

2020

Measuring Intelligence with the Sandia Matrices: Psychometric Review and Recommendations for Free Raven-Like Item Sets

Alexandra M. Harris
University of Georgia; Northwestern University


Jeremiah T. McMillan
University of Georgia

Benjamin Listyg
University of Georgia

Laura E. Matzen
Sandia National Laboratories

Nathan Carter
University of Georgia

Follow this and additional works at: <https://scholarworks.bgsu.edu/pad>

 Part of the [Cognitive Psychology Commons](#), [Industrial and Organizational Psychology Commons](#), [Other Psychology Commons](#), and the [Quantitative Psychology Commons](#)

Recommended Citation

Harris, Alexandra M.; McMillan, Jeremiah T.; Listyg, Benjamin; Matzen, Laura E.; and Carter, Nathan (2020) "Measuring Intelligence with the Sandia Matrices: Psychometric Review and Recommendations for Free Raven-Like Item Sets," *Personnel Assessment and Decisions*: Vol. 6 : Iss. 3 , Article 6.

DOI: <https://doi.org/10.25035/pad.2020.03.006>

Available at: <https://scholarworks.bgsu.edu/pad/vol6/iss3/6>

This Measurement and Measures is brought to you for free and open access by the Journals at ScholarWorks@BGSU. It has been accepted for inclusion in Personnel Assessment and Decisions by an authorized editor of ScholarWorks@BGSU.

MEASURING INTELLIGENCE WITH THE SANDIA MATRICES: PSYCHOMETRIC REVIEW AND RECOMMENDATIONS FOR FREE RAVEN-LIKE ITEM SETS

Alexandra M. Harris^{1,2}, Jeremiah T. McMillan², Benjamin Listyg²,
Laura E. Matzen³, and Nathan Carter²

1. Northwestern University
2. University of Georgia
3. Sandia National Laboratories

ABSTRACT

KEYWORDS

intelligence, Raven's Progressive Matrices, Sandia Matrices

The Sandia Matrices are a free alternative to the Raven's Progressive Matrices (RPMs). This study offers a psychometric review of Sandia Matrices items focused on two of the most commonly investigated issues regarding the RPMs: (a) dimensionality and (b) sex differences. Model-data fit of three alternative factor structures are compared using confirmatory multidimensional item response theory (IRT) analyses, and measurement equivalence analyses are conducted to evaluate potential sex bias. Although results are somewhat inconclusive regarding factor structure, results do not show evidence of bias or mean differences by sex. Finally, although the Sandia Matrices software can generate infinite items, editing and validating items may be infeasible for many researchers. To aid implementation of the Sandia Matrices, we provide scoring materials for two brief static tests and a computer adaptive test. Implications and suggestions for future research using the Sandia Matrices are discussed.

The Raven's Progressive Matrices (RPMs; Raven et al., 1998) are widely used measures of analytical intelligence (Arthur Jr. & Woehr, 1993) in part because they are nonverbal. The RPMs¹ are matrix completion problems that require participants to solve patterns among objects. The *Sandia Matrices* are software-generated matrix completion problems designed to function similarly to the RPMs (Matzen et al., 2010). Although the Sandia Matrices software (Benz & Dixon, 2010) was created to remedy the limited number of RPMs, another advantage over the RPMs is that Matzen and colleagues have made the software—including

a large bank of pre-generated items—available for free (<https://github.com/LauraMatzen/Matrices>). Because the RPMs are proprietary, they are often cost prohibitive for researchers. Consequently, the Sandia Matrices are likely to have some of the advantages of the RPMs without its monetary disadvantages.

Matzen et al. (2010) provided an extensive review of the Sandia Matrices development and item properties relative to the RPMs in their introductory norming study. However, an in-depth psychometric review is needed prior to widespread implementation. The first aim of this study is to provide a psychometric review of select Sandia Matrices items. We begin by using item response theory (IRT) to review item parameters and screen for potentially problematic items.² Then, given the intended similarity between the Sandia Matrices and RPMs, we briefly review issues historically evaluated in the RPMs: (a) dimensionality and (b) potential sex differences, including measurement bias and

1 Although some researchers may use the RPMs as a generic term to refer to matrix-type problems generally, we use RPMs in the current paper to refer explicitly to the branded, proprietary Raven's tests. We use "matrix-type" to refer to problems that use a matrix-type format but are not a specific Raven's test.

2 Relative to traditional psychometric approaches to test construction (e.g., classical test theory), IRT offers a number of advantages such as more precise reliability estimates, more readily interpretable difficulty parameters, and sample independence such that item parameters can be used to estimate latent trait scores in new samples (see Reise & Henson, 2003 and Zickar & Broadfoot, 2009 for further review of the benefits of IRT).

Corresponding author:
Alexandra M. Harris
Email: alexandra.harris@northwestern.edu

score (i.e., trait estimate) differences.

Additionally, although the Sandia Matrices software can generate infinite combinations of items, the time involved in generating and curating new items may hinder implementation for many researchers. Matzen et al. (2010) acknowledge many items generated for the norming study required manual alterations to ensure appropriate distractors. As such, the second aim of this study is to identify appropriate sets of pregenerated stimuli and provide corresponding scoring information, including the IRT parameter estimates from our psychometric review. Although other free matrix-type cognitive ability measures similar to the RPMs exist (e.g., International Cognitive Ability Resource Team, 2014), IRT-based psychometric information is rarely available. By using the items recommended here and the associated parameters, researchers who lack the resources to generate and curate new items or the sample sizes necessary for IRT scoring can still benefit from the enhanced precision of IRT estimates.

Thus, we culminate our psychometric review of Sandia Matrices items by recommending two 10-item sets that can be administered in a paper-and-pencil format. Additionally, due to the utility and efficiency of computerized adaptive testing for psychological assessment (van der Linden & Glas, 2010), we provide code for administering a computer adaptive test (CAT). Materials for both the 10-item sets and the CAT are provided such that researchers can administer and score the Sandia Matrices for as few as a single participant. Finally, we report mean raw scores (i.e., proportion correct) and standard deviations for the final items sets so that researchers who do not wish to use IRT parameter estimates can calculate standardized scores.

Dimensionality

One of the commonly debated properties of the RPMs is their dimensionality. Although the RPMs are intended as a unidimensional measure of analytical intelligence, some researchers have proposed that they also assess visuospatial abilities and are therefore two dimensional (e.g., Dillon et al., 1981). This argument stems from a taxonomy that separates the rules underpinning RPM solutions into verbal-analytic and visuospatial-based strategies (Carpenter et al., 1990; DeShon et al., 1995). Despite its popularity, support for a two-dimensional structure driven by distinct cognitive processes has been weak thus far (Vigneau & Bors, 2008; Waschl et al., 2016).

A primary reason that researchers are concerned about the influence of visuospatial processing on the RPMs is that men generally demonstrate advantages over women in spatial ability tasks (Voyer et al., 1995). Some researchers have proposed that group differences between men and women on the RPMs (Lynn & Irwing, 2004) are attributable to those items that invoke visuospatial processes (Colom et al., 2004). Thus, despite inconclusive evidence regarding

the factor structure of the RPMs, we evaluate the factor structure of the Sandia Matrices in order to better understand potential sex differences.

Relative to the RPMs, the Sandia Matrices utilize a narrower set of rules to inform solutions but can still be mapped on to the taxonomies used to distinguish RPM rules. The Sandia Matrices include object relation (OR) and logic items (see Figure 1). OR problems involve simple transformations (e.g., shape, shading, orientation) across the matrix and are subdivided by the number of transformations that participants must track (one, two, or three relations; here called OR-1, OR-2, and OR-3 respectively). In contrast, logic problems involve conjunction and disjunction rules. Table 1 summarizes approximately how these transformations correspond to four rules defined by DeShon et al. (1995).

Given the debate regarding a two-factor structure in the RPMs, it is possible that the strategies used to solve Sandia Matrices problems would similarly produce a two-factor structure. However, in the Sandia Matrices, the two types of processes (i.e., verbal analytic vs. visuospatial) roughly correspond to the two problem types (i.e., OR and logic). According to DeShon et al.'s taxonomy, all of the Sandia Matrices logic problems involve visuospatial processing, whereas OR problems involve primarily verbal-analytic processing unless they include rotation. Consequently, evidence of a two-factor structure for the Sandia Matrices may stem from a distinction between either underlying visuospatial and verbal-analytic processes or simply the two problem types. Thus, this study evaluates the dimensionality of the Sandia Matrices by considering a unidimensional model as well as alternative two-dimensional models: visuospatial versus verbal-analytic processing and OR versus logic problems.

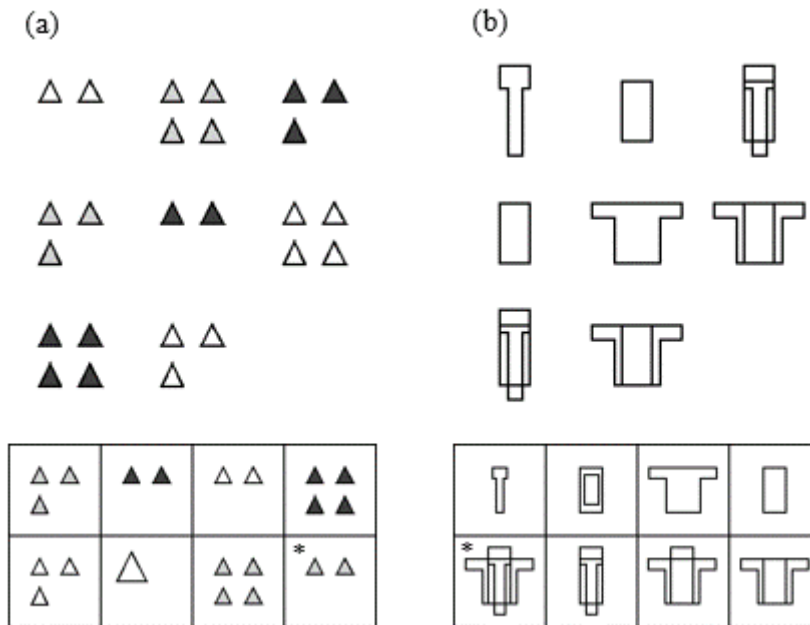
Sex Differences

As mentioned above, one of the primary reasons for investigating a two-dimensional visuospatial versus verbal-analytic structure is gender difference implications. Gender differences may manifest either as bias in item parameters such that men and women with the same trait scores show different likelihoods of answering correctly (i.e., measurement bias) or score differences even after accounting for biased items (i.e., trait estimate differences). Although some researchers have found no evidence of sex bias in the RPMs (Waschl et al., 2016), others suggest that sex differences on the RPMs persist even after accounting for measurement bias in items that invoke spatial processing (Abad et al., 2004). Thus, we evaluate gender differences by first conducting measurement equivalence (ME) analyses (Drasgow, 1984) to determine whether item parameters are different between groups (i.e., show bias). After accounting for potential bias, we compare trait estimates across genders.

TABLE 1.
Correspondence Between Sandia Matrices Transformation Types and DeShon et al. (1995)'s Taxonomy

Problem type	Transformation types	DeShon et al. (1995) rules
Object relation (OR)	Shape, shading, orientation	Distribution of three, constant in a row, quantitative pairwise progression, rotation
Logic	Conjunction, disjunction	Superimposition, superimposition with cancellation

FIGURE 1.
Sandia Matrices Item Types: (a) Object Relations and (b) Logic



Note. *Correct answer.

METHOD

Participants

Sample 1 participants were workers on Amazon Mechanical Turk ($N = 1,276$, $M_{age} = 34.56$, $SD_{age} = 11.47$, 66.9% female, 82.2% White). Sample 2 participants were undergraduates at a large university in the southeastern United States ($N = 338$; $M_{age} = 19.18$, $SD_{age} = 1.47$, 65.4% female, 77.8% White). Participants completed an online questionnaire including Sandia Matrices and demographic items. These final samples include only participants who passed a variety of attention check items and took longer than 5 minutes to complete the survey.

Sandia Matrices

In their norming study, Matzen et al. (2010) found that Sandia Matrices item types advance in difficulty from OR-1 to logic problems. To target a range of difficulties, in

Sample 1 we administered 5 items of each Sandia Matrices item subtype (OR-1, OR-2, OR-3, and logic) for a total of 20 items. Within each item type, we further selected items according to the proportion correct (i.e., 0%, 25%, 50%, 75%, 100%) as reported by Matzen et al. (2010). In Sample 2, we selected additional OR-3 and logic items to increase the number of items expected to show moderate to high difficulty for a 25 total items. Participants completed all items in both samples. All Sandia Matrices items include eight response options, and all items were selected from those included in Matzen et al.'s (2010) norming study. Items were coded according to type (OR vs. logic) and whether the rules used to solve them required visuospatial, verbal-analytic, or both processes. Responses were coded as correct or incorrect.

Data Analysis

Confirmatory multidimensional IRT. Confirmatory

multidimensional item response theory (CMIRT) analyses were conducted using the R package “mirt” (Chalmers, 2012) to compare three possible factor structures: unidimensional; visuospatial versus verbal-analytic processing (two-factor); OR versus logic problems (two-factor). OR items that utilized strategies thought to invoke both verbal-analytic and visuospatial processes were set to load onto both factors.

Sex differences. ME analyses were also conducted using the R package “mirt.” To conduct ME analyses (Drasgow, 1984; Stark et al., 2006), we compared the fit of three models for the alternative two-factor structures in each sample: a fully freed model in which all item parameters, means, and factor correlations were allowed to vary between genders; a partially constrained model in which item parameters were constrained but means and factor correlations were allowed to vary across genders; and finally a fully constrained model in which all item parameters, means, and factor correlations were set to be equal across genders (i.e., a one group model). Improved fit of the partially constrained model relative to the fully freed model would suggest the Sandia Matrices do not show measurement bias, and improved fit of the fully constrained relative to the partially constrained model would further suggest the Sandia Matrices do not show structural or trait estimate differences across genders.

Recommended item sets. Before determining which items to include in our static test sets and CAT item bank, we conducted multiple groups analysis to test for meaningful group differences in means or item parameters between the two samples. We then considered item parameters as estimated by the final models to select items for recommendation. In the two 10-item sets, we selected items to represent a range of difficulty (*b*) parameters and approximately balance item types between the two sets.

RESULTS

Because to our knowledge no prior studies have conducted a model-based psychometric evaluation of Sandia Matrices items, we first reviewed model-data fit for unidimensional models and reviewed all items for problematic properties. Both the 2-parameter logistic model (2PLM) and 3-parameter logistic model (3PLM) are appropriate for dichotomously scored multiple choice data. The 2PLM includes two item parameters: the item discrimination, *a*, which is conceptually similar to factor loadings; and the item difficulty, *b*, which is defined as the trait level at which persons have a .50 probability of getting the item correct (Embretson & Reise, 2000). The 3PLM includes these parameters as well as a third that accounts for “guessing” (Waller, 1989), *c*, which is the probability of a correct answer for a person with infinitely low ability. A number of items demonstrated substantial guessing parameters, which suggests that guessing is a concern. Thus, to account for items with large guessing parameters, we chose to proceed

with the 3PLM for the purposes of psychometric review.

Global fit statistics were calculated using the M_2 statistic to derive global absolute (i.e., RMSEA) and relative (i.e., TLI and CFI) model-data fit statistics. Item fit was determined using MODFIT (Stark, 2001) to calculate χ^2/df ratios, where good model-data fit is indicated by χ^2/df ratios less than 3 (Drasgow et al., 1995). The 3PLM showed acceptable model-data fit in both Sample 1 ($M_2(150)=353.30$, $p < .001$; RMSEA = .03; TLI = .96; CFI = .97; χ^2/df ratio < 3 for all items) and Sample 2 ($M_2(250)=376.04$, $p < .001$; RMSEA = .04; TLI = .93; CFI = .94; χ^2/df ratio < 3 for all items).

Although model-data fit was acceptable overall, some items exhibited extreme item parameters that warranted further review. In Sample 1, items B5_1, B5C1, and D2E4 exhibited discrimination parameters of 7.61, 6.99, and 28.84 respectively. Such high discrimination parameters suggest that responses to these items may be similarly influenced by factors other than cognitive ability (i.e., demonstrate local dependence). Closer evaluation of these items revealed that all three utilized the same shading progression strategy (Figure 2). For each item, the darkest shape was the correct answer, yet a large proportion of respondents chose the lightest shape. This pattern suggests that many respondents thought the problem followed a symmetry rule or repeated. The elimination of these three items resulted in a 17 final items for Sample 1.

In Sample 2, item A3D4E1 showed a relatively high guessing parameter of 0.261. Because the Sandia Matrices include eight response options, we would expect guessing parameters to be approximately at or below .125. Thus, a high value suggests that participants with very low cognitive ability could guess the correct answer to item A3D4E1 at a rate greater than expected by chance. Closer review revealed that this item had only two competitive distractors, which were identical except for the size of the stimuli. Because some participants may have had difficulty discerning the differences in stimuli sizes for reasons other than intelligence level (e.g., size of electronic screen), we chose to eliminate this item from further analyses. Although other items in Sample 2 also showed somewhat high parameter estimates, we discerned no obvious content-related reasons. Eliminating item A3D4E1 resulted in 24 final items for Sample 2.

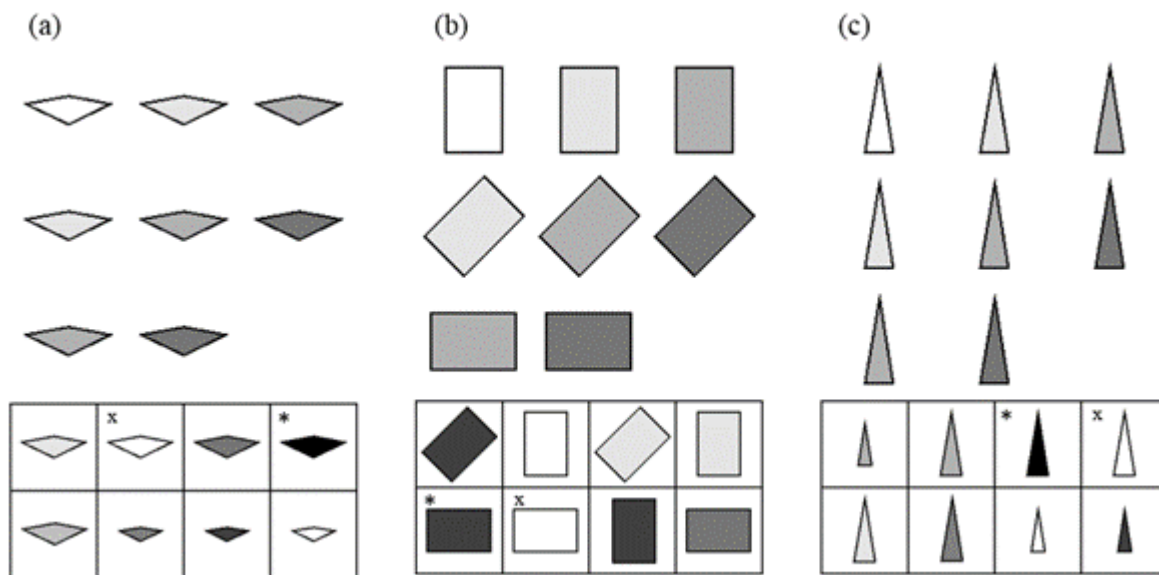
Table 2 displays descriptive statistics for each sample, including coefficient alpha after removing problematic items. The unidimensional 3PLM showed acceptable model-data fit (see Table 3) and χ^2/df ratio < 3 for all remaining items in both samples.

Dimensionality

Table 3 presents fit statistics for the three factor structures evaluated using confirmatory CMIRT analyses. In both samples, Akaike’s information criterion (AIC) and Bayesian information criterion (BIC) show larger values for the unidimensional model than either the visuospatial

FIGURE 2.

Items Eliminated Due to Extreme Discrimination Parameters: (a) B5_1, (b) B5C1, (c) B5_3



Note. *Correct answer. xIncorrect, commonly selected answer.

TABLE 2.

Descriptive and Reliability Statistics for Items Included in Analyses

Sample	<i>k</i>	<i>N</i>	<i>M</i>	<i>SD</i>	α
Sample 1	17	1,276	0.63	0.16	.68
Sample 2	24	338	0.64	0.16	.77

Note. *k* = number of items.

TABLE 3.

Model-Data Fit Statistics for Tested Factor Structures

Model	AIC	BIC	M_2	<i>df</i>	<i>p</i>	RMSEA	RMSEA 95% CI		TLI	CFI	<i>r</i>
							Lower	Upper			
Sample 1 (N = 1,276)											
Unidimensional	19084.12	19346.85	151.29	102	.001	.019	.012	.026	.984	.988	-
Visuospatial vs. verbal-analytic	19067.77	19345.95	150.35	99	<.001	.020	.013	.026	.983	.988	.80
Object-relation vs. logic	19070.10	19337.97	153.98	101	<.001	.020	.013	.026	.983	.987	.84
Sample 2 (N = 338)											
Unidimensional	7405.34	7680.59	355.13	228	<.001	.041	.032	.049	.928	.941	-
Visuospatial vs. verbal-analytic	7322.39	7620.59	250.44	222	<.001	.019	.000	.031	.984	.987	.82
Object-relation vs. Logic	7340.33	7619.41	274.02	227	<.001	.025	.011	.035	.974	.978	.63

Note. *r* = factor correlations.

versus verbal-analytic model or the OR versus logic model. However, for both samples, AIC was lower for the visuospatial versus verbal-analytic model, and BIC was lower for the object-relation versus logic model. Moreover, RMSEA confidence intervals of all three models were nearly identical in Sample 1 and overlapped between the alternative two-factor models in Sample 2. Thus, although there is some evidence that a two-factor structure fits better than a one-factor structure, *which* two-factor structure is not clear. It is possible that the improved fit of a two-factor structure relative to a one factor is attributable to the distinction between OR and logic item types as opposed to the underlying processing strategies.

Sex Differences

As noted above, one of the primary reasons a two-factor structure is a concern for the Sandia Matrices is because a verbal-analytic versus visuospatial distinction might suggest sex differences. To evaluate whether these factor structures might impact sex differences (i.e., whether any of the factors exhibited evidence of measurement bias or trait estimate differences), we conducted ME analyses. In all cases, the fully constrained (i.e., one group) model fit better than the models in which parameters were free to vary across genders (see Table 4). Because results did not point to a clear two-factor structure that could not be explained by simple differences in item types, nor was their evidence that the factors had meaningful consequences for measurement bias or score differences by gender, we chose to proceed using a unidimensional model for the remainder of our analyses.³

Finally, to determine whether there were sex differences in Sandia Matrices scores derived using the unidimensional 3PLM model or raw scores, a t-test was performed for using latent trait scores and proportion correct in both samples. Results are shown in Table 5. No significant sex differences were found in either sample, regardless of scoring approach.

Recommended Item Sets

Before determining our recommended item sets, we conducted multiple groups analysis to determine whether there were any meaningful differences in group means or item parameters between our two samples. We first omitted one item that persisted in showing an extreme discrimination parameter in the final unidimensional model for Sample 2 (Y_{11} , $a = 5.24$). Removal of this item resulted in a final item set of 26 items. Next, we compared the fit of a 2PLM and 3PLM model for our combined samples. AIC and BIC support fit of the 2PLM (AIC = 24412.62; BIC =

24692.72) relative to the 3PLM (AIC = 24423.52; BIC = 24843.66). Additionally, when estimated with the 3PLM, over one-fourth of items exhibited discrimination parameters over 4.0, which indicates overfitting. Given evidence of overfitting with the 3PLM and inconsistent support for either the 2PLM or 3PLM, we proceeded with the 2PLM for all remaining analyses.

To conduct multiple groups analysis, we compared a model in which group means and item parameters were allowed to freely vary between samples (i.e., fully freed baseline model) with the fully constrained 2PLM. AIC and BIC support fit of the constrained model (fit reported above) relative to the fully freed model (AIC = 24477.33; BIC = 25032.14). Thus, we proceeded with a model that utilized both samples as a single group ($N = 1,614$).

Items were selected such that each recommended 10-item set would reflect a range of difficulty (b) parameters and that the proportion of each item type would be similar between the two sets. Further, we avoided selecting items with particularly high discrimination (a) parameters (e.g., above 2.0) to avoid overly weighting any one item. Empirical reliability for the full 26 items was .70, and empirical reliability for both 10-item measures was above .95. Figure 3 illustrates the test information (TIF) for the full 26 items, and Figure 4 illustrates TIFs for the two 10-item measures. Table 6 includes parameter estimates for all 26 items estimated across samples, an indication of item-set assignment for each of the 10-item measures, as well as mean raw scores and standard deviations for all item sets.

Additionally, we aimed to construct a CAT that researchers could use to efficiently assess intelligence in just a few items. CAT is an iterative assessment procedure whereby item locations are matched as closely as possible to respondent ability levels. The standard process in a CAT is to start by assuming an individual has average/moderate ability, present a single item, update the estimate of ability based upon the respondent's response and a given response model, select and present the next item that maximizes information at that ability level, and so on until the termination criterion has been reached (i.e., a set length or a set standard error of measurement). CATs provide maximum utility when there are many candidate items that can be matched precisely to any estimated ability level (i.e., items span a wide range of locations; Flaugher, 2000). All 26 items were included in the item bank for CAT items. All 26 item stimuli as well as R code for scoring the 10-item measures from participant responses and administering the CAT are included in the online supplementary materials.

DISCUSSION

This study aimed to provide the first modern psychometric review of the Sandia Matrices as well as to recommend two 10-item measures and construct a CAT for use by researchers. Specifically, we reviewed two psychometric issues historically evaluated in the RPMs: dimensionality

3 To evaluate potential measurement bias at the item-level we also conducted differential item functioning (DIF) analyses (Meade & Lautenschlager, 2004). No more than one item demonstrated possible evidence of DIF in any sample (i.e., less than the 10% that would be expected due to type 1 error using a liberal significance criterion of .10).

TABLE 4.

Model-Data Fit Statistics Two-Dimensional Factor Structures Across Genders

Model	AIC	BIC	M_2	df	p	RMSEA	RMSEA 95% CI		TLI	CFI
							Lower	Upper		
Sample 1 (N = 1,255)										
Visuospatial vs. verbal-analytic										
Fully freed baseline	18735.52	19279.82	225.19	200	.107	.010	.000	.016	.992	.994
Partially constrained	18716.57	19086.29	261.84	234	.102	.010	.000	.016	.992	.993
Fully constrained	18695.37	18962.38	149.79	101	.001	.020	.013	.026	.984	.988
Object-relation vs. logic										
Fully freed baseline	18734.95	19299.79	220.07	196	.115	.010	.000	.016	.992	.994
Partially constrained	18713.68	19093.66	259.26	232	.106	.010	.000	.016	.992	.993
Fully constrained	18693.88	18971.16	149.77	99	.001	.020	.013	.027	.984	.988
Sample 2 (N = 338)										
Visuospatial vs. verbal-analytic										
Fully freed baseline	7399.83	7965.64	479.04	452	.183	.013	.000	.023	.985	.988
Partially constrained	7369.09	7965.64	526.67	500	.198	.013	.000	.023	.986	.988
Fully constrained	7340.33	7751.39	274.02	227	.018	.025	.011	.035	.974	.978
Object-relation vs. logic										
Fully freed baseline	7385.20	7989.24	468.23	442	.187	.013	.000	.023	.985	.988
Partially constrained	7350.56	7751.98	512.38	495	.285	.010	.000	.021	.991	.992
Fully constrained	7322.39	7620.59	250.44	222	.092	.019	.000	.031	.984	.987

Note. Fully freed baseline model: item loadings, item thresholds, means, and trait correlations allowed to vary across genders. Partially constrained model: item loadings and item loadings constrained; means and trait correlations allowed to vary. Fully constrained: item loadings, item thresholds, means, and trait correlations constrained (i.e., single group model). In Sample 1, 12 participants did not report gender yielding a total sample size of 1,255 for gender analyses.

TABLE 5.

Sex Differences in Sandia Matrices Scores

Score type	Male		Female		t -test
	M	SD	M	SD	
Sample 1	$(n = 401)$		$(n = 854)$		
Proportion correct	0.62	0.18	0.63	0.15	$t(1253) = -1.00, p = .319$
Latent trait estimates	-0.03	0.96	0.01	0.78	$t(1253) = -0.84, p = .404$
Sample 2	$(n = 117)$		$(n = 221)$		
Proportion correct	0.65	0.18	0.63	0.15	$t(336) = 0.77, p = .440$
Latent trait estimates	0.03	1.02	-0.01	0.85	$t(336) = 0.40, p = .688$

Note. In Sample 1, 12 participants did not report gender yielding a total sample size of 1,255 for gender analyses.

TABLE 6.

Final Item Parameters Estimated in Multiple Groups Analysis and 10-Item Set Assignment

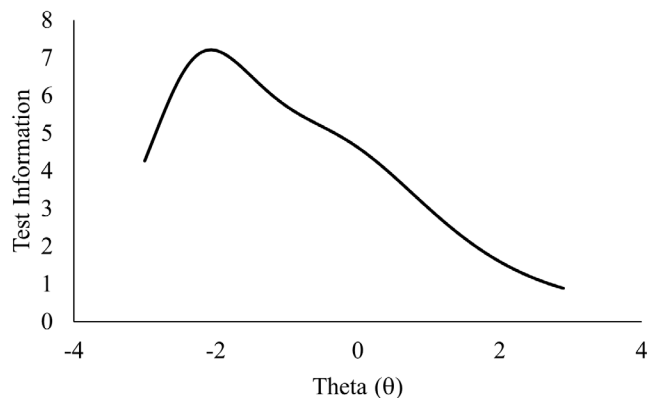
Item #	Name	Type	Subtype	<i>a</i>	<i>b</i>	Recommended 10-item sets		Correct answer
						Test 1	Test 2	
1	A4_1	1R	Shape	2.416	-2.386			2
2	A1B4C2	3R	Shading	1.995	-2.306	1		8
3	E4_2	1R	Number	1.623	-2.273		1	7
4	D4_2	1R	Size	1.618	-2.222			5
5	B4D1E2	3R	Shading	0.655	-2.057			5
6	B3E4	2R	Shading and number	1.718	-1.881	2		8
7	B2D3	2R	Shading and size	1.282	-1.852		2	5
8	A4D1E2_2	3R	Shape	1.727	-1.784			2
9	A3E2	2R	Shape and number	1.322	-1.726	3		3
10	B1D3E4	3R	Size and number	0.459	-1.591			1
11	D2E4	2R	Size and number	1.059	-1.215		3	8
12	X_9	Logic	OR	1.547	-0.520		4	5
13	A3C4E5_1	3R	Shape, orientation, and number	1.015	-0.460			3
14	Z_9	Logic	XOR	1.508	-0.442	4		5
15	A2C5D4	3R	Orientation and size	0.904	-0.184		5	3
16	X_5	Logic	OR	1.441	-0.090	5		8
17	Z_14	Logic	XOR	1.479	0.154		6	8
18	A3C4E5_3	3R	Shape, orientation, and number	0.893	0.243	6		2
19	X_14	Logic	OR	1.082	0.745		7	5
20	Z_11	Logic	XOR	0.685	0.782	7		5
21	Z_8	Logic	XOR	1.278	0.888		8	4
22	Y_13	Logic	AND	0.488	1.245	8		8
23	A3D2E4	3R	Shape and number	0.736	2.636	9		5
24	B4D5E3_3	3R	Shading, size, and number	0.454	2.712		9	7
25	A3C2D5	3R	Shape and size	0.436	3.020		10	6
26	X_18	Logic	OR	0.332	4.269	10		6
<i>M (SD)</i>						.56 (.19)	.66 (.19)	

Note. In full set of 26 items, $M = .65$, $SD = .16$.

and sex differences. Present results suggest that the Sandia Matrices may show a two-factor structure, although it is unclear whether that two-factor structure is an artifact of item types or influenced by differences in the underlying cognitive processes required to solve the items. Notably, these results are consistent with prior research that suggests test

artifacts are more likely to account for the two-dimensional structure of the RPMs than are differences in required cognitive strategies (Vigneau & Bors, 2008). Regardless, the primary concern for the influence of a two-dimensional structure is rooted in potential sex differences on visuospatial items. Our results do not show evidence of sex

FIGURE 3.
Test Information Function for Final Set of 26 Items



differences on the Sandia Matrices, regardless of the factor structures tested here. Thus, we believe that proceeding with a unidimensional model is sufficient for scoring the recommended item sets. Further, there was no evidence of score differences by sex.⁴

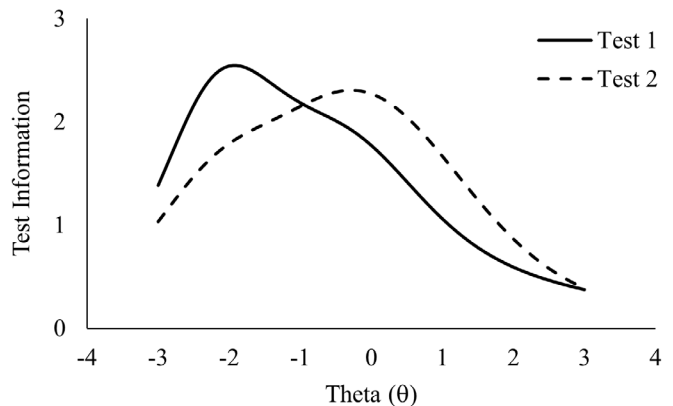
Nonetheless, results highlight other potential concerns. First, several items were removed prior to conducting key analyses due to evidence of extreme item parameters. In all cases, these items seemed to include a single competitive distractor. Even after eliminating these three problematic items, use of the 3PLM was warranted for additional item review due to evidence of substantial c (i.e., “guessing”) parameters. Notably, items used in this study were pregenerated and had already been reviewed to ensure appropriate distractors (Matzen et al., 2010). Thus, we caution against new software-generated problems without manually checking or manipulating distractors. Even using pregenerated and edited items without first conducting a thorough IRT-based psychometric review may yield misleading results.

Here, we have recommended two compilations of items with relatively reasonable parameters. The provided R code allows researchers to administer the recommended items sets or CAT to as few as a single participant and still derive theta estimates using the IRT parameters provided here. We expect that these measures and corresponding review of item properties will substantially aid researchers in implementing the Sandia Matrices in their own studies.

Limitations and Future Directions

This study utilized participants recruited from multiple sources, including Amazon Mechanical Turk and an undergraduate participant pool. Although the diversity of sources bolsters confidence in our findings, these populations may have intelligence distributions that differ from the average

FIGURE 4.
Test Information Function for Recommended 10-item Sets



adult in the United States. We encourage future research to explore characteristics of the Sandia Matrices in the broader population.

Additionally, the factor analytic approach used here is not necessarily appropriate for fully testing the types of cognitive strategies underlying the Sandia Matrices items. Given that the primary aim of investigating dimensionality in this study was to better understand potential sex differences, the limited evidence of sex differences, and the consistency of our approach with other studies investigating the dimensionality of the RPMs (see Waschl et al., 2016 for a review), we believe the analytic approach used here was sufficient for our purposes. Nonetheless, researchers interested in exploring cognitive strategies specifically should consider more advanced analysis approaches that were beyond the scope of this study (see Embretson et al., 1986; Mislevy & Verhelst, 1990).

Finally, additional studies might also consider how item types, including types of transformations and combinations, influence discrimination and location parameters to better inform construction of other Sandia Matrices item sets or use of the software in generating additional items. To fully supplant the RPMs with the Sandia Matrices, researchers will need to understand how to compile sets of Sandia Matrices items equivalent to both the Raven’s Standard Progressive Matrices and the Advanced Raven’s Progressive Matrices.

Conclusion

Although the RPMs are an extremely popular measure of intelligence, their proprietary status represents a limitation for many researchers. This study offers an initial IRT-based psychometric evaluation of Matzen et al. (2010)’s free alternative that shows no evidence of sex differences. We hope that the multiple, curated item sets recommended here will spur additional exploration of the Sandia Matrices as well as greater implementation of intelligence measurement in psychological research.

4 At the suggestion of a reviewer, we also tested for possible age effects in each sample. Neither sample showed a significant correlation between age and latent trait estimates (Sample 1: $r = 0.00$, $p = .880$; Sample 2: $r = -0.08$, $p = .115$).

REFERENCES

- Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences*, 36, 1459-1470. doi: 10.1016/S0191-8869(03)00241-1
- Arthur JR, W., & Woehr, D. J. (1993). A confirmatory factor analytic study examining the dimensionality of the Raven's Advanced Progressive Matrices. *Educational and Psychological Measurement*, 53, 471-478. doi: 10.1177/0013164493053002016
- Benz, Z., & Dixon, K. (2010). Sandia Generated Matrix Tool (SGMT) v 1.0 (Version 00) [Computer Software]. <https://www.osti.gov/servlets/purl/1231293>
- 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, 404-474. doi: 10.1037/0033-295X.97.3.404
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48, 1-29. doi: 10.18637/jss.v048.i06
- Colom, R., Escorial, S., & Rebollo, I. (2004). Sex differences on the progressive matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences*, 37, 1289-1293. doi: 10.1016/j.paid.2003.12.014
- 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. doi: 10.1016/0160-2896(95)90023-3
- Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement*, 41, 1295-1302. doi: 10.1177/001316448104100438
- Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, 95, 134-135. doi: 10.1037/0033-2909.95.1.134
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143-166. doi: 10.1177/014662169501900203
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum.
- Embretson, S., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement*, 23, 13-32. doi: 10.1111/j.1745-3984.1986.tb00231.x
- Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 37-60). Routledge.
- International Cognitive Ability Resource Team. (2014). The international cognitive ability resource. <https://icar-project.com/>
- Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, 32, 481-498. doi: 10.1016/j.intell.2004.06.008
- Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, 42, 525-541. doi:10.3758/BRM.42.2.525
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7, 361-388. doi: 10.1177/1094428104268027
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195-215. doi: 10.1007/BF02295283
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford Psychologists Press.
- Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment*, 81, 93-103. doi: 10.1207/S15327752JPA8102_01
- Stark, S. (2001). MODFIT: A computer program for model-data fit. Unpublished manuscript, University of Illinois at Urbana-Champaign.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1292-1306. doi: 10.1037/0021-9010.91.6.1292
- van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. Statistics for Social Behavioral Sciences.
- Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, 36, 702-710. doi: 10.1016/j.intell.2008.04.004
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, 117, 250-270. doi: 10.1037/0033-2909.117.2.250
- Waller, N. G. (1989). The effect of inapplicable item responses on the structure of behavioral checklist data: A cautionary note. *Multivariate Behavioral Research*, 24, 125-134. doi: 10.1207/s15327906mbr2401_8
- 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. doi: 10.1016/j.paid.2015.12.008
- Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 37-61). Routledge.

RECEIVED 08/26/19 ACCEPTED 07/10/20