


Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test

Assessment
19(3) 354–369
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191112446655
<http://asm.sagepub.com>


Warren B. Bilker¹, John A. Hansen¹, Colleen M. Brensinger¹,
Jan Richard¹, Raquel E. Gur¹, and Ruben C. Gur¹

Abstract

The Raven's Standard Progressive Matrices (RSPM) is a 60-item test for measuring abstract reasoning, considered a nonverbal estimate of fluid intelligence, and often included in clinical assessment batteries and research on patients with cognitive deficits. The goal was to develop and apply a predictive model approach to reduce the number of items necessary to yield a score equivalent to that derived from the full scale. The approach is based on a Poisson predictive model. A parsimonious subset of items that accurately predicts the total score was sought, as was a second nonoverlapping alternate form for repeated administrations. A split sample was used for model fitting and validation, with cross-validation to verify results. Using nine RSPM items as predictors, correlations of .9836 and .9782 were achieved for the reduced forms and .9063 and .8978 for the validation data. Thus, a 9-item subset of RSPM predicts the total score for the 60-item scale with good accuracy. A comparison of psychometric properties between 9-item forms, a published 30-item form, and the 60-item set is presented. The two 9-item forms provide a 75% administration time savings compared with the 30-item form, while achieving similar item- and test-level characteristics and equal correlations to 60-item based scores.

Keywords

Raven's Standard Progressive Matrices, intelligence testing, mental abstraction, analogical reasoning, scale reduction, predictive model

The Raven's Standard Progressive Matrices (RSPM) instrument is a multiple-choice test used to assess mental ability associated with abstract reasoning, which Cattell (1963) termed *fluid intelligence*. The test consists of increasingly difficult pattern matching tasks and has little dependency on language abilities. Carpenter, Just, and Shell (1990) describe the associated problem-solving technique as the ability to manage a hierarchy of goals and subgoals as each problem is decomposed into manageable segments of pairwise comparisons. Originally published in 1938 (unpublished thesis; Raven, 1936, 1938), the standard form (RSPM) consists of five sets of 12 matrices presented in black and white. In total, the psychometric properties of the 60 RSPM items have been thoroughly analyzed and are used as an indicator of general intelligence throughout the world (Raven, 1989, 2000). In addition to the set of standard Raven's items, a set of 36 more difficult items has been developed, The Raven's Advanced Progressive Matrices Test (APM), as well as a set of color items, The Raven's Coloured Progressive Matrices Test (Raven, Raven, & Court, 1998).

Although there are many advantages in using a well-tested psychometric instrument such as the RSPM, one limitation is the amount of time required for administration of the full set of items. This limitation has been noted in a number of studies that evaluate the properties of reduced sets of items from both the RSPM and APM (Arthur & Day, 1994; Bors & Stokes, 1998; Hamel & Schmittmann, 2006; Wytek, Opgenoorth, & Presslich, 1984). Williams and McCord (2006) reported the average administration time for a computerized version of the 60-item RSPM to be 17 minutes. The RSPM, as well as the APM, can be administered as a speeded test. However, this too has its drawbacks. Raven, Raven, and Court (1993) noted that as a speeded test, *intellectual efficiency* is assessed. Hamel and Schmittmann

¹University of Pennsylvania, Philadelphia, PA, USA

Corresponding Author:

Warren B. Bilker, Perelman School of Medicine at the University of Pennsylvania, Department of Biostatistics, 601 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA
Email: warren@upenn.edu

(2006), however, noted that for the untimed test, practice and experience with previous items play a large role in performance. For example, those examinees completing items more quickly have greater recency and familiarity with more items per unit time as they attempt each subsequent item in the test. Furthermore, Raven, Raven, and Court (1998) found that the untimed version of the APM has a one-dimensional factor structure, whereas a timed version invariably includes a dimension for speed. Our development of a short-form RSPM was motivated by the need for a brief assessment with multiple forms for use in large-scale longitudinal or treatment studies in genomically informative populations and in patients with cognitive deficits. The need for abbreviated testing has become even more pressing as cognitive measures have become increasingly used as endophenotypes in large-scale genomic studies. In such studies, the cognitive assessment is a component of a more comprehensive evaluation, and efficiency is critical. The computerized format is also essential in such studies, which need to incorporate, evaluate and analyze massive amounts of data.

Previous attempts at reducing the time for administration included both reducing the number of matrix stimuli as well as setting a fixed time limit for the test administration (the speeded test). Correlations between scores based on reduced sets and the total APM test ranged from $r = .66$ (Arthur & Day, 1994) to $r = .91$ (Bors & Stokes, 1998). Investigating the effects of a fixed administration time, Hamel and Schmittmann (2006) found a correlation of $r = .75$ between 20-minute limit scores and nontimed scores, using a test-retest interval of 2 months on the APM test.

Arthur and Day (1994), Bors and Stokes (1998), and Wytek et al. (1984) all used item-reduction strategies based on both classical test theory (CTT) and item response theory (IRT). Specifically, Arthur and Day (1994) used APM scores from a sample of 202 university community residents to evaluate item-total correlations, item difficulty, and conditional alpha coefficients. Based on item difficulty, 12 groupings of three items were created. From each grouping, the item with greatest item-total correlation was selected. Ties were broken by selecting the more difficult item or, in the case of equal difficulty, the item with the greater conditional alpha coefficient. Cronbach's α for the 12-item short form was $r = .72$ compared with $r = .84$ for the original APM. In a second approach to shortening the same test, Bors and Stokes (1998) used scores from a sample of 506 undergraduates to rank order items by their item-total correlations, removing redundant items using item factor scores that were based on a single factor solution derived from a tetrachoric correlation matrix. The resulting 12-item short form had a Cronbach's α of $r = .73$. A third approach, by Wytek et al. (1984), involved choosing items based on iterative selection from Rasch model

statistics. The model estimates were based on a sample of 300 outpatients and "testable" inpatients in a psychiatric clinic. A single split-half reliability coefficient of the 30 selected items was $r = .95$. Moving beyond item-level characteristics, an alternative approach to item reduction involves determining the best combination of items and item weights that can predict a full-instrument score.

The primary goal of this study was to identify a parsimonious subset of the 60 RSPM items that can be used in a reduced scale that will be highly predictive of the 60-item RSPM scale total. The secondary goal was to identify a second subset of the RSPM items that were not included in the first subset and that were also highly predictive of the 60-item RSPM scale total. The second subset will allow for a follow-up test without repeating items. These two subsets of items, Form A and Form B, can be used in place of the 60-item RSPM scale and can be administered in a much shorter time frame.

Method

Data

The sample used for this study included 180 consecutive participants assessed on the RSPM at the University of Pennsylvania's Brain Behavior Laboratory. The sample included individuals ranging in age from 16 to 77 years, with a mean age of 33.9 years ($SD = 12.6$). Participants had an average of 14.5 years ($SD = 2.6$) of education and 54.4% were male. It included large groups of healthy volunteers and patients with schizophrenia, and smaller groups of relatives of patients and nonpsychotic patients with Axis I disorders. Table 1 presents the participant demographics. Suitability for participation in the study at the Penn Schizophrenia Research Center was determined based on specified inclusion and exclusion criteria established by standardized screening and assessment. Participants underwent medical, neurological, and psychiatric evaluations. The comprehensive intake evaluation applied established procedures (Gur, Ragland, Moberg, Turner, et al., 2001a; Gur, Ragland, Moberg, Bilker, et al., 2001b), including the Diagnostic Interview for Genetic Studies (Nurnberger et al., 1994). All participants provided signed informed consent in accordance with the university's institutional review board policies. For those less than 18 years of age, assent and parental approval were required. They were all proficient in English, medically and neurologically healthy, and those with a psychiatric diagnosis met the *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (American Psychiatric Association, 1994) criteria based on consensus diagnosis.

A split sample approach was used for the scale reduction, where $n = 90$ participants were used to fit the models,

Table 1. Study Participant Demographics

Variable	Group	Overall (N = 180)
		n (%)
Group	Normal control	60 (33.3%)
	Family	35 (19.4%)
	Other Axis I	12 (6.7%)
	Schizophrenia	73 (40.6%)
Sex	Male	98 (54.4%)
	Female	82 (45.6%)
Race	White	133 (73.9%)
	Black	38 (21.1%)
	Asian	2 (1.1%)
	Hispanic	7 (3.9%)
Mean (SD/Range)		
Age (years)	33.9 (12.6/16, 77)	
Age of onset (years)	23.5 (7.0/15, 45)	
Duration of illness (years)	9.4 (7.9/0, 29)	
Education (years)	14.5 (2.6/9, 22)	
Total Raven's Correct (RSPM 60-item scale)	41.3 (13.2/2, 59)	

Note. RSPM = Raven's Standard Progressive Matrices.

the model construction data set, and the remaining 90 participants were used for validation. A cross-validation procedure was used to verify the results as described below.

In prior research, a method was developed to reduce the 21-item Carpenter Quality of Life Scale (QLS; Bilker et al., 2003). In the QLS, each scale item has an ordinal value, ranging from 0 (*severe impairment*) through 6 (*high functioning*). To determine an optimal combination of a reduced set of items, all combinations of QLS items were considered for all subsets containing 1 to 10 items. This can be shown to be

$$\sum_{i=1}^{10} \binom{21}{i} = 1,048,575$$

combinations of items and associated predictive models to be considered. The maximum of 10 items considered represent a reduction to less than one half of the 21 scale items included in the full-scale administration. For each combination of items, an ordinary linear regression model was fit to predict the total score of the items not included in the model. The sum of the QLS items included in the model, the observed items of the proposed reduced scale, was then added to the predicted partial total to obtain the predicted total 21-item QLS score for the particular combination of items. The predictive ability of the combination of items was then assessed using the Pearson correlation of the predicted total with the true observed total (intraclass correlation was also assessed to allow for systematic deviations,

which were not present). It was shown that a linear combination of a subset of only 7 of the 21 items, where the weights are the estimated beta coefficients from the regression model, had a Pearson correlation of $r = .9831$ with the observed total of the 21 QLS items. This result was then validated with additional data.

A modified approach was required for the reduction of the RSPM for multiple reasons. First, the RSPM consists of 60 items. Consideration of up to 20 items in the reduced version using the brute force approach applied to the QLS reduction would require

$$\sum_{i=1}^{20} \binom{60}{i} = 7,776,048,412,324,713$$

(nearly 8 quadrillion) item combinations and associated regression models to be fit. Clearly, an alternative approach is needed. We present an algorithm that greatly reduces the number of item combinations that need to be considered. Second, each RSPM scale item is coded as correct or incorrect (binary) rather than an ordinal score as in the QLS. The score of the RSPM is the total number of correct items, yielding a total score of 0 through 60. That is, the total score is the count of correct items. The model of choice for count data is Poisson regression, rather than ordinary linear regression, which was used for the QLS reduction. However, the distribution of the number of correct items, the score, for the RSPM is the mirror image of a Poisson distribution, with a long left tail rather than a long right tail. (For $n = 180$ combined sample: mean = 41.3, median = 45.5, range = 2-59, skewness = -0.94, see Figure 1A; For $n = 90$ model construction sample: mean = 41.0, median = 44.0, range = 2-59, skewness = -1.04, see Figure 1B.) To accommodate the use of Poisson regression, the distribution is reversed, modeling the number of incorrect items, which has the long right tail of the Poisson distribution rather than the number of correct items. The predictions of the total number of incorrect items from the Poisson regression models, with each combination of items considered for each subject, were then converted back to the number of correct items by simply subtracting the prediction from 60 (60-number incorrect).

Consider a specific subset of items to be used to predict the RSPM scale total. A Poisson regression model was fit to predict the total number of incorrect items for those items not in the subset, using each item of the subset as a predictor in the model. Each observed item has a binary value (0 = *correct* and 1 = *incorrect*). For example, consider the case where Items 1 to 5 are the subset of items being evaluated. In this case, Items 1 to 5 are the predictors of the total number of incorrect items in Items 6 to 60. Let Y be the total number of correct items for the full scale of RSPM Items 1 to 60, R be the total number of incorrect items among Items 6 to 60 (to be predicted), and V be the number of incorrect items in Items 1 to 5 (observed). The initial goal is to

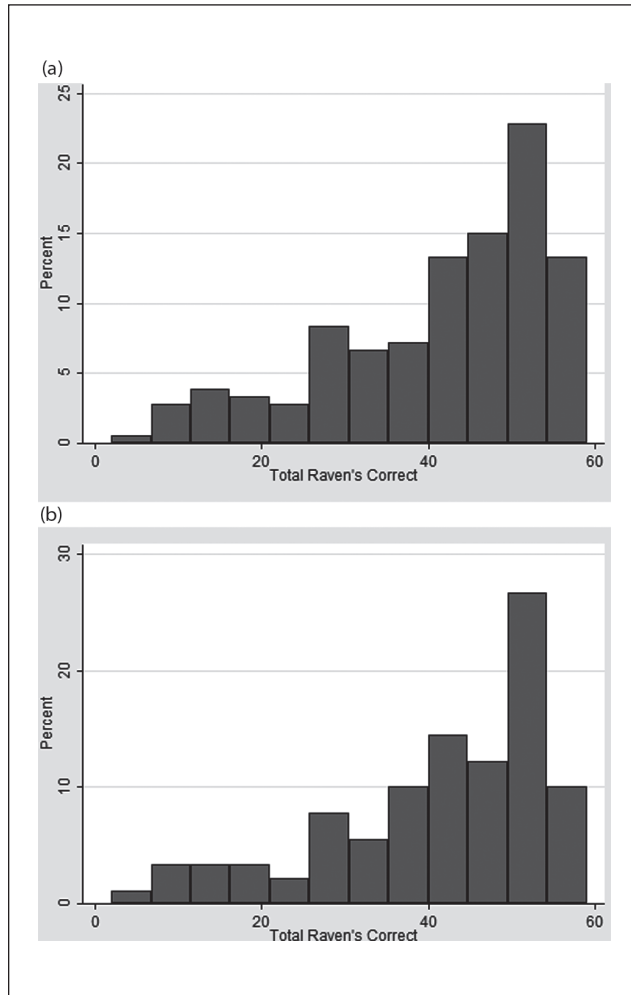


Figure 1. (A) Total Raven's Correct: Combined data set ($n = 180$). (B) Total Raven's Correct: Model building data set ($n = 90$)

predict the total incorrect of the unobserved Items, 6 to 60, using a model with the observed Items, 1 to 5, as the predictors. The Poisson regression model is used to estimate R , and the number of incorrect items among items not included in the predictive model (Items 6-60) is as follows:

$$\log(R) = \alpha + \beta_1 I_1 + \beta_2 I_2 + \beta_3 I_3 + \beta_4 I_4 + \beta_5 I_5,$$

where I_j is 1 if item j is incorrect and 0 if correct, β_j is the coefficient for item j , and α is the intercept. The subscripting for individual participants is not included in the formula for ease of presentation. After fitting the Poisson regression model, the predicted value for the total number of incorrect items in Items 6 to 60, \hat{R} , is determined as $\hat{R} = \exp(\hat{\alpha} + \hat{\beta}_1 I_1 + \hat{\beta}_2 I_2 + \hat{\beta}_3 I_3 + \hat{\beta}_4 I_4 + \hat{\beta}_5 I_5)$ where the beta coefficient estimates are obtained from the fitted model. \hat{R} is taken to be 0 in cases where the estimate is negative. Additionally, \hat{R} is rounded to the nearest integer value to best represent a count.

The predicted value for Y , the total number of correct RSPM items (Items 1-60), is $\hat{Y} = 60 - (V + \hat{R})$, 60 minus the total number of incorrect among Items 1 to 5 (the observed items for this example) minus the predicted total number of incorrect for Items 6 to 60 (items not observed in this example). In practice this is rounded in the following form $\hat{Y} = 60 - (V + \text{Round}(\hat{R}))$ to achieve an integer value for the predicted number correct.

For each model, the predicted values of the total number of correct RSPM items are determined for each individual. The Pearson correlation coefficient of the predicted number of correct items with the true observed number of correct items is then computed as the metric from which to assess each subset of items. Note that the intraclass correlation coefficient (ICC) is also obtained and is presented in the results tables. The two correlation coefficients are nearly identical in all cases. Additionally, for each model and number of items considered, predictions of the RSPM total score were obtained for the validation data set applying the model obtained from the model construction data set, and the Pearson correlation (and ICC) coefficients were obtained for the validation data.

The algorithm used to determine the optimal model for each number of items used to predict the scale total must be able to rule out most of the nearly 8 quadrillion possible combinations of 1 to 20 items from the 60 items in the RSPM as not being sufficiently close to optimal for further consideration. The following algorithm was used where, for each combination of items considered, the Poisson regression model approach described above was applied to estimate the total number of correct RSPM items for each subject, and the Pearson correlation coefficient was determined for each combination of items considered.

The parameters for algorithm are as follows:

N = number of items in scale

M = Maximum number of items selected (models with 1 to M items considered)

S = Number of top models selected for each number of items considered at each stage of model construction (Select top S models that include 1 item, 2 items, etc.)

Algorithm (The value of each parameter specific to the RSPM reduction is shown in parentheses)

1. Consider all N (60) models, each having 1 item as a predictor. For each model, use the above approach to determine the predicted number of items correct, \hat{Y} , and determine the Pearson correlation of the predicted number of correct items from the model with the true observed number of correct items.

2. Select S (30) single items yielding the highest correlations of the predicted number of correct items with the true observed number of correct items.
3. For each of the S (30) top single items selected in Step 2, create 2-item combinations with each of the remaining $N - 1$ ($60 - 1$) items.
4. Select S (30) 2-item combinations yielding the highest correlations of the predicted number of correct items with the true observed number of correct items.
5. For each of the S (30) top 2-item combinations selected in Step 4, create 3-item combinations with each of the remaining $N - 2$ ($60 - 2$) items, eliminating any duplicates.
6. Select S (30) 3-item combinations yielding the highest correlations of the predicted number of correct items with the true observed number of correct items.
7. For each of the S (30) top 3-item combinations selected in Step 6, create 4-item combinations with each of the remaining $N - 3$ ($60 - 3$) items, eliminating any duplicates.
8. Continue procedure to add items, until the top S (30) models containing M (20) items are generated.
9. Choose the model with M (20) items that yields the highest correlation of the predicted number of correct items with the true observed number of correct items.

Note that, at each stage, there is the possibility for duplicates of some item combinations. However, it is not possible to elaborate these in advance since they are specific to the scale being reduced. Thus, it is not possible to obtain an exact number of models to be fit using this algorithm. The maximum number of models fit, without elimination of duplicates, can be shown to be

$$N + S(M - 1)(N - M / 2).$$

For the RSPM reduction this is

$$60 + 30(20 - 1)\left(60 - \frac{20}{2}\right) = 28,560 \text{ models.}$$

Note that when this new method is applied to the original data from the QLS reduction, the identical results are obtained as in the original QLS reduction study (Bilker et al., 2003). However, the clock time needed to complete the computer runs to fit all of the needed models for the QLS reduction, on the same computer platform, was 15 days for the original QLS reduction approach and 1 hour for the modified approach presented here.

At completion of the procedures described above to determine the top model (subset of items with highest

Pearson correlation) for each number of items considered, 1 to 20 RSPM items, the task was to determine the number of items to be used as the “optimal” subset of items for the Form A subset of RSPM items. In determining the “optimal” number of items for the reduced scale, there is a trade-off between parsimony and correlation. As identified in the QLS reduction study (Bilker et al., 2003), formal approaches, such as bootstrapping and permutation methods, for testing the statistical significance of the increases in correlation of adding each predictor find statistical significance even for clinically unimportant and minuscule increases in correlation, and are thus not helpful because of their excessive sensitivity for this particular problem. The same was true for a paired t test to test the decrease in absolute error associated with each additional RSPM item as a predictor. Instead, an informal approach was used to examine the relative increases in correlations for increasing numbers of predictors, and that approach will be applied for the RSPM reduction. For each increase in the number of items used for prediction, the percentage decrease in the Pearson correlation is computed. The *a priori* rule was to add items until the increase in the percentage gain in Pearson correlation for adding an additional item was less than 0.5% for the model construction data set.

Two separate subsets of the 60-item RSPM were sought, Form A and Form B. Form A was determined using the above procedures to select the optimal parsimonious subset of 60 RSPM items. Form B was then determined using the same procedures, excluding the items selected for inclusion in Form A.

Validation of the results is an important aspect of any project involving prediction. Multiple approaches to validation were used to assess the final optimal parsimonious subset of RSPM items selected and the associated model. First, a bootstrap approach was used to assess the Pearson correlation of both the model construction and validation data sets, and the difference between the two correlations. In this approach, separate bootstrap data sets were selected from the model construction data set ($n = 90$) and the validation data set ($n = 90$). The following procedure was applied to the bootstrap data sets. Using the final optimal item subset from Form A, the Poisson regression model was refit to the bootstrap model construction data set (same final items as Form A but with parameter estimates based on the bootstrap sample), and estimates of the number of correct RSPM items and the correlations with the true RSPM scale total were obtained. Additionally, the same model from the bootstrap sample was used to predict the scale total from the bootstrap validation sample, as well as the correlations with the true scale total. The drop in correlation between the model construction and validation samples was then computed. This bootstrap process was repeated 10,000 times, allowing for an estimate of the distribution of the drop in correlation between the model construction and validation data sets.

Table 2. Descriptive Statistics for the Predicted Total Raven's Score, Form A

Number of Predictors	Items in Top Model	Model Construction Sample (<i>n</i> = 90)			Validation Sample (<i>n</i> = 90)		
		ρ	ICC	Mean \pm SD (Minimum, Maximum), 40.98 \pm 13.28 (2, 59)	ρ	ICC	Mean \pm SD (Minimum, Maximum), 41.64 \pm 13.11 (9, 59)
1	31	0.7751	0.7516	41.09 \pm 10.23 (20, 46)	0.8164	0.7964	39.36 \pm 11.40 (20, 46)
2	17, 52	0.8832	0.8747	40.97 \pm 11.44 (13, 50)	0.7250	0.7180	41.80 \pm 11.17 (13, 50)
3	17, 50, 52	0.9196	0.9173	40.73 \pm 12.26 (13, 51)	0.7964	0.7876	42.23 \pm 11.20 (13, 51)
4	17, 22, 51, 52	0.9416	0.9409	40.97 \pm 12.61 (8, 51)	0.8508	0.8513	41.49 \pm 12.53 (8, 51)
5	24, 31, 40, 52, 53	0.9576	0.9578	41.08 \pm 12.96 (12, 53)	0.8744	0.8732	41.11 \pm 13.97 (12, 53)
6	10, 24, 31, 40, 52, 53	0.9666	0.9669	41.03 \pm 13.07 (10, 53)	0.8811	0.8821	41.58 \pm 13.31 (10, 53)
7	14, 24, 31, 40, 51, 52, 53	0.9720	0.9722	41.07 \pm 13.09 (4, 53)	0.8882	0.8821	40.97 \pm 14.75 (4, 53)
8	11, 24, 28, 43, 48, 49, 53, 55	0.9778	0.9779	41.03 \pm 13.08 (11, 56)	0.9066	0.9028	41.31 \pm 14.48 (11, 56)
9	11, 24, 28, 36, 43, 48, 49, 53, 55	0.9836	0.9835	40.99 \pm 12.98 (10, 56)	0.9063	0.9027	41.36 \pm 14.47 (10, 56)
10	9, 11, 24, 28, 43, 48, 49, 53, 55, 58	0.9856	0.9857	40.97 \pm 13.15 (7, 57)	0.9230	0.9183	41.31 \pm 14.58 (7, 57)
11	2, 11, 19, 24, 28, 43, 48, 49, 53, 55, 58	0.9881	0.9882	40.98 \pm 13.15 (3, 57)	0.9156	0.9120	41.39 \pm 14.45 (3, 57)
12	2, 11, 19, 24, 28, 36, 43, 48, 49, 53, 55, 58	0.9900	0.9901	40.98 \pm 13.20 (3, 57)	0.9194	0.9155	41.47 \pm 14.50 (3, 57)

Note. ICC= intraclass correlation coefficient.

A second validation approach used cross-validation to assess both the final subset of items selected for Forms A and B and the corresponding correlations, while reproducing the complete model building process. The 180 observations were randomly resplit into two sets of 90 observations (model construction and validation) 200 times. For each random split, the complete item selection procedure was performed, with the exception that the final subsets were fixed at 9 items to match the number of items actually selected for Forms A and B. This allows assessment of the impact of the specific split that was used for Forms A and B, as well as providing another validation of the item subsets selected for Forms A and B considering the total $n = 180$ observations as a representative universe of the subjects for whom the scale reduction is sought.

Results

Two 9-item short forms were derived from the full 60-item RSPM. For Form A, the correlations (and ICCs) for the top models for each number of RSPM item predictors are presented in Table 2, for both the model construction and validation data sets. For brevity, we present the results for 1 to 12 predictors in the tables. The item numbers selected for the top model for each number of predictor items are also presented. It can be seen that, for as few as three items, the Pearson correlation between the predicted and actual total RSPM scores exceeds 0.9 for the model construction data set and this is true for 8 items for the validation data set. The correlations for the model construction data set are also presented in Figure 2. There is a 0.93% gain in correlation for the top models increasing from 5 to 6 item predictors, 0.56% increasing from 6 to 7 predictors, 0.59%

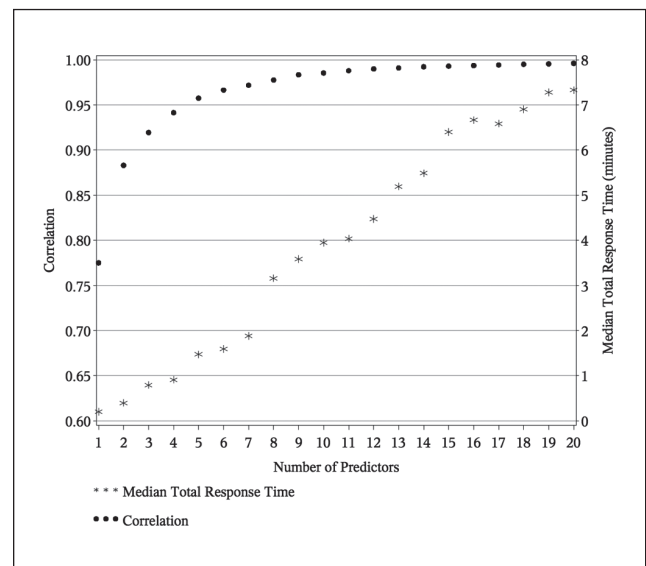


Figure 2. Correlations between actual and predicted total Raven's score, Form A

moving from 7 to 8, 0.59% moving from 8 to 9, and 0.20% moving from 9 to 10 predictors. Applying the a priori rule to add predictors until the gain in correlation fell below 0.5%, the "optimal" number of items for Form A is 9. The correlation of the predicted RSPM total based on Form A with the actual total is 0.9836. A plot of the Actual Total Raven's score versus the Form A Predicted Total Raven's score, with 9 items, for the model construction data set is shown in Figure 3.

The items included in Form A are 11, 24, 28, 36, 43, 48, 49, 53, and 55 from the original 60-item Raven's. The beta coefficients for the top model for each number of items

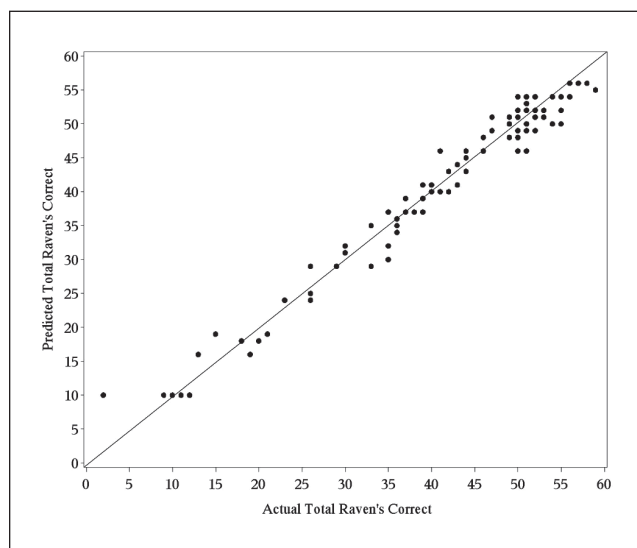


Figure 3. Scatterplot for Raven's Form A

used for prediction in Form A are given in Table 3. Define the values $I_{11}, I_{24}, \dots, I_{55}$ to be 0 if the particular item was correct and 1 if incorrect (since it is the number of incorrect items that is modeled). The prediction of the total RSPM score (total predicted correct) has

$$\hat{R} = \exp(1.323 + 0.198 * I_{11} + 0.216 * I_{24} + 0.237 * I_{28} + 0.142 * I_{36} + 0.374 * I_{43} + 0.304 * I_{48} + 0.178 * I_{49} + 0.458 * I_{53} + 0.289 * I_{55}),$$

$$V = I_{11} + I_{24} + I_{28} + I_{36} + I_{43} + I_{48} + I_{49} + I_{53} + I_{55},$$

and the predicted scale total of $\hat{Y} = 60 - (V + \hat{R})$ (\hat{Y} is increased to 0 if negative).

The performance of the top models for each number of items in the development of Form A was also assessed for two subgroups, patients with schizophrenia and normal controls. The correlations of the predicted totals with the actual totals for these subgroups are given in Table 4. It is seen that the correlations for the schizophrenia group are similar to those obtained in the development of Form A, while for the normal controls the validation correlation is slightly less but is still 0.8720 for 9 items. Thus, predictions from the Form A model perform well for these subgroups.

Recall that the RSPM items considered for Form B are those not included in Form A. For Form B, the correlations (and ICCs) for the top models for each number of RSPM item predictors are presented in Table 5, including the item numbers selected for each top model. As with Form A, for as few as three items, the Pearson correlation between the predicted and actual total RSPM scores exceeds 0.9 for the model construction data set. The correlation for the validation data set achieves 0.9 for 10 items. The correlations for the model construction data set are also presented in Figure 4.

For Form B, there is a 0.82% gain in correlation for the top models increasing from 5 to 6 item predictors, 0.71% increasing from 6 to 7 predictors, 0.51% moving from 7 to 8, 0.35% moving from 8 to 9, and 0.31% moving from 9 to 10 predictors. Applying the stopping rule strictly for Form B would result in 8 items selected. However, for the purpose of having an equal number of items in each form, this was extended to 9 items for Form B. The correlation of the predicted RSPM total with the actual total for 9 items for Form B is 0.9782. A plot of the actual versus predicted total Raven's for 9 predictors in Form B is shown in Figure 5.

The items included in Form B are 10, 16, 21, 30, 34, 44, 50, 52, and 57 from the original 60-item Raven's and excluding the items included in Form A. The regression coefficients for the top model for each number of items used for prediction in Form B are given in Table 6. The prediction of the total score has

$$\hat{R} = \exp(1.875 + 0.168 * I_{10} + 0.212 * I_{16} + 0.247 * I_{21} + 0.189 * I_{30} + 0.203 * I_{34} + 0.135 * I_{44} + 0.243 * I_{50} + 0.316 * I_{52} + 0.193 * I_{57}),$$

$$V = I_{10} + I_{16} + I_{21} + I_{30} + I_{34} + I_{44} + I_{50} + I_{52} + I_{57},$$

and the total predicted scale total of $\hat{Y} = 60 - (V + \hat{R})$ (\hat{Y} is increased to 0 if negative).

The performance of the top models for each number of items in the development of Form B was assessed for two subgroups, patients with schizophrenia and normal controls. The correlations of the predicted totals with the actual totals for these subgroups are given in Table 4. The correlation for the schizophrenia group for the model construction data set is near $r = .98$, whereas the other correlations considered all remain at about $r = .83$ or more. Thus, predictions from the Form B model also perform well for the subgroups.

The relationships between age, years of education, and sex with each of the RSPM forms considered were evaluated. The Pearson correlations between age and the full 60-item RSPM, the 60-item total RSPM predicted by the 9-item Form A, and the 60-item total RSPM predicted by the 9-item Form B are -0.26 , -0.22 , and -0.14 , respectively. For education, these correlations are $.51$, $.49$, and $.47$, respectively. For males, the mean total scores are 40.3, 40.8, and 40.6, respectively, and for females, the mean total scores are 42.5, 41.7, and 42.8. Thus, these correlations and means are similar for all versions of the test.

Overall, the reduced scales, Form A and Form B, perform well in terms of their correlations with the full 60-item RSPM. To consider the possibility of floor and ceiling effects of the short forms, the Pearson correlations were estimated for tertiles of the full 60-item RSPM, based on the observed values from the 180 participants. The tertiles were 0 to 38, 39 to 50, and 51 to 60. It is important to note

Table 3. Coefficients for Top Models, Form A

Num	Pred																							Intercept
	ravi2	ravi9	ravi10	ravi11	ravi14	ravi17	ravi19	ravi22	ravi24	ravi28	ravi31	ravi36	ravi40	ravi43	ravi48	ravi49	ravi50	ravi51	ravi52	ravi53	ravi55	ravi58		
1										1.025													2.645	
2						0.750													0.813				2.252	
3						0.541													0.612				2.161	
4						0.405											0.468		0.363	0.530			2.173	
5									0.342		0.314		0.285						0.450	0.366			2.000	
6			0.157						0.314		0.272		0.236						0.441	0.366			2.001	
7									0.267		0.291		0.278					0.161	0.465	0.252			1.979	
8				0.159					0.222	0.193				0.352	0.349	0.195				0.512	0.293		1.437	
9				0.198					0.216	0.237		0.142		0.374	0.304	0.178				0.458	0.289		1.323	
10					0.112				0.207	0.196				0.371	0.408	0.215				0.470	0.291	0.306	1.081	
11	0.144			0.094			0.093		0.229	0.154				0.402	0.437	0.230				0.439	0.263	0.264	1.070	
12	0.127			0.131			0.083		0.228	0.194		0.094		0.423	0.403	0.219				0.408	0.268	0.252	0.969	

Table 4. Correlations for the Predicted Total Raven's for Subgroups

Number of Items	Model Construction Sample		Validation Sample	
	<i>n</i> = 31	<i>n</i> = 27	<i>n</i> = 30	<i>n</i> = 28
Form A	Schizophrenia	Normal Control	Schizophrenia	Normal Control
1	0.7096	—	0.8255	0.2684
2	0.7952	0.7869	0.6669	0.6303
3	0.8549	0.8669	0.7090	0.7543
4	0.9186	0.8913	0.8208	0.6669
5	0.9518	0.8372	0.8413	0.7085
6	0.9660	0.8402	0.8370	0.7145
7	0.9637	0.8867	0.8647	0.6874
8	0.9837	0.9324	0.9270	0.8539
9	0.9849	0.9449	0.9274	0.8720
10	0.9876	0.9678	0.9323	0.8746
11	0.9865	0.9717	0.9343	0.8645
12	0.9888	0.9739	0.9367	0.8935
Form B	Schizophrenia	Normal control	Schizophrenia	Normal control
1	0.7096	—	0.8255	0.2684
2	0.7952	0.7869	0.6669	0.6303
3	0.8549	0.8669	0.7090	0.7543
4	0.9186	0.8913	0.8208	0.6669
5	0.9471	0.8023	0.8631	0.7281
6	0.9636	0.8659	0.8635	0.5782
7	0.9664	0.8523	0.8413	0.7457
8	0.9703	0.9083	0.8789	0.6505
9	0.9793	0.8877	0.8651	0.8291
10	0.9837	0.8972	0.8820	0.8513
11	0.9877	0.8981	0.9158	0.8540
12	0.9877	0.9289	0.9057	0.8676

Table 5. Descriptive Statistics for the Predicted Total Raven's Score, Form B

Number of Predictors	Items in Top Model	Model Construction Sample (<i>n</i> = 90)			Validation Sample (<i>n</i> = 90)		
		ρ	ICC	Mean \pm SD (Minimum, Maximum), 40.98 \pm 13.28 (2, 59)	ρ	ICC	Mean \pm SD (Minimum, Maximum), 41.64 \pm 13.11 (9, 59)
1	31	0.7751	0.7516	41.09 \pm 10.23 (20, 46)	0.8164	0.7964	39.36 \pm 11.40 (20, 46)
2	17, 52	0.8832	0.8747	40.97 \pm 11.44 (13, 50)	0.7250	0.7180	41.80 \pm 11.17 (13, 50)
3	17, 50, 52	0.9196	0.9173	40.73 \pm 12.26 (13, 51)	0.7964	0.7876	42.23 \pm 11.20 (13, 51)
4	17, 22, 51, 52	0.9416	0.9409	40.97 \pm 12.61 (8, 51)	0.8508	0.8513	41.49 \pm 12.53 (8, 51)
5	15, 21, 34, 50, 57	0.9550	0.9548	41.04 \pm 12.79 (10, 53)	0.8985	0.8946	41.51 \pm 11.80 (10, 53)
6	14, 22, 31, 40, 51, 52	0.9629	0.9632	41.14 \pm 13.15 (0, 52)	0.8705	0.8627	40.74 \pm 14.94 (0, 52)
7	10, 16, 21, 30, 34, 50, 52	0.9698	0.9700	41.08 \pm 13.12 (7, 53)	0.8699	0.8595	42.76 \pm 11.42 (7, 53)
8	12, 14, 21, 31, 40, 50, 51, 52	0.9748	0.9750	41.10 \pm 13.09 (4, 53)	0.8901	0.8891	41.37 \pm 14.01 (4, 53)
9	10, 16, 21, 30, 34, 44, 50, 52, 57	0.9782	0.9781	40.96 \pm 12.96 (7, 53)	0.8978	0.8884	42.30 \pm 11.35 (7, 53)
10	10, 12, 16, 21, 30, 34, 44, 50, 52, 57	0.9812	0.9812	40.99 \pm 13.05 (8, 54)	0.9077	0.8993	42.32 \pm 11.47 (8, 54)
11	3, 10, 16, 21, 30, 34, 45, 50, 52, 54, 57	0.9833	0.9833	41.06 \pm 13.07 (6, 54)	0.9221	0.9106	42.60 \pm 11.34 (6, 54)
12	10, 12, 16, 21, 30, 34, 40, 44, 50, 51, 52, 57	0.9854	0.9854	40.97 \pm 13.07 (8, 54)	0.9249	0.9211	42.03 \pm 11.93 (8, 54)

Note. ICC= intraclass correlation coefficient.

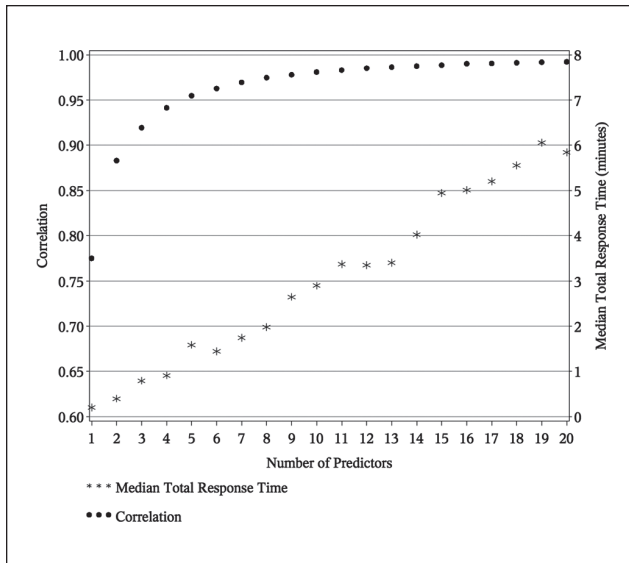


Figure 4. Correlations between actual and predicted total Raven's score, Form B

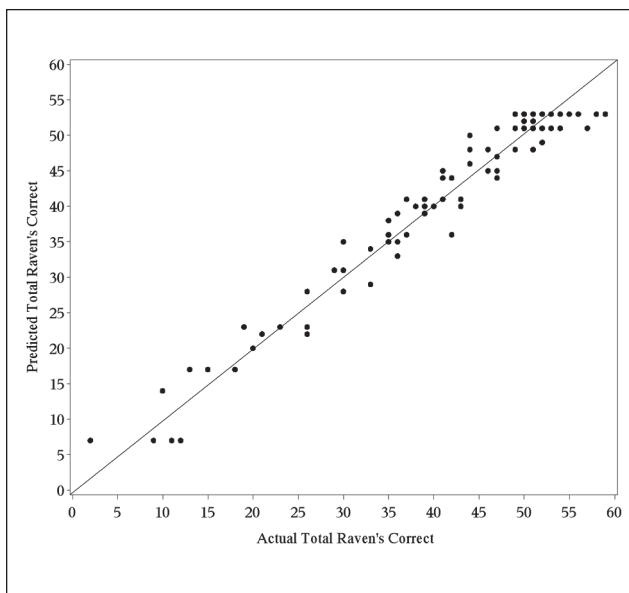


Figure 5. Scatterplot for Raven's Form B

that the correlations within segments of one of the variables being correlated may be smaller than the overall correlation, since local linearity relationships dominate these local correlations, especially for narrow segments, as in this case. The Pearson correlations of the full 60-item RSPM with Form A for the first, second, and third tertiles are .75, .80, and .67, respectively. For Form B, the correlations are .84, .69, and .39. Thus, there does not appear to be a floor effect for either of the reduced forms. There does appear to be a ceiling effect, particularly for Form B, resulting in a reduced correlation in the upper tertile. Since the upper tertile is

composed of a very narrow relatively high range (51 to 60) and the major emphasis of the Ravens test is to look at average to lower-than-average ranges of abilities, this is not a major concern.

The first validation approach was to assess an estimate of the distribution of the drop in correlation between the model construction and validation data sets using a bootstrap approach with 10,000 bootstraps. Considering the final item subset of 9 items from Form A, the mean drop in Pearson correlation was $r = .0761$ ($SD = 0.0239$, 95% confidence interval [CI] = 0.0293, 0.1229). Note that the observed drop for 9 items was $r = .0773$ (.9836-.9063). This provides an estimate of the range of plausible values of the drop in correlation for the specific split used for constructing the Forms A and B presented. Considering the final item subset of 9 items from Form B, the mean drop in Pearson correlation was $r = .0866$ ($SD = 0.0235$, 95% CI = 0.0404, 0.1327). The observed drop for 9 items for Form B was $r = .0804$ (.9782-.8978).

The second validation approach used cross-validation to assess both the final subset of items as well as the corresponding correlations and how they vary across the distribution of possible splits. This included 200 repetitions of the data splitting, and model fitting process, including item selection, and estimation of the correlations for both the model construction and validation data sets from each split. Over the 200 splits for the model construction data set, the mean correlation for Form A was $r = .9795$ ($SD = 0.0026$, 95% CI = 0.9744, 0.9846) and was $r = .9256$ ($SD = 0.0182$, 95% CI = 0.8899, 0.9613) for the corresponding validation data sets. For Form B the mean correlation was $r = .9749$ ($SD = 0.0031$, 95% CI = 0.9688, 0.9810) and was $r = .9205$ ($SD = 0.0199$, 95% CI = 0.8815, 0.9595) for the corresponding validation data sets. Thus, the correlations for the construction data sets for Forms A and B have little variation over random splitting, indicating that the particular split did not strongly impact the correlations. The variation for the validation data sets is larger, as would be anticipated.

Also, for the cross-validation, the 9 items selected for Forms A and B were allowed to vary, allowing assessment of the variation in items selected for each form across random splitting. The results are displayed in Table 7. It was shown from the cross-validation results that the correlations for Forms A and B vary little across random splits. However, this part of the simulation shows that the items included in the final Forms A and B do differ somewhat across the random splits. For example, Item 36 was selected for final Form A but was selected only for 6.5% of the Form A's across the 200 splits. At the other extreme, Item 24 was selected for the final Form A and was selected in 62.0% of the Form A splits. A similar phenomenon occurs for Form B. Since this occurs while the correlation remains stable, this is a strong indication that there are many different subsets of items that could have made up Forms A and B, with very similar correlations. Knowledge of these relationships

Table 6. Coefficients for Top Models, Form B

Num	Pred																					Intercept
	ravi3	ravi10	ravi12	ravi14	ravi15	ravi16	ravi17	ravi21	ravi22	ravi30	ravi31	ravi34	ravi40	ravi44	ravi45	ravi50	ravi51	ravi52	ravi54	ravi57		
1											1.025										2.645	
2						0.750												0.813			2.252	
3						0.541										0.468		0.612			2.161	
4						0.405			0.406								0.363	0.530			2.173	
5					0.366			0.343				0.352				0.391				0.380	1.968	
6				0.228					0.261		0.273		0.235				0.311	0.540			2.133	
7		0.144				0.249		0.274		0.192		0.232				0.278		0.449			2.003	
8			0.130	0.197				0.262			0.159		0.220			0.184	0.238	0.485			1.994	
9		0.168				0.212		0.247		0.189		0.203		0.135		0.243		0.316	0.193		1.875	
10		0.163	0.084			0.197		0.246		0.195		0.189		0.120		0.230		0.327	0.177		1.820	
11	0.096	0.112				0.184		0.233		0.221		0.135			0.153	0.214		0.217	0.232	0.132	1.839	
12		0.140	0.107			0.107		0.245		0.148		0.126	0.125	0.103		0.165	0.145	0.308	0.175		1.806	

Table 7. Cross Validation of Correlations and Item Subset Selection, Forms A and B

Item No.	Form A	Form B	Pct. A	Pct. B	Item No.	Form A	Form B	Pct. A	Pct. B
1			4.0	6.0	31			19.5	28.5
2			4.0	1.0	32			31.0	35.0
3			7.0	7.0	33			13.0	14.5
4			4.0	2.5	34		*	10.5	33.5
5			3.0	8.0	35			12.0	16.0
6			1.0	3.0	36	*		6.5	16.5
7			25.5	10.5	37			11.0	8.0
8			5.5	6.0	38			8.5	11.0
9			3.5	7.5	39			2.0	9.0
10		*	5.5	8.5	40			6.0	6.0
11	*		17.5	15.5	41			10.5	9.0
12			15.0	19.0	42			4.5	7.0
13			6.0	4.5	43	*		26.5	14.5
14			1.5	12.5	44		*	18.0	16.0
15			20.0	7.5	45			28.0	23.0
16		*	7.5	12.0	46			19.5	24.5
17			11.5	11.0	47			11.0	11.0
18			5.5	14.0	48	*		17.5	24.0
19			6.5	13.5	49	*		24.5	17.0
20			7.0	7.5	50		*	16.5	18.0
21		*	28.0	29.0	51			39.5	23.5
22			22.5	21.5	52		*	12.5	21.0
23			7.0	23.5	53	*		46.0	42.0
24	*		62.0	24.5	54			19.0	34.5
25			3.0	1.0	55	*		29.0	33.5
26			5.5	9.0	56			14.0	27.5
27			5.0	2.0	57		*	82.5	14.5
28	*		25.5	24.0	58			10.0	15.0
29			17.5	18.5	59			2.0	2.0
30		*	11.5	14.0	60			0.5	0.5
Column			Description						
Form A			*Indicates that item no. is included in final Form A						
Form B			*Indicates that item no. is included in final Form B						
Percentage A			Percentage of 200 splits that include item in Form A						
Percentage B			Percentage of 200 splits that include item in Form B						

facilitate the understanding of individual item contributions to the score on the full-length measure and can be used to inform item selection when using test reduction strategies based on CTT and IRT.

To compare the results of our computational technique with those from other test reduction strategies, we stepped through conventional CTT and IRT reduction strategies and have highlighted the results. In addition, the comparison includes references to the 30-item RSPM by Wytek et al. (1984), where the research team used both CTT and the Rasch model to inform their item reduction. The following computations and estimates were produced using the full sample of $n = 180$ participants detailed above. Full and short forms will be referenced as Full RSPM (60-item

RSPM), Wytek30 (Wytek 30-item reduced scale), Form A (9-item reduced RSPM scale), and Form B (9-item reduced RSPM scale), respectively.

Beginning with a reduction of items of minimal variance, a difficulty index was computed flagging items with a probability of correct response below .10 and above .90 for removal. Ten items were removed. Next, item-total correlations were inspected. No “identity” items with associations greater than $r = .80$ were found, and two items having correlations below $r = .20$ were dropped. Not surprisingly, of the 12 items dropped, none were found in the subsets of Wytek30, Form A, or Form B.

Beyond the basic identification of item difficulty and total score associations, we employed IRT to estimate both

Table 8. Ravens Standard Progressive Matrices Item and Test Characteristics

Version	Maximum Information, <i>M</i> (SD)	Information Error, <i>M</i> (SD)	Maximum Effectiveness ^a , <i>M</i> (SD)	Loading, <i>M</i> (SD)	Threshold, <i>M</i> (SD)	Reliability Index ^b , <i>M</i> (SD)
Full RSPM	1.05 (0.79)	0.37 (0.29)	0.24 (0.26)	0.72 (0.12)	-0.79 (1.19)	0.27 (0.13)
Wytek et al. Wytek30	1.37 (0.91)	0.44 (0.33)	0.37 (0.29)	0.78 (0.08)	-0.56 (0.68)	0.35 (0.10)
Bilker et al. Form A	1.09 (0.92)	0.34 (0.27)	0.38 (0.37)	0.73 (0.11)	-0.02 (0.77)	0.33 (0.12)
Bilker et al. Form B	1.23 (0.53)	0.38 (0.15)	0.38 (0.26)	0.77 (0.07)	-0.57 (0.74)	0.35 (0.11)

Note. RSPM = Raven's Standard Progressive Matrices.

a. Maximum effectiveness reflects the information conveyed by an item in a population with a normal ability distribution. It is the maximum product of the item's information function and the corresponding normal density function.

b. Reliability = $\sigma^2/(\sigma^2 + MSE)$ and represent actual rather than lower bound estimates of reliability. All values are estimated using our $n = 180$ data set.

Table 9. Ravens Progressive Matrices Reduction Summary

Version	Full/Short Length	Percentage of Item Reduction	Full/Short, α Reliability	Full/Short, Score Correlations	Full/Short, Time ^a	Percentage of Time Saved
Arthur and Day: APM selection sample	36/12	67%	.84/.72	.90	45/NA	NA
Arthur and Day: APM validation sample	36/12	67%	.86/.65	.66	34/15	44%
Bors and Stokes: APM	36/12	67%	.84/.73	.92	45 ^b /NA	NA
Wytek et al.: Wytek30 ^c	60/30	50%	.97/.96	.98	17/9	47%
Bilker et al.: Form A	60/9	85%	.96/.80	.98	17/4	76%
Bilker et al.: Form B	60/9	85%	.96/.83	.98	17/3	82%

Note. APM = Raven's Advanced Progressive Matrices Test; NA = not applicable.

a. Approximate average in minutes.

b. As indicated in administration manual.

c. Correlation and time values were not reported in research article and are computed using values from our $n = 180$ data set.

item and test characteristic for our comparison of short forms and for comparison to the full-length test. Using the Full RSPM (60 items) items both one- and two-parameter logistic (1PL, 2PL) models were fit to the data. A three-parameter model was not considered given that the estimation technique would not produce model estimates with satisfactory error given the limited sample size. Comparing the $-2 \log$ likelihood statistics between the 1PL and 2PL models found a better statistical fit in the latter, $\chi^2(59)$ deviance = 341, $p < .001$. In addition to the χ^2 deviance tests among $-2 \log$ likelihood statistics, the 2PL model enhanced the fit with respect to average slopes ($M_{1PL} = 1.03$, $M_{2PL} = 1.14$), and the number of items with significant misfit statistics ($n_{1PL} = 8$, $n_{2PL} = 2$; Bock, 1972). Item estimates associated with the full and reduced-length tests are presented in Table 8.

Remarkable similarities are found comparing summary statistics of the item estimates among the reduced-length tests. For example, means of maximum information, maximum effectiveness, loading, and reliability are all comparable. Average item error is slightly less for Form A and Form B. Interestingly, the average threshold for Wytek30 and Form B is slightly lower than that of Form A, indicating that the latter is slightly more difficult, though its median difficulty is -0.39 . In addition, threshold variability, the spread of item difficulty, is mostly the same though the

9-item forms appear to have a slightly wider distribution. Overall, comparing these estimates suggest that while our procedure for selecting short-form items differs from those based on CTT and IRT, the resulting item characteristics are quite similar. In fact, when comparing shortened tests in aggregate, Form A and Form B appear to have distinct advantages over the Wytek30. For comparison, summary statistics from full- and reduced-form tests are presented in Table 9.

From Table 9, the greatest detractor to Form A and Form B is their Cronbach's alpha reliability, $r = .80$ and $r = .83$, respectively, compared with $r = .96$ for the Full RSPM. This is to be expected given the limited number of items. However, in comparison with the shortened APM tests including a total of 12 items, our results are notable, especially given that scores based on the 9-item sets have an equal correlation to the full-length test when comparing with a short form of 30 items. This difference of 21 items is perhaps the greatest strength of our technique—time savings. In comparison with the Wytek30, saving 47% of administration time, the item sets Form A and Form B have an approximate savings of 76% to 82%.

The median total response time for the top model (highest correlation) for each number of items considered for Form A (1-20 items) was estimated, and these are presented

Table 10. Categorization of PMAT and Raven Item Content

Rule ^a	Form A	Form B	Wytek30 ^b	Full RSPM
Gestalt Completion: One piece of a repeating pattern or picture is missing.	1 (11%)	1 (11%)	1 (3%)	12 (20%)
Missing Reflection: Symmetry of figure horizontal or vertical.	1 (11%)	2 (22%)	4 (13%)	12 (20%)
Addition or Subtraction: A figure or element from one cell is added or subtracted (juxtaposed/superimposed) from another cell to produce third cell.	2 (22%)	3 (33%)	8 (27%)	12 (20%)
Quantitative Progression: A quantitative increment/decrement to figure occurs between a series (two or three) of sequential cells.	2 (22%)	1 (11%)	4 (13%)	7 (12%)
Movement or Rotation: Figure changes position in cell, rotates, deforms, expands/contracts.	1 (11%)	1 (11%)	5 (17%)	6 (10%)
Distribution of One/Two/Three Values: Distribution of a constant or multiple figural elements presented in a row or column. The distribution of elements is altered with progression over the cells.	2 (22%)	1 (11%)	8 (27%)	11 (18%)

Note. RSPM = Raven's Standard Progressive Matrices.

a. Many items can be classified into multiple categories.

b. Items from Wytek et al., 30-item short form. Several items can be classified into multiple categories. Here items are classified based on the most dominant feature, as judged by the research team.

in Figure 2. Consider the top model with 5 items selected in the development of Form A. These were items 24, 31, 40, 52, and 53 (see Table 2). The total response time for this subset of items used for a reduced test is estimated to be the total of the individual item response times for these 5 items. The total response time for these items are summed for each individual in the model building sample and the median of these total response times was determined. For this case, the estimated median response time is 1.47 minutes. The estimated median response time was determined for each total number of items considered for Form A (Figure 2, right axis) and this process was repeated for Form B (Figure 4, right axis). Clearly, the actual total response time is dependent on the complete sequence of items used. However, this does provide a metric to compare the different number of items in the short forms.

Finally, we would be remiss if our presentation of this reduction methodology did not include a comparison of validity. Issues pertaining to validity are the Achilles heel to all short forms, ours notwithstanding. To address content validity we classified the 60 RSPM items into six abstract reasoning categories. The six general classifications were informed by categories proposed by Carpenter et al. (1990), Deshon, Chan, and Weissbein (1995), Jacobs and Vandeventer (1972) as well as by categories defined for the geometric analogies presented in Mulholland, Pellegrino, and Glaser (1980). We then counted the items in each category for the Full RSPM, Wytek30, Form A and Form B (see Table 10).

Content representation of items in Form A, Form B, and Wytek30 were all in similar proportions with the Full RSPM for each taxonomic category. The short forms had a slightly smaller proportion of items in the category of Gestalt Completion; however, this was considered acceptable as these items are nearly self-evident, providing a practice or familiarity experience for the participant (Raven,

& Court, 2000, 2003). Although an argument can be made that a single item will not represent sufficiently the entire content domain, some support across these classifications of abstract reasoning is found in score correlations among the six groups ranging from $r = .57$ to $r = .82$. These substantial correlations reflect the fact that many items require multiple types of abstract reasoning to obtain the correct solution.

Discussion

Two nine-item short forms were derived from the full set of items of the Raven's Standard Progressive Matrices Test. Each of the short forms, A and B, had correlations to the long form of $r = .9836$ and $r = .9782$, respectively. Both forms performed well in subgroups of patients with schizophrenia and healthy adults. Additionally, each of these reduced scales fared well in the validation approaches used for these predictive models. The reduction from 60 to 9 items represents an 85% reduction in the number of items to be administered.

Importantly, psychometric properties associated with two 9-item short-form Raven's tests were found to be comparable with those of the full-length Raven's Progressive Matrices test as well as with another short-form test using a reduced set of 30 items. Aside from test-level qualities, item-level characteristics of the 9-item forms were comparable with those found in the 30-item form, and except for the easiest items, all were representative of item characteristics found in the 60-item test. The two 9-item forms were distinctly different from the 30-item form in the vast time savings, >75%, for administration while preserving equal correlations to full-length scores. Finally, like all short-form instruments both 9-item forms had a reduced number of items available to represent the six general categories of abstract reasoning. Both forms A and B, however, included

a similar proportion of content as the full 60-item test. Further support for the content validity of these reduced-item tests is found with an average correlation of $r = .71$ across reasoning domains where multiple categories of abstract reasoning are often present in each item.

A new methodological approach was developed for the reduction of the RSPM from 60 items to the resulting 9 items for each of two forms, A and B. The approach differs from that used for the reduction of clinical rating scales such as the QLS in multiple ways. First, a Poisson regression model was considered to allow for the positive integer nature of the outcome, the scale total. The number of incorrect items among those selected was modeled by a Poisson distribution, due to the negative skew of the distribution of the number correct (scale total). The predicted number incorrect was then used to obtain the predicted number correct. This permitted application of a Poisson regression modeling approach to predict the actual scale total, the number correct. To accommodate the larger number of items, the algorithm developed for this reduction allowed consideration of a small fraction of the total number of possible combinations of items. When this modified algorithm for the RSPM reduction is applied to the data used for the QLS reduction, the exact same top models are obtained as from the consideration of all possible combinations, which was not possible for the 60 items of the RSPM. This approach can be applied to other time consuming behavioral assessment instruments.

The present methodology has several limitations. Importantly at present, it requires considerable computational resources. Even after the algorithm for narrowing down the number of models to be tested, implementation of the present analysis, including the validation analyses, required nearly 2 full weeks of CPU time on our rather advanced systems. This difficulty will be ameliorated with advanced CPU speeds, but already now the time savings in large-scale studies from using our methods are substantial. Another limitation of the method is that it can accentuate floor and ceiling effects inherent in the original test. In our case, there appeared to be some ceiling effects more pronounced for Form B. This problem could be resolved by including a harder item that is not part of the RSPM. The disadvantage of adding a harder item is that it could exacerbate the frustration of poorly performing individuals, which is of special concern in clinical populations. Application of adaptive testing using IRT could address this concern. Finally, the split data set of 180 participants, with 90 participants in the model construction phase, is not large. In future studies, it would be valuable to compare these results with those based on a larger sample. However, the results from the detailed validations, as well as assessment of the psychometric properties of both of the 9-item forms compared with the full 60-item scale, indicate very good performance of the reduced scales.

Using the technique put forth in this article researchers can produce shortened sets of items that will result in substantial savings in time while preserving item and test characteristics represented in the full-length test. Indeed, the time savings can make a difference in the decision to include such a measure in large-scale genomic studies, where cognition is not the primary target, but where a measure similar to RSPM could arguably help in assessing the impact of particular disorders. For example, a genomic study of risk for diabetes may include thousands of participants for whom blood samples are collected and information is obtained on dietary and lifestyle habits. Every minute added to data collection adds about 17 hours of testing for 1,000 subjects. The proposed methodology could be applied to shortening other traditional tests so as to expand the range of measures of cognitive abilities that can be incorporated in large-scale genomic or treatment studies. Such efforts are essential if psychological assessment is to be included in the genomics revolution that is overtaking biomedical research.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:

This work was supported by NIH Grants P50-MH064045 and R01-MH084856.

References

- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.) Washington, DC: Author.
- Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, 54, 394-403.
- Bilker, W. B., Brensinger, C., Kurtz, M. M., Kohler, C., Gur, R. C., Siegel, S. J., & Gur, R. E. (2003). Development of an Abbreviated Schizophrenia Quality of Life Scale using a new method. *Neuropsychopharmacology*, 28, 773-777.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.
- 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.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97, 404-431.

- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1-22.
- Deshon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence*, 21, 135-155.
- Gur, R. C., Ragland, J. D., Moberg, P. J., Bilker, W. B., Kohler, C., Siegel, S. J., & Gur, R. E. (2001). Computerized neurocognitive scanning: II. The profile of schizophrenia. *Neuropsychopharmacology*, 25, 777-788.
- Gur, R. C., Ragland, J. D., Moberg, P. J., Turner, T. H., Bilker, W. B., Kohler, C., . . . Gur, R. E. (2001). Computerized neurocognitive scanning: I. Methodology and validation in healthy people. *Neuropsychopharmacology*, 25, 766-776.
- 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.
- Jacobs, P. I., & Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement*, 32, 235-248.
- Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12, 252-284.
- Nurnberger, J. I., Blehar, M. C., Kaufmann, C. A., York-Cooler, C., Simpson, S. G., Harkavy-Friedman, J., . . . Reich, T. (1994). Diagnostic interview for genetic studies. Rationale, unique features, and training. NIMH Genetics Initiative. *Archives of General Psychiatry*, 51, 849-859.
- Raven, J. C. (1936). *Mental tests used in genetic studies: The performances of related individuals in tests mainly educative and mainly reproductive* (Unpublished master's thesis). University of London, England.
- Raven, J. C. (1938). *Raven's progressive matrices (1938): Sets A, B, C, D, E*. Melbourne, Victoria, Australia: Australian Council for Educational Research.
- Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement*, 21(1), 1-16.
- Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, 41, 1-48.
- Raven, J., Raven, J. C., & Court, J. H. (1993). *Raven manual section 1: General overview*. Oxford, England: Oxford Psychologists Press.
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual section 4: Advanced progressive matrices*. Oxford, England: Oxford Psychologists Press.
- Raven, J., Raven, J. C., & Court, J. H. (2000). *Standard progressive matrices*. London, England: Psychology Press.
- Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual section 1. General overview*. San Antonio, TX: Pearson Assessment.
- Williams, J., & McCord, D. (2006). Equivalence of standard and computerized versions of Raven Progressive Matrices Test. *Computers in Human Behavior*, 22, 791-800.
- Wytek, R., Opgenoorth, E., & Presslich, O. (1984). Development of a new shortened version of Raven's Matrices Test for application and rough assessment of present intellectual capacity within psychopathological investigation. *Psychopathology*, 17, 49-58.