chb.2004.03.005

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.