Intelligence 59 (2016) 181–198

Contents lists available at ScienceDirect

Intelligence

Error patterns on the Raven's Standard Progressive Matrices Test

Maithilee Kunda [a,*,1], Isabelle Soulières [b], Agata Rozga [a], Ashok K. Goel [a]

[a] School of Interactive Computing, Georgia Institute of Technology, 85 Fifth Street NW, Atlanta, GA 30318, USA
[b] Department of Psychology, University of Quebec in Montreal C.P. 8888 succursale Centre-ville, Montréal, (Québec) H3C 3P8, Canada

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form 3 May 2016
Accepted 28 September 2016
Available online 5 November 2016

ABSTRACT

Although many psychometric tests, like the Raven's Progressive Matrices test, are commonly evaluated according to total score, additional variables can lend insight into the underlying cognitive processes of the test takers. We examine conceptual errors on the Raven's Standard Progressive Matrices (SPM) test. We present a new, complete classification of error types on the SPM using a two-kind coding scheme. We also present a new method for analyzing group errors patterns on these kinds of tests. We present two examples of this analysis using our SPM error classification. The first looks at the errors made by an artificial intelligence model of Raven's problem solving. The second example looks at the errors made by children and adults who are typically developing or have been diagnosed with autism. We close by discussing implications of this error classification and analysis method for the interpretation of SPM scores, towards a better understanding of the diversity of cognitive processes involved in Raven's problem solving.

1. Introduction

The Raven's Progressive Matrices (RPM) are a widely used series of intelligence tests[2] that contain multiple-choice visual analogy problems, such as the problem shown in Fig. 1. Each problem contains a matrix of geometric figures with one figure missing. The correct missing figure must be selected from among the given answer choices. Each test is divided into multiple sets of problems, and the "progressive" nature of the test comes from gradual increases in problem difficulty both within and across sets.

The RPM tests were originally designed by John C. Raven in the 1930s to measure eductive ability, or the ability to extract and understand information from a complex situation (Raven, 1936). Over time, the RPM was found to exhibit strong correlations with many other intelligence tests, leading it to be considered one of the best single-format psychometric measures of Spearman's general intelligence factor *g* (Snow, Kyllonen, & Marshalek, 1984). As a result, the RPM tests are widely used in clinical, educational, and scientific settings as a measure of general intelligence, or sometimes specifically of nonverbal intelligence.

Performance on an RPM test is typically measured in terms of overall score, i.e. number correct, which can then be used as an index into normative test data to determine a percentile ranking for the test-taker (Raven, Raven, & Court, 2003). While total score is certainly an important measure, this measure flattens the notion of individual differences to differences of *degree* and not of *kind*. This assumption about the nature of individual differences in intellectual ability can be problematic when interpreting test scores. For example, suppose two individuals are solving the test in two completely different, but equally effective, ways. In this situation, identical test scores do not imply cognitive sameness. Depending on the goals of the specific testing situation, these strategy differences, i.e. cognitive differences of *kind*, could be very important but remain hidden by the standard RPM scoring mechanism. For instance, RPM scores are often used for group matching in experimental studies. Even if two groups show similar distributions of RPM scores, it may not always be correct to assume that the two groups are cognitively homogenous.

Interestingly, the importance of individual differences in problem-solving strategies, as opposed to just problem-solving performance, was recognized early on by Alfred Binet, considered by many to be a key figure in the origins of intelligence testing (Fancher, 1985). He observed strategy differences between his two young daughters and made similar observations in his studies of intellectual savants. For example, he found that of two calculating prodigies, one seemed to perform his calculations using auditory mental representations while the other seemed to use visual mental imagery. However, as intelligence testing became more widely adopted, Binet's nuanced position was

* Corresponding author.
E-mail addresses: mkunda@vanderbilt.edu (M. Kunda), soulieres.isabelle@uqam.ca (I. Soulières), agata@gatech.edu (A. Rozga), ashok.goel@cc.gatech.edu (A.K. Goel).
[1] Present Address: Department of Electrical Engineering and Computer Science, Vanderbilt University, PMB 351679, 2301 Vanderbilt Place, Nashville, TN 37235-1679, USA.
[2] Terminology: Raven's Progressive Matrices (RPM) refers to the general family of tests. Specific test versions include: Standard Progressive Matrices (SPM), intended for children and adults in average ability ranges; Colored Progressive Matrices (CPM), intended for use with children, the elderly, or other individuals falling into lower ability ranges; and Advanced Progressive Matrices (APM), intended for higher-ability individuals.

http://dx.doi.org/10.1016/j.intell.2016.09.004
0160-2896/

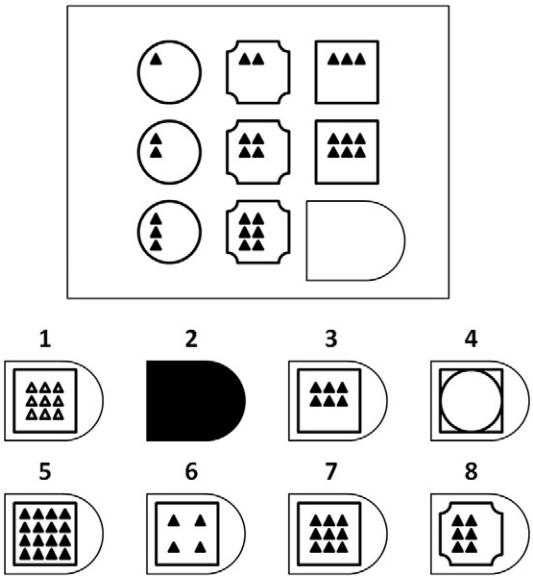


Fig. 1. Example of an RPM-like problem. The correct answer is #7.

lost in favor a view of intelligence as a unitary, unidimensional construct (e.g., Goddard, 1920).

The unidimensional conceptualization of intelligence and psychometric testing embodies one of two implicit assumptions about the relationship between cognitive abilities and test performance. The weaker assumption is that it does not matter how an individual solves particular test items; it just matters whether they can solve them. The stronger assumption is that test items themselves elicit very particular strategies, and these strategies taps into some facet of "general intelligence." In other words, all individuals solve psychometric tests in the same way, and the resulting scores measure quantitative variations in a unidimensional cognitive ability.

In reference to the stronger assumption, it is increasingly apparent that not all individuals exhibit the same qualitative forms of cognition, and that the same psychometric test can often elicit very different strategies from different people (Kunda & Goel, 2011). In reference to the weaker assumption, one way to conceptualize the existing divide is between "correlational and experimental approaches to human cognitive activity" (Keating, 1984, p. 17), where the former refers to traditional psychometric approaches, and the latter refers to approaches that are grounded in information-processing accounts of cognition. Understanding the strategies that underlie individual performance on psychometric tests like the RPM will not only lend insight into the cognitive abilities particular tests are measuring but will also shed light on the nature of individual differences of kind and not just of degree.

Some strides have been made towards trying to incorporate strategy differences into psychometric theory. Keating and Bobbitt (1978) observe that many sources of variance can contribute to differences in measurement outcomes, including: 1) differences in cognitive processing efficiency, 2) differences in strategy use, or 3) differences in metacognitive abilities of strategy selection. Hunt (1980) addresses in some detail the impacts of strategy differences on behavioral measures, though he does observe that the question of how the existence of strategy differences can be reconciled with the fairly robust statistical evidence of a *g* factor for intelligence is an important issue. Mislevy and Verhelst (1990) explicitly incorporate strategy differences into item response theory, though this approach does not seem to have been widely applied.

On the RPM in particular, numerous studies have leveraged findings from cognitive psychology to propose detailed models of the reasoning processes that people use to solve test items. Some of these models exist only as theoretical constructs, for example as process descriptions or algorithms, while other models have been implemented as working computational systems, following the research paradigm of artificial intelligence. Among the theoretical models, several propose that individual differences in RPM performance are a function of the ability to apply different types or numbers of conceptual transformations, like rotations or distributions of elements, together with perceptual factors related to the encoding or salience of stimuli (e.g., Embretson, 2004; Primi, 2001). Other theoretical models incorporate qualitative differences in strategy, such as whether people approach items using more gestalt vs. rule-based expectations (Hunt, 1974; Kirby & Lawson, 1983), or whether people choose answers based on prediction vs. elimination (Mulholland, Pellegrino, & Glaser, 1980).

In terms of working computational models, one influential model proposed that ability differences are largely a function of a rule storage and goal maintenance in working memory (Carpenter, Just, & Shell, 1990), while others have explored RPM performance in terms of strategies for analogy construction (Lovett, Forbus, & Usher, 2010) or rule induction (Little, Lewandowsky, & Griffiths, 2012; Rasmussen & Eliasmith, 2011). One study proposed an interesting extension of rules from prior modeling papers (Carpenter et al., 1990; Little et al., 2012) in which rules can apply over sensory dimensions as well as visual ones, and matrix reasoning problems are solved by a humanoid robot (Schenck et al., 2014). The computational model we discuss in Section 3.2 was designed to investigate differences in representational strategies for the RPM, i.e. whether problem information is represented visually versus verbally by the test-taker (Kunda, McGreggor, & Goel, 2013) .

While all of these models of RPM problem solving propose interesting cognitive possibilities, the ensuing challenge for researchers becomes how to empirically test model predictions by examining actual human RPM performance. Neuroimaging studies offer one window into how RPM items elicit particular patterns of brain activity (e.g., Soulières et al., 2009), but the gap between the specificity of cognitive processing models and the generality of neuroimaging results is still quite large. In the realm of behavioral observation, many researchers have moved beyond looking only at total score in order to obtain more detailed information about test-taker performance. Most of these behaviorally-focused efforts can be grouped into four categories: 1) item difficulty analysis, 2) item type analysis, 3) reaction time analysis, and 4) error analysis.

Item difficulty analysis examines whether test-takers exhibit systematic differences in terms of which items are easily solved and which items are more difficult. These studies generally follow the methods of item response theory, including Rasch analysis. Standardizations of the RPM, like psychometric standardizations in general, rely heavily on item analysis to produce test items that are intended to function in the same way in terms of difficulty for many different individuals. Several studies have looked for differences in RPM item functioning for different populations, including patients with neuroses (Halstead, 1943), patients with senile dementia (Eysenck, 1945), children with Down's syndrome (Facon & Nuchadee, 2010), children with Williams syndrome (Van Herwegen, Farran, & Annaz, 2011), and children on the autism spectrum (Dawson et al., 2007). For the most part, these studies have not found significant group differences in RPM item functioning. These findings, however, do not imply that item functioning is indeed the same across all test-takers, only that item functioning does not seem to vary in a systematic way across the different experimental groups that have been studied. For example, Halstead (1943) found that scores in both patients with neuroses as well as in an individually score-matched control group showed considerable variation in relative item functioning across sets in the SPM. In line with this observation about individual variation, the standard scoring procedures given in the SPM manual (Raven et al., 2003) suggest that each individual test-taker's scores for each set should be compared to the normative scores per set for the same total score. If these set-specific scores vary too much from the norm, then the manual labels this test-taker's score as "inconsistent" and cautions against conventional interpretations of

his/her test performance. In other words, if an individual solves all of the difficult RPM items correctly and misses all of the easy items, then his/her score should not be interpreted in the same way as an individual who has solved all of the easy items correctly but missed all of the difficult items, even if both individuals have the same total score.

Item type analysis investigates whether RPM problems can be divided into subgroups based on the kinds of cognitive strategies they elicit or require. These problem groupings are often obtained from factor analyses, e.g. (Lynn, Allik, & Irwing, 2004; van der Ven & Ellis, 2000). One method for problem grouping for the APM uses the outputs of a computational cognitive model, based on which of a pre-defined set of rules the model chooses to solve a given problem (Carpenter et al., 1990). These rules include concepts like, "constant in a row," and "distribution of three values." Under this approach, individual RPM problems can be grouped by the type of rule used to solve them or by how many rules need to be applied.

Reaction time analysis has been used in two different ways. First, by altering the presentation of RPM-like problems so that the matrix and answers are shown to the test-taker in different stages, reaction times can indicate the use of certain cognitive strategies (Bethell-Fox, Lohman, & Snow, 1984). In addition, several studies have looked at reaction time as an additional dimension over which to make group comparisons (e.g. Halstead, 1943; Soulières et al., 2009).

Error analyses investigate what happens when RPM problems are answered incorrectly, and in particular look at which of the given distracters the test-taker has selected as his/her chosen answer. The basic hypothesis behind error analyses is that test-takers make mistakes systematically. In other words, error analyses assume that when an error is made, it is not because the test-taker cannot solve the problem at all and is just guessing randomly. Instead, the test-taker is misled by a specific distracter for a definite reason, or the test-taker is making at least an educated (and not random) guess. Error analyses generally take one of three different forms, which we describe below.

First, several studies have investigated how the spatial layout of distracters affects the frequency of their selection. Raven himself found that spatial layout had considerable effect on test-taker behavior, with distracters placed close to the matrix blank much more likely to be selected than those placed further away (Miller & Raven, 1939). Second, many studies have looked at the selection frequencies of individual distracters across groups, e.g. (Eysenck, 1945; Forbes, 1964; van der Ven & Ellis, 2000). However, results from these studies are hard to generalize, because the data are relevant only to a specific problem and the specific distracters provided with that problem. A third and more general way of looking at errors has been to classify distracters as being of a particular conceptual type, which was begun in Raven's early studies of his test (Miller & Raven, 1939).

It is this third method—looking at conceptual types of errors made on the RPM—that is the focus of our study. The intuition behind looking at conceptual types of errors is that, when a test-taker chooses a particular distracter, it is because a certain class of cognitive process has led him/her to believe that that distracter is the correct one. In other words, test-takers using a particular kind of strategy will tend to choose particular kinds of distracters. In one interesting study that provides support for this idea, Kirby and Lawson (1983) developed a series of ambiguous RPM-like items in which different answer choices were deemed correct depending on the strategy one was employing. This study serves as evidence that different strategies can systematically lead to different answer choices, though in this case both answers are deemed correct.

While many studies have looked at types of errors on the RPM (see Section 1.2), the impacts of these studies have been limited by a lack of standard error type taxonomies and analysis methods. We believe this kind of error analysis has enormous potential to add useful information to RPM test interpretations (Sigel, 1963). We focus specifically on the SPM, for which no established classification of conceptual error types exists.

For the rest of the introduction, we first summarize the types of conceptual errors that have been identified for RPM problems, with examples of each conceptual type of distracter. Then, we present a review of previous studies from the RPM literature that have used different types of error analyses. Finally, we identify key shortcomings in the extant literature, and we describe the purpose of, and motivation behind, the present study.

1.1. Conceptual error types in RPM problems

The idea behind conceptual error types is that some particular, incorrect solution strategy may yield the choice of certain distracters in a systematic way. Thus, distracters should be classifiable according to which incorrect solution strategy they embody.

We have synthesized listings of conceptual error types from the CPM and APM manuals (Raven et al., 2003: manual sections 2, p. 5, and 4, p. 10, respectively) to produce an integrated list of four fundamental types of RPM errors: 1) *Repetition*, 2) *Difference*, 3) *Wrong Principle*, and 4) *Incomplete Correlate*. These four error types are described below, and examples of distracters representing each of these error types can be found in Fig. 1. We also give the corresponding labels for each error type as they originally appear in the CPM and APM manuals.

Repetition (R) errors occur when the chosen distracter copies a matrix entry adjacent to the blank space. Choosing an R distracter may represent some degree of perseveration or fixation on the problem matrix, such that an answer is selected using perceptual matching between the matrix entries closest to the blank space and the available answers. These adjacent entries may also be the last viewed before the test-taker moves on to look at the answer choices. Answer choices #3 and #8 in Fig. 1 are examples of R distracters. These errors are labeled as "Repetition of the pattern" in the CPM manual and as "Repetitions" in the APM manual.

Difference (D) errors occur when the chosen distracter is qualitatively different in appearance from the other choices. D distracters include those that are completely blank, as well as those that have extraneous shapes that are not found anywhere else in the problem. A D distracter can also be the most complex-seeming answer choice, either combining all of the matrix entries together into a single agglomeration of elements or taking some feature from the matrix and increasing its value until it surpasses all the other entries and answer choices. All of these variations share the quality that a D distracter is likely to visually "pop" from among the other answer choices. Answer choices #2 and #5 in Fig. 1 are examples of D distracters. These errors are labeled as "Difference" in the CPM manual and as "Over-determined choices / Confluence of ideas" in the APM manual.

Wrong principle (WP) errors occur when the chosen distracter is a copy or composition of elements from the problem matrix. A WP distracter might be chosen if the test-taker fails to identify the relationship from the matrix and instead combines the entries according to some other rule or relationship. Answer choices #4 and #6 in Fig. 1 are examples of WP distracters. These errors are labeled as "Inadequate individuation" in the CPM manual and as "Arbitrary lines of reasoning / wrong principle" in the APM manual.

Incomplete correlate (IC) errors occur when the chosen distracter is almost, but not quite, correct. For example, some IC distracters represent a rotation or reflection of the correct answer. Other IC distracters differ from the correct answer in a single feature dimension, e.g. they might have four elements instead of three, or straight elements instead of curvy ones, or have the correct shape but the wrong texture. Alternately, an IC distracter might be only missing one element when compared to the correct answer. Oftentimes, an IC distracter might be correct in terms of a single row or column in the matrix, e.g. looking just at the right-most column or just at the bottom-most row, but when both rows and columns are taken into account, it no longer fits the matrix pattern. These kinds of errors are made when a test-taker more or less "gets" the problem, in terms of identifying and

understanding the relevant matrix relationships, but fails to fully account for all of the problem details when selecting an answer. Answer choice #1 in Fig. 1 is an example of an IC distracter. These errors are labeled as "Incomplete correlate" in the CPM manual and as "Incomplete solutions / incomplete correlate" in the APM manual.

1.2. Previous studies of error types in the RPM literature

In this section, we provide a review of past literature that has looked at error patterns on the RPM as a way to gain additional insight into the different cognitive strategies that test-takers use, beyond looking at total score or item difficulty measures. The following studies all fall into three broad classes of error analyses:

1. **Spatial position** analyses refer to analyses of errors that look at whether test-takers are influenced by the spatial position of distracters when they make errors.
2. **Selection frequency** analyses refer to analyses of errors that look at the statistics of which specific distracters are chosen most or least often, for given RPM problems.
3. **Conceptual type** analyses refer to analyses of errors that group distracters into different conceptual types and then look at whether test-takers show tendencies towards making certain types of errors over others. Examples of conceptual types of RPM errors include repetition (R), difference (D), wrong principle (WP), and incomplete correlate (IC), as described in Section 1.1.

For each study included in this review, Table 1 provides demographic information, the RPM version that was used, which type(s) of error analyses were conducted, and a brief summary of results. Studies are presented in chronological order. More detailed information about the methods and results from each study is given in Appendix A, also in chronological order.

This review reveals a rich history of research into conceptual error types on the RPM. One recurring theme is that the spatial position of distracters exerts a surprisingly large influence on test-takers. However, building on many early studies, modern versions of the RPM appear to have minimized the influence of spatial position by balancing the correct answer and different distracter types across different positions (see Fig. 4 and Fig. 5 for our analysis of position and error type on the SPM). Another theme is the prominence of distracters that represent errors of repetition (R). Repetition errors appear to have been studied earlier and more often than the other error types and occur in many different populations of different ages, ability levels, diagnostic status, etc.

One factor that has been explored less extensively in RPM error analyses is the role played by the entire field of distracters in the likelihood of participants making certain errors versus others. Green and Kluever (1992) observed that inter-distracter relationships, such as distinctness or whether certain distracters are mirror images of others, are often correlated with item difficulty. To give a simple example, if all the distracters are identical except for one, then for sure the unique distracter must be the correct answer. Along these lines, White and Zammarelli (1981) examined the extent to which RPM items can be solved solely by examining patterns of variation within the set of distracters itself and found that many RPM items could be solved this way.

We close with one caveat related to this last observation. Many previous analyses of RPM error types did not correct for the baseline distributions of different error types in the test itself. So, for example, if a set of RPM problems has more distracters that represent repetition errors than other error types, this baseline distribution alone could be responsible for observations of "high" rates of repetition errors. This consideration is less important when comparing whether two groups show error patterns that differ from one another, but accounting for baseline distributions is nevertheless important in conducting meaningful evaluations of error response data. Babcock (2002) presented an analysis method that accounts for baseline error types, and we give a similar method in Section 3.1.

1.3. Purpose and motivation of our study

As evidenced by the summary of the literature given in Section 1.2, many researchers seek to conduct analyses of conceptual error types on the RPM tests. These analyses have revealed important differences in the types of conceptual errors made by individuals of different ability levels (e.g. Babcock, 2002) or developmental conditions (e.g. Gunn & Jarrold, 2004), among other findings. These analyses suggest that different groups of individuals may rely on different cognitive strategies to solve RPM items. Using this kind of error analysis to gain additional information about the cognitive characteristics of RPM test-takers greatly increases the value of the RPM as a cognitive assessment, especially since conducting an error analysis requires no change to the test itself or to the standard format of test administration.

Two key problems in conducting an error analysis, however, are that (1) there is no established classification of conceptual error types for the SPM, as there are for the CPM and APM, and (2) there are no standardized methods for how to compare the resulting data on test-takers' errors. The main contributions of this paper are our proposals for solving these two problems.

In Section 2, we present a new method for classifying distracters on the SPM according to conceptual type. We also give the results of applying this classification process, including measures of inter-rater reliability as well as a high-level description of the resulting SPM error classification.

In Section 3, we describe a new method for analyzing group differences in error type, and we present two concrete examples of applying this method on SPM data. First, we analyze the errors made by a computational cognitive model that solves RPM problems and discuss how this error analysis deepens our understanding of the model and its problem-solving performance. Second, we compare the errors made on the SPM by both children and adults who are typically developing or have been diagnosed with autism. While both examples use our SPM error taxonomy, this method could be applied to different error taxonomies and/or to any of the RPM family of tests.

In Section 4, we summarize the implications of this study and discuss how this kind of error analysis can improve the interpretation of group differences in RPM performance.

2. A new classification of error types for the SPM

As described in Section 1.1, both the APM and CPM manuals contain error taxonomies that classify distracters into one of four different types of conceptual errors (Raven et al., 2003). However, the SPM manual does not contain information on such error types. Vodegel Matzen et al. (1994) attempted to use the classifications from the APM manual to categorize distracters for sets C through E of the SPM, but inter-rater reliability between two coders was found to be only around 70%. The authors concluded that classification of SPM distracters seemed "problematic," as no explicit methodology for constructing distracters seems apparent in the test itself or in the existing SPM literature (Vodegel Matzen et al., 1994, p. 1).

To fill this gap, we developed a new taxonomy of conceptual error types for the SPM along with a classification system for assigning SPM distracters to these error types. As described in Section 1.1, we observed that the four types of errors given for the APM and CPM are conceptually similar despite having different names. We combined these ideas to identify four different types of conceptual errors that can be made on RPM problems: (1) repetition, (2) difference, (3) wrong principle, and (4) incomplete correlate. Descriptions of these error types are given in Section 1.1, and concrete examples of each error type can be found in

Table 1
Summary of published literature on RPM error patterns. Notes: Demographic characteristics are listed mostly using modern terminology, with occasional references to the original terminology from each study. Ages are given in years, either as a range ("min-max") or as mean and standard deviation ("M (SD)"). Test editions do not necessarily match modern versions. Error types in each study were not always labeled using our terminology; we have assigned labels based on our best interpretations of the original study intent: spatial position of distracters (Pos.), selection frequencies (Freq.), repetition (R), difference (D), wrong principle (WP), and incomplete correlate (IC). More detailed descriptions of procedures and results for each study are given in Appendix A.

			Error Types						
Reference	Participants	Items	Pos.	Freq.	R	D	WP	IC	Findings
Miller & Raven (1939)	younger TD children (age 5.7–7.5); older TD children	SPM	x		x			x	R errors common on difficult problems. R errors also affected by distracter position.
Raven (1939)	children (n = 308; age 3–9), (n = 38; age 13–14), (n = 53; age 8–9), (n = 341; age 9–14), (n = 43; age 9–10), (n = 178; age 13–14), (n = 56, age 5–9); adult students (n = 24); adult soldiers (n = 44); both low and typical ability levels.	SPM, CPM board form	x		x	x	x	x	R and IC errors common for all participants. Some participants made WP and D errors. Low-ability children made mostly R errors.
Halstead (1943)	adult neurotic patients; TD adults	SPM		x	x	x		x	Low ability participants tended to make R errors and "perceptual" errors, and also persisted in one type of error for several items in a row. High ability participants tended to make IC errors.
Eysenck (1945)	adults with senile dementia (n = 100; age 73 (6.5)); TD adults (n = 2790, same sample as Halstead 1943)	SPM board form	x	x	x				R errors very common, accounting for >75% of errors in both groups when available. R errors slightly higher in adults with senile dementia.
Sigel (1963)	two groups TD female children (n = 42; age 9) (n = 53; age 10); three groups TD male children (n = 34; age 9) (n = 49; age 10) (age 7–11)	SPM Set B, SPM	x		x	x	x	x	Patterns of error types not related to total set score. Influence of distracter position is higher for younger children, and is more pronounced for R distracters.
Bromley (1953)	adult psychiatric cases, classified as senile dementia, paranoid, depressive, or organic (n = 35; age 61 (11.3))	SPM	x	x	x	x	x	x	R errors most common, followed by WP and D, and then IC.
Forbes (1964)	adolescents and adults (n = 2256; age 15–30)	APM		x	x	x	x	x	IC errors most common (>50% of errors) in participants of average and above-average ability. WP errors somewhat more common in participants of low ability. R and D errors not very common, and less common with increasing ability.
Weatherick (1966)	TD adults (n = 236; age 20–69); compared with sample from Bromley 1953	SPM		x	x				When most frequent errors are different between groups, R errors more common in adults with senile dementia.
Vejleskov (1968)	TD children (age 12.0 (0.4))	SPM		x				x	On certain problems, IC errors common in female participants.
Carter (1970)	TD adults with high verbal ability (n = 12; age 33.3, range 20–48); TD adults with lower verbal ability (n = 12; age 35.4, range 27–48); paranoid schizophrenic patients (n = 12; age 35.5, range 22–48); nonparanoid schizophrenic patiens (n = 12; age 31.9, range 19–45); all groups half male, half female	SPM subset		x					Participants asked to rank all distracters for given RPM problems in terms of similarity to the correct answer and to the elements in the problem matrix. Results reported as part of aggregate analyses of several tasks, used to examine factor structure of test as well as discriminability of groups.
Jacobs & Vandeventer (1970)	TD children from the USA, first grade (n = 81) and third grade (n = 20); TD children and adults (Eskimo) from Baffin Island (n = 114; age 10–40+); TD children and adults (Temne) from Sierra Leone (n = 119; age 10–40+)	CPM						x	Proportion of IC errors (vs. other error types) correlated with score in lower-scoring groups, but not in higher-scoring groups.
Guttman (1974)	TD children (age 8–24), parents (age 35–61), and siblings and cousins (n = 408 total)	SPM		x					On each item, 2–3 distracters more frequently chosen than others. Parent/child and sib/sib pairs choose same distracter with same rate as across whole sample (Sets C-E only).
Thissen (1976)	TD children in junior high (n = 570)	subsets of SPM and CPM		x				x	Likelihood of choosing particular distracters varies by ability and by similarity of distracter to correct answer.
Horner & Nailling (1980)	Adult males with left brain damage (n = 12; mean age 54.1), right brain damage (n = 12; mean age 58.4), or no brain damage (n = 12; mean age 56.3)	CPM			x	x	x	x	R is most frequent error type across all groups. (Did not control for base error type distributions.)
Vodegel-Matzen et al. (1994)	TD children (n = 1655, age 8.5–12.5)	SPM, Sets C-E			x	x	x	x	IC errors most common for all participants, followed by WP. R and D errors were not common. R errors relatively more common for lower-scoring participants.
van der Ven & Ellis (2000)	TD children (n = 710, age 12–15)	SPM, Set C and Set E		x				x	Certain problems elicit certain error types, including IC as well as not taking whole matrix into account or lacking resistance to distracters.
Babcock (2002)	TD adults (n = 818, age 17–99, mostly 18–30 or 60+)	APM			x	x	x	x	Medium and high-ability participants make more IC errors than other types. Low-ability participants do not show an error type preference.
Gunn & Jarrold (2004)	TD children (n = 213; age 7.4 (1.5)); children with Down syndrome (n = 39; age 13.0 (2.4)); children with moderate learning disabilities (n = 171; age 11.8 (2.8))	CPM			x	x	x	x	Children with Down syndrome make fewer R errors and more WP and D errors than other groups. Subgroups group-matched on total score show same pattern. In TD group, age is positively correlated with R and IC errors, and negatively correlated with WP and D errors. Pattern in similar in children with Down syndrome, but not significant.

(continued on next page)

Table 1 (*continued*)

Reference	Participants	Items	Error Types						Findings
			Pos.	Freq.	R	D	WP	IC	
Fajgelj et al. (2010)	TD children (n = 2334; age 4–11)	CPM	x	x	x				Younger children more often favor particular distracters (i.e. distracter chosen by >20% of participants). Younger children make more R and position-based errors. Older children show some influence of distracter position but not significantly. R errors and distracter position often have overlapping effect.
Matzen et al. (2010)	TD adults (n = 80; age 20, range 17–40)	SPM subset and SPM-like problems			x		x	x	Errors types vary by number, direction, and types of transformations in problem matrix. R errors common.
Van Herwegen et al. (2011)	Groups individually matched on total score: TD children (n = 53; age 5 (1.3)); children with Williams syndrome (n = 53; age 18.3 (9.8))	CPM			x	x	x	x	No group differences in error types or change in error types with age. With increasing age, R errors increase, D and WP errors decrease, and IC errors do not change.

the distracters for the problem shown in Fig. 1. In the following subsections, we describe:

1. Specific criteria for classifying distracters into one of these four types (Section 2.1).
2. The two-stage coding method that we developed for applying these criteria to code SPM distracters into different types (Section 2.2).
3. Results of the coding process, including inter-coder agreement (Section 2.3).
4. Descriptions of the types of errors observed to be present on the SPM (Section 2.4).

2.1. Criteria for classifying RPM distracters

From the four conceptual types of errors that we identified (repetition, difference, wrong principle, and incomplete correlation), we developed a list of specific criteria, given in Table 2, that can be used to classify particular distracters from the SPM into one of these four types. These criteria are similar to those given in the CPM manual (Raven et al., 2003: manual section 2, p. 5) but include additional criteria to cover all of the distracters contained in the SPM, which is a much longer test than the CPM and displays more diversity in the distracters that are presented.

2.2. Two stage coding method

A major difficulty in performing an error classification for an existing test like the SPM is ambiguity. Certain distracters might fit multiple classification criteria, and it is not always clear how to best resolve these ambiguities (Vodegel Matzen et al., 1994). For example, in Fig. 1, choosing answer #3 can be viewed either as a *Repetition* error, as it duplicates the matrix entry immediate above the blank space, or as an *Incomplete Correlate* error, as it differs from the correct answer only through missing a single row of triangles.

Observation of these kinds of ambiguities on the SPM led us to a key observation, which is that the four error types listed in Table 2 in fact represent two orthogonal classifications of distracters. We define these as *Kind I* and *Kind II* errors:

1. **Kind I errors:** *Repetition*, *Difference*, and *Wrong Principle* errors all have to do with how a particular distracter is related to information in the matrix and in the other answer choices, completely apart from the content of the correct answer choice. In particular, these errors assume that the test participant is attending to irrelevant or erroneous aspects of the problem, and that he/she not able to discover even a partial solution to the problem.
2. **Kind II errors:** *Incomplete Correlate* errors, on the other hand, have to do with how a particular distracter is related to the correct answer choice. These errors assume that the participant correctly guesses some part of the solution but does not quite attain the correct answer.

From this observation, we developed a two-stage manual coding method for classifying distracters from the SPM into error types. In the first stage, all answer choices (distracters as well as the correct answer choice) for a given problem are classified according to whether they

Table 2
Criteria for classifying distracters on the SPM into four error types. These criteria are adapted from those used for the CPM (Raven et al., 2003: manual section 2, p. 5), with some new criteria added to cover additional distracter features represented in the SPM.

Error type	Code	Criteria
Repetition (R)	1. R-Left	Repetition of matrix entry to left of blank space
	2. R-Top	Repetition of matrix entry above blank space
	3. R-Diag	Repetition of matrix entry to top-left of blank space
Difference (D)	4. D-Blank	Filled completely white or black
	5. D-Union	Union of matrix entries or aspects of them, such that union has more components than any single matrix entry
	6. D-Plus	Maximizes some feature value or makes it more complex
	7. D-Diff	Differs qualitatively from matrix and other answers, or contains information not found anywhere in matrix
Wrong Principle (WP)	8. WP-Copy	Copy of matrix entry not adjacent to blank space
	9. WP-Flip	Rotation/reflection of matrix entry
	10. WP-Matrix	Other transformations or combinations of matrix entries or aspects of them, including negative images
Incomplete Correlate (IC)	11. IC-Neg	Negative (color-inversion) of correct answer
	12. IC-Fill	Change only in fill, texture, or style from correct answer
	13. IC-Flip	Rotation/reflection of correct answer
	14. IC-Layout	Change only in spatial layout of elements from correct answer
	15. IC-Scale	Change in size and/or scale (including feature-wise scaling) from correct answer
	16. IC-Num	Change only in number of discrete elements (allowing for slight changes in layout) from correct answer
	17. IC-Inc	Incomplete version of correct answer, with missing element or portion

represent *Kind I* errors, i.e. whether they fit any of criteria #1–10 in Table 2. In this stage, the entire problem, both matrix and answer choices, is relevant to the coding, and no information about which answer choice is correct is relevant. In the second stage, distracters are classified according to whether they represent *Kind II* errors, i.e. whether they fit any of criteria #11–17 in Table 2. In this stage, the classification depends only on the relationship of each distracter to the correct answer, and no information about the problem matrix is relevant.

After both stages of coding are complete, every single distracter will have received a *Kind I* code, and some distracters will have received *Kind II* codes. The process we used to resolve conflicts between codes in order to assign each distracter a final code is described in Section 2.3.

2.3. Error coding results

The SPM is divided into five sets labeled A through E, each containing 12 problems. Each problem in Sets A and B contains six answer choices, and each problem in Sets C, D, and E contains eight answer choices. Thus, there are a total of 432 answer choices contained in the entire SPM, 60 of which represent the correct answers and 372 of which are distracters.

We had two independent coders perform the SPM error coding. Each coder received a coding protocol, which has been reproduced in its entirety in Appendix B. This protocol contains descriptions of each error type, an example problem illustrating the various types and classification criteria, and finally, an instruction sheet with ordered codes and criteria to use for each stage of the coding procedure.

For the first stage, coders were given a full copy of the test booklet in which no answers had been marked. For the second stage, coders were given a different copy in which the correct answers had been marked and the matrix portions of each problem had been cut off, so only the answer choices were visible.

The initial agreement between the two coders was 82% for the first stage ($n = 432$ answer choices coded) and, by coincidence, 82% for the second stage ($n = 372$ answer choices coded; 60 answer choices representing correct answers did not require codes). Kappa coefficients were calculated to test for independence between raters. The kappa values were 0.79 for the first stage and 0.67 for the second stage.

Then, the two coders met to negotiate and discuss instances of differing code assignments. There were several systematic disagreements that were easily resolved by making the coding criteria more specific. For example, the D-Union criterion (Criterion #5 in Table 2) was modified to specify that this type of distracter had to have more elements in it than any entry in the matrix, which was not originally part of the criterion. All of these changes have been incorporated into the final criteria listed in Table 2.

After the negotiation and criteria-revision phase, the coders revised their code assignments and agreement was re-calculated. Post-negotiation agreement was 95% for the first stage and 98% for the second stage. Remaining disagreements were resolved based on consideration of the conceptual type of error intended to be captured.

After final, post-resolution codes were assigned to each individual distracter, each distracter was classified according to one of the four over-arching error types. Of the 372 distracters given in the SPM, 242 of them received only a *Kind I* code; for these distracters, the classification into types is straightforward using the information in Table 2. The remaining 130 distracters received both *Kind I* and *Kind II* codes.

To resolve final error classifications for distracters that received both codes, we used the following rules, which were determined after the initial coding was complete based on which specific code pairings were observed. These rules attempt to resolve observed code pairings by considering the complexity of operations involved across both *Kind I* and *Kind II* codes, resolving conflicts in favor of distracters that copy, rotate, or flip information, as opposed to operations that affect features like texture or number of elements, and also in favor of principled, as opposed to arbitrary, hypothesized lines of reasoning. The rules are applied in order as follows:

1. *Kind I* repetition errors take precedence over any *Kind II* error. *Kind I* repetition errors involve a simple perceptual copy of adjacent matrix entries, whereas all *Kind II* errors involve somewhat successful efforts towards a correct solution using various other, often more complex, reasoning operations. This rule resolves 40 of the 130 paired-code distracters, leaving 90 to be resolved.
2. Any *Kind II* error takes precedence over *Kind I WP-Matrix* errors. *Kind I WP-Matrix* errors involve some arbitrary combination of matrix elements that is not, for example, a simple copy or flip (*WP-Copy* or *WP-Flip*) or something that is visually salient (difference errors). Thus, if a distracter has received codes for being any kind of incomplete correlate as well as *WP-Matrix*, we assume that it is chosen as part of an incomplete but properly-directed solution attempt, rather than some apparently equally complex but arbitrary reasoning process. This rule resolves 46 of the 90 remaining distracters, leaving 44 to be resolved.
3. *Kind II IC-Flip* errors take precedence over any *Kind I* error. *Kind II IC-Flip* errors represent a simple rotation or flip of the correct answer, whereas all the conflicting codes for remaining *IC-Flip* distracters represent either wrong principle or difference errors, which assume either arbitrary lines of reasoning or more complex visual operations such as combining elements across all distracters. This rule resolves 12 of the 44 remaining distracters, leaving 32 to be resolved.
4. *Kind I WP-Copy* or *WP-Flip* errors take precedence over any *Kind II* error. If a distracter is a copy or rotation/flip of a matrix entry as well as an incomplete correlate that uses some more complex operation, we assume it is chosen as part of the application of the simpler copy/rotate type of operation and represents a wrong principle error. This rule resolves 26 of the 32 remaining distracters, leaving 6 to be resolved.
5. *Kind I D-Plus* or *D-Diff* errors take precedence over any *Kind II* error. If a distracter takes some feature and maximizes it (*D-Plus*) or changes the distracter content entirely (*D-Diff*), we assume these are more likely reasoning operations than complex variants of the correct solution (*Kind II* errors that alter the number, texture, or layout of information relative to the correct answer). The remaining 6 distracters are thus coded as difference errors.

2.4. SPM error classification results

In this section, we present observations and analyses related to how SPM distracters are distributed with respect to the four error types. We also performed an analysis of how error types are distributed across individual answer positions on the test.

Fig. 2 illustrates the proportion of answer choices falling into each error type category on each individual SPM problem. As this figure shows, the proportion of answer choices falling into each error category varies drastically from problem to problem. These proportions can be interpreted as the chances that a random guesser will produce responses corresponding to each category and are important to consider when looking at the patterns of errors generated by a test-taker. For example, as shown in Fig. 2, every single distracter on problem 2 corresponds to a *Difference* error; thus, the fact that 100% of test-taker errors on this problem fall into the *Difference* category is trivial.

Next, we calculated the expected distribution of errors that would be made by a random guesser, similar to the method used by Babcock (2002). In other words, suppose that a test-taker, for any problem that he/she cannot solve, randomly guesses among the available distracters. What distribution of errors will he/she produce?

We computed this distribution as follows. For each error type $\boldsymbol{t}$, let $\boldsymbol{P_t}$ be the proportion of that error type expected through random guessing,

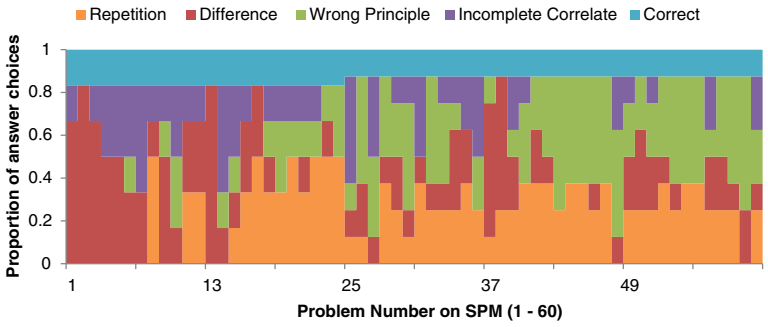


Fig. 2. Proportion of answer choices falling into four error categories plus correct answer category, for each SPM problem. These proportions can be interpreted as the probability that a test-taker making random guesses will produce a response corresponding to each of these categories. Note that the chance of producing a certain response varies from problem to problem. For many problems, one or more possible error categories can be entirely absent. The chances of randomly guessing the correct answer is 0.17 (1 out of 6 possible answer choices) for the first 24 problems and 0.125 (1 out of 8 possible answer choices) for the latter 36 problems.

and let p_{ti} be the proportion of that error type represented by distracters for each problem i. Then, for any subset of problems i_1 to n, P_t can be computed as:

$$P_t = \left(\frac{1}{n}\right)\sum_{i=i_1}^{n} p_{ti}.$$

Results from this calculation are shown in Fig. 3. The expected distribution of errors is not uniform across the four error types for any individual set of the SPM or across the test as a whole. This variability is important to consider when analyzing a test-taker's error patterns on the SPM, as we describe in the next section.

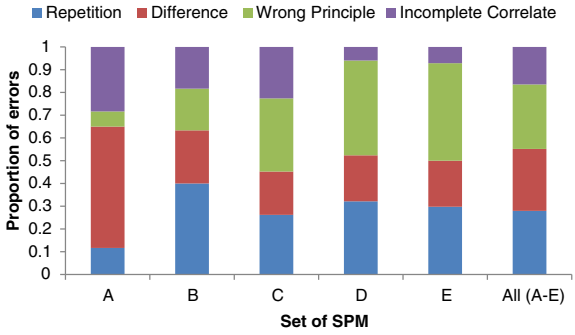


Fig. 3. Proportion of errors expected to be produced by a random guesser on each set of the SPM. These values are computed according to the per-problem distributions of error types represented by given distracters, within each set A through E and also across all five sets.

Finally, we performed an analysis of how the SPM error types that we observed are distributed across absolute answer positions on the test. As the SPM literature has shown, absolute answer position can strongly influence the distracters chosen by a test-taker. In particular, many people tend to choose answers in positions that are close to the blank matrix entry (Eysenck, 1945; Miller & Raven, 1939). Thus, if certain error types are found with greater frequency in these favored positions, then this position-choosing tendency could be a confounding factor in attempts to analyze the conceptual error patterns themselves.

Fig. 4 and Fig. 5 show the distribution of error types across answer choices 1 through 6 for Sets A and B and across answer choices 1 though 8 for Sets C, D, and E. The error frequencies were too small to perform a standard chi-square test of independence, and so data were analyzed using a simulated p-value, with the statistical software package R. While there is some variation in the distribution of error types across answer position, the distributions are not significantly different for 2×2 matrices, $\chi^2(N = 144) = 6.71$, $p = 0.99$, or for 3×3 matrices, $\chi^2(N = 288) = 16.46$, $p = 0.96$. Therefore, we assume that the interaction between position and error type is minimal.

3. Analyzing group differences in error patterns on the SPM

The previous studies looking at conceptual error patterns on the RPM tests have used a variety of analysis methods. Because these methods are so different, it is difficult to compare results from one study to the next. In particular, some studies control for factors that other studies omit. For example, some studies take into account the uneven distribution of error types in each RPM problem (e.g. Babcock, 2002; Vodegel Matzen et al., 1994), while other studies do not (e.g. Gunn & Jarrold, 2004).

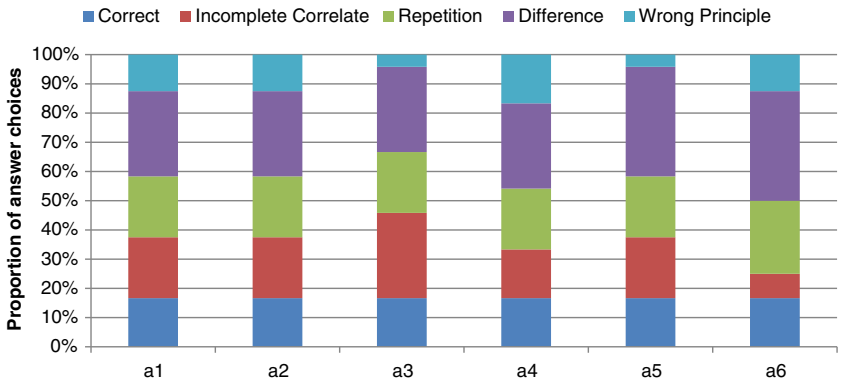


Fig. 4. Proportion of error types found across each answer position for 2×2 matrices.

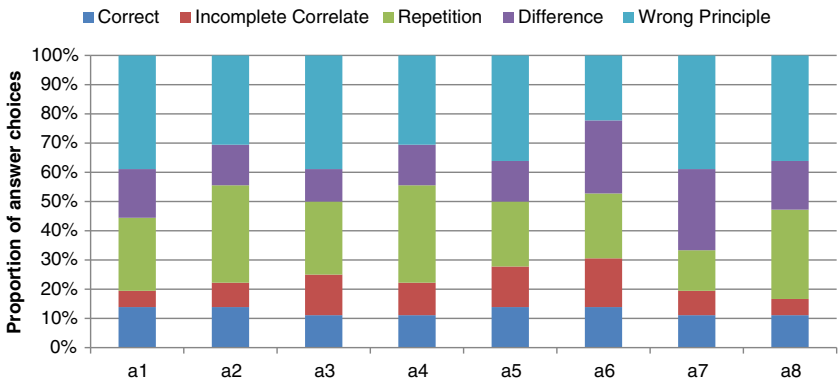


Fig. 5. Proportion of error types found across each answer position for 3 × 3 matrices.

In this section, we present a new proposal for how to analyze group differences in error patterns on the RPM tests. We describe this method in Section 3.1, and then we give two concrete examples of applying this method to real data, using the SPM error classification that we discussed in Section 2. While both examples use error data from the SPM, this method could also be used on any of the RPM tests, or with any alternate error classification. Section 3.2 presents an analysis of errors made by a computational artificial intelligence (AI) model that was designed to solve RPM problems, and we discuss how this error analysis provides additional insight into the model's performance, beyond looking just at total score alone. Finally, in Section 3.3, we present an analysis of errors made by human test-takers, divided into groups of children and adults who are either typically developing or who have been diagnosed with autism spectrum disorder (ASD).

3.1. Method for analyzing RPM error patterns

When analyzing errors made by an individual on the RPM, there are three assumptions that we use to narrow the scope of the analysis. The first assumption is that when a test-taker chooses the correct answer choice, they are choosing this answer for the right reason and have correctly solved the problem. This assumption is important because it is possible that correct answers are sometimes chosen by random guessing or even by incorrect lines of reasoning. For example, sometimes the correct answer on a problem is also a repetition of one of the adjacent entries in the matrix. So if a test-taker has the strategy of always choosing an answer choice that is a repeated entry, i.e. making a Repetition-type error, he/she may occasionally hit upon the correct answer, even though she/he was actually using an erroneous strategy. This kind of mistake is difficult to detect, and since the chances are likely to be low of choosing the correct answer by mistake, we disregard this kind of error and assume that when the correct answer has been chosen, it is chosen for the right reasons. Thus, we focus our error analysis only on the incorrect answers that have been chosen by a test-taker.

Second, we assume that the effects of spatial layout are negligible. In fact, as summarized in Section 1.2, there have been many studies showing that the spatial layout of distracters has a considerable effect on the errors that participants make. However, we have shown in Section 2.4 that the conceptual types of errors present on the SPM are distributed more or less equally across spatial positions, and so we omit further consideration of spatial position as a covariate in our analyses.

Third, we assume that if a test-taker is choosing a distracter at random, that choice should not count as making a conceptual error of any particular type. Again, this event cannot be directly measured, but it means that we must take into account the existing distribution of error types on each problem when computing error frequencies. For example, if a problem only has distracters of the Repetition type, then the fact that a test-taker makes a Repetition-type error on that problem means nothing whatsoever. In other words, we wish to control for the existing distribution of error types on each problem, which, as we have shown in Fig. 2 and Fig. 3, is not uniform.

Here, we present a method for analyzing the errors made by a group of individuals on the RPM that meets all of these goals. The basic idea behind our method is to first compute the expected errors that would be made by a group of random guessers across all problems on the test, and then compare the actual errors made by the participant group to this baseline. The result is a measure of how much the participant group deviates from this random-guessing baseline, on each different conceptual error type.

As before, let $\boldsymbol{p_{ti}}$ be the proportion of each error type $\boldsymbol{t}$ represented by the distracters for test problem $\boldsymbol{i}$. For any group that has taken the test, let $\boldsymbol{g_{ti}}$ be the number of errors of type $\boldsymbol{t}$ that the group has made on problem $\boldsymbol{i}$, and let $\boldsymbol{g_i}$ be the total number of errors made by the group on problem $\boldsymbol{i}$. Then, the product $\boldsymbol{g_i p_{ti}}$ gives the expected number of errors of type $\boldsymbol{t}$ that would have been made by the group if all incorrect answers were chosen by randomly guessing from among the incorrect distracters. The difference $\boldsymbol{d_{ti}} = \boldsymbol{g_{ti}} - \boldsymbol{g_i p_{ti}}$ gives a measure of how much the actual errors made by the group differ from a random-guessing baseline. A positive value for $\boldsymbol{d_{ti}}$ means that the group makes errors of type $\boldsymbol{t}$ more frequently than chance, and a negative value means that the group makes these errors less frequently than chance. The average value $\boldsymbol{D_t}$ of $\boldsymbol{d_{ti}}$ across all test problems $\boldsymbol{i}$ gives an overall measure of the group's tendency (in terms of deviation from chance) to make a particular type of error. We summarize this method as follows:

$$D_t = \frac{\sum_i \left(g_{ti} - \left(\sum_t g_{ti}\right) p_{ti}\right)}{\sum_{t,i} g_{ti}}$$

The resulting values of $\boldsymbol{D_t}$ can then be used to evaluate a group's error patterns on the test, either as a comparison to random guessing or in comparison to another group.

Though derived independently, our method is very similar to the one used by Babcock (2002). We consider the convergence of our respective research studies on such similar methods to be a positive point in favor of this general type of method, the core feature being that errors are measured in relation to the baseline error type distributions that are already present in each set of distracters.

3.2. Analysis of errors made by ASTI computational model

Using the method presented in Section 3.1, we analyze the errors made on the SPM by a computational artificial intelligence (AI) model of RPM problem solving. In previous work (Kunda, 2013; Kunda et al., 2013), we presented a computational model of problem solving on

the RPM called the Affine and Set Transformation Induction (ASTI) model. This model was constructed in order to investigate problem solving on the RPM using visual mental representations. All previous extant computational RPM models had relied on propositional forms of representation (e.g. Carpenter et al., 1990), despite a breadth of evidence from human studies suggesting that problem solving can proceed using either visual or verbal forms of representation (see Kunda et al., 2013 , for a summary of these studies).

The ASTI model also has implications for a recent study of RPM performance in individuals diagnosed with autism, which found that these individuals seemed to use predominantly visual strategies (Soulières et al., 2009), in line with other empirical evidence showing a visual cognitive bias in autism (Kunda & Goel, 2011).

The ASTI model uses purely visual representations in the form of pixel-based images along with affine and set transformations designed to emulate the types of operations observed in studies of human mental imagery. The model uses a constructive matching approach (Bethell-Fox et al., 1984): first, it examines different subsets of the matrix entries (each an individual image), under each of these transforms to induce a "best-fit" overall transform. Then, the ASTI model applies this best-fit transformation to the remaining matrix entries to generate a predicted answer image. Finally, this predicted answer is compared to each answer choice to select the best match.

The current version of the ASTI model correctly answers 50 out of 60 problems on the SPM (Kunda, 2013). One difficulty with high performing computational models such as ASTI is that it is not immediately clear how errors made by the model might be analyzed in a meaningful way, as error data can only be collected on 10 of the 60 problems.

We use a method for obtaining error data from a computational RPM model through model ablation (Cohen & Howe, 1988). In particular, the ASTI model uses affine transforms (rectilinear rotations and reflections), as well as addition, subtraction, and pair-wise image composition (union, intersection, etc.). The model also inspects the matrix according to rows, columns, and diagonals. By ablating the model through removing access to subsets of these mechanisms, we can observe the errors made by general classes of ASTI configurations. Table 3 lists mechanisms used for 2×2 matrices (found in Sets A and B of the SPM) and 3×3 matrices (found in Sets C through E of the SPM). Ablating combinations of these mechanisms yields 96 different model configurations, whose total scores range from 15 to 50 correctly solved problems.

We performed the error analysis described in Section 3.1 on our artificial model "group" of 96 test-takers. While not all of these 96 configurations necessarily represent "cognitively plausible" models of RPM problem solving, they do illustrate the space of solution strategies that is embodied by the current model implementation. For this reason, we do not attempt to directly compare model errors against those of humans. Future work with the model will involve identifying subsets of "cognitively plausible" configurations for use in human comparisons.

Results from this analysis are shown in Fig. 6. As shown in this figure, it is immediately apparent that the model makes more errors of type Repetition than would be expected by chance, and fewer errors of type Difference. By itself, this is an interesting finding, because it shows that our model has some definite strategic biases in how it goes about attempting to solve SPM problems. In particular, the ASTI model never guesses randomly, but always chooses the most similar answer choice to what it has imagined the correct answer to be. In contrast, a model that produces random guesses whenever it has low confidence in its selection of the correct answer would be expected to show error tendencies close to zero, using this analysis method.

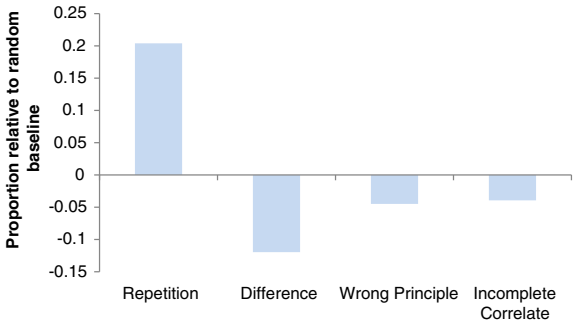


Fig. 6. Proportion of errors of different conceptual types made by ASTI model. This graph shows data aggregated across all SPM problems and in comparison to a random-guessing baseline.

Why does the ASTI model make so many Repetition errors? The model uses elements of the matrix as building blocks to construct its prediction of what the correct answer image will be. In particular, the model most often specifically uses the matrix entries that are adjacent to the blank space as the primary building blocks. Thus, it makes sense that the model shows such a strong tendency towards making Repetition errors. On the other hand, why does the model make relatively few Difference errors? Recall that Difference errors are defined by distracters that have very little in common with any of the matrix entries. Thus, since the model composes its answer prediction directly from the existing matrix, it makes sense that the model will seldom generate an answer prediction that matches one of the Difference distracters.

These new observations about the model are valuable for informing the continued design of the model as a tool for understanding human cognition. In particular, by observing the range of errors that humans can make, as discussed in the literature review in Section 1.2, we can identify mechanisms that the model should include in order to be able to explain these human findings in computational terms. For example, repetition errors in humans are thought to stem from visual priming or perseveration, and so future versions of the model should include some parameter that can control the extent to which the model will fixate on the matrix entries that it has seen; currently, the model's level of visual fixation is essentially static. Likewise, Difference errors in humans are likely influenced by the visual salience of particular distracters. The ASTI model currently does not include any mechanisms of attention based on visual salience, which represents a second valuable direction of investigation.

Both of these observations about the ASTI model, i.e. its tendency towards perseveration-like behavior and its lack of mechanisms for salience-based visual attention, were made possible by observing the results of our analysis of the errors made by the model. Neither observation is apparent just from looking at the total SPM scores achieved by the model. Furthermore, as we continue to refine and evaluate the ASTI model relative to human performance, this kind or error analysis provides a valuable dimension of comparison between the two.

3.3. Analysis of errors made by children and adults across developmental conditions

As a second example, we present an analysis of the errors made by children and adults who are either typically developing (TD) or have been diagnosed with autism spectrum disorder (ASD). Whereas the

Table 3
Component mechanisms in the ASTI model used for ablation experiments. Different configurations of the model were created by removing one or more of these mechanisms. A total of 96 ablated model configurations were developed and tested on the SPM.

Type	Image sets	Transforms
2×2 matrices	1. Rows	1. Identity
	2. Columns	2. Rotation/reflection
		3. Addition/subtraction
3×3 matrices	1. Rows	1. Identity
	2. Columns	2. Rotation/reflection
	3. Diagonals	3. Addition/subtraction
		4. Composition

Table 4
Participant demographics. Participants are divided into groups according to age (children and adults) and diagnostic status (autism (ASD) or typically developing (TD)). Group differences between the TD and ASD groups were evaluated using a two-sample t-test. Superscripts indicate significant group differences in the mean.

	Children		Adults	
	TD	ASD	TD	ASD
N	54	105	52	42
Age in years: mean (SD)	11.96 (3.40)	11.02 (2.99)	22.98[1] (4.28)	26.80[1] (6.72)
Full scale IQ: mean (SD)	109.82[2] (10.35)	84.38[2] (20.03)	106.91[1] (11.76)	97.61[1] (16.40)
SPM score: mean (SD)	42.61[1] (9.79)	37.57[1] (12.13)	50.69 (5.38)	48.50 (9.71)

Note: Not all participants had FSIQ data available.
[1] $p < 0.01$.
[2] $p < 0.001$.

RPM scores of TD individuals are usually strongly correlated with their Wechsler IQ scores, individuals with autism have demonstrated RPM scores much higher than their Wechsler scores (Bölte et al. 2009; Dawson et al. 2007; Mottron, 2004). The reason for this difference is not well understood, although fundamental differences in cognitive strategy (Kunda & Goel, 2011) and the underlying neural activation (Soulières et al., 2009) have been proposed.

Looking at differences in RPM errors patterns between different diagnostic groups has been done previously for individuals with Down's syndrome (Gunn & Jarrold, 2004) and also for individuals with Williams syndrome (Van Herwegen et al., 2011), but not for individuals with ASD. We present the following analysis as an initial step into this important area of scientific investigation.

Table 4 summarizes demographic information for the participants who were included in this study. Participants included 106 TD individuals and 153 individuals with autism (AUT). Data were obtained from previous studies done at the Hôpital Rivière-des-Prairies in Montreal, Canada. Participants diagnosed with autism received a best-estimate multidisciplinary diagnosis after evaluation with standard diagnostic instruments, the ADOS and ADI-R (Le Couteur, Lord, & Rutter, 2003; Lord et al., 1999). Five participants in the autism group who had not given an answer for one or more problems on the SPM were excluded from this analysis. One additional participant in the autism group was also excluded, as he had selected answer choice "1" for more than half of the problems.

Participants were grouped into children and adults using a cutoff of 17 years for the maximum age of the children groups. Data available for each participant included age, Wechsler full-scale IQ (FSIQ), and the answer choice given for each SPM problem. Using a two-sample *t*-test, we observed significant group differences in FSIQ and SPM score but not in age for the children, and significant group differences in age and FSIQ but not in SPM score for the adults ($\alpha = 0.05$).

Fig. 7 shows the results of our error analysis for adult participants across the entirety of the SPM test. The differences overall are not that large; the greatest magnitude comes for TD adults who make around 7% fewer Repetition errors than would be expected by chance through random guessing. However, we do see different error patterns overall between the two groups. Adults with ASD seem to make more Repetition errors than TD adults, though still less than what would be expected by chance. Adults with ASD also make slightly more Difference errors than chance or the TD group. Adults with ASD also make fewer errors of type Wrong Principle or Incomplete Correlate, in comparison to the TD group.

Fig. 8 shows the results of our error analysis for child participants across the entirety of the SPM test. The differences overall are smaller here than for the adult data shown in Fig. 7; here, the errors only deviate from chance by about 2%. It is interesting that the strongest group differences in both figures occur for the Repetition type error, and for both children and adults, the ASD group makes more of this type of error.

There are two potential confounds in our analyses that may affect results. First, several studies have found that RPM error patterns often differ as a function of overall ability, i.e. test-takers with lower overall scores not only make more errors but in fact different types of errors than high-scoring participants do. As shown in Table 4, both children and adults in our sample show significant group differences in overall IQ, and the children show significant group differences in RPM score. Thus, it may provide additional insight to compute these error patterns analyses while including IQ and/or SPM score as an additional variable.

Second, our error analysis treats errors on all test problems equally. However, especially given the progressive, set-based organization of the RPM tests, it may be that different groups of test-takers will show certain patterns of errors for certain types of problems. Thus, including some subdivision of test problems as part of the error analysis could also be valuable.

4. Contributions and next steps

The main motivation for this work stems from the view that conceptual types of errors made on the Raven's Progressive Matrices (RPM) family of tests can serve as an important additional measure of behavioral performance, above and beyond total score. To this end, this paper makes two primary contributions.

The first major contribution is the new classification of error types on the SPM using a two-stage approach. This classification should have considerable utility for further studies of human or machine SPM

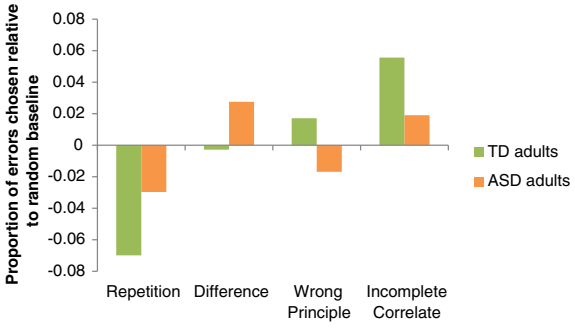


Fig. 7. Proportion of errors of different conceptual types made by adults who are either typically developing (TD) or have been diagnosed with autism spectrum disorder (ASD). This graph shows data aggregated across all SPM problems and in comparison to a random-guessing baseline.

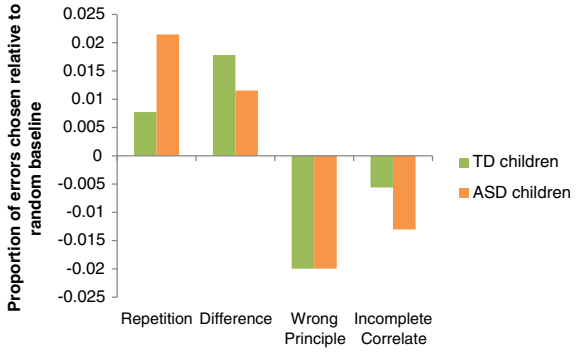


Fig. 8. Proportion of errors of different conceptual types made by children who are either typically developing (TD) or have been diagnosed with autism spectrum disorder (ASD). This graph shows data aggregated across all SPM problems and in comparison to a random-guessing baseline.

performance, and it adds a significant new dimension of information for the RPM family of tests. Both the CPM and APM tests already have such error classifications, which are published in the corresponding test manuals, but the SPM previously did not.

The second major contribution is the method presented for performing group-level analyses of the error patterns exhibited on any RPM test. The method that we present accounts for the per-problem differences in existing error type distributions. This method can be applied to look at the error patterns shown by a single group, in comparison to a random-guessing baseline, or to compare the error patterns made by two or more groups. We have provided two different examples of how this method can be used. The first example analyzes the errors made by a computational model of RPM problem solving, called the ASTI model. The second example analyses the errors made by children and adults who are either typically developing or have been diagnosed with autism spectrum disorder (ASD). Our analysis shows some interesting differences between groups; for example, both children and adults diagnosed with autism seem to make more errors of the Repetition type than the corresponding typically developing groups do.

There are several open issues highlighted by our study that suggest important steps for future research. We discuss next steps in three areas of research: 1) SPM error classification, 2) methods for analyzing RPM error patterns, and 3) computational models of RPM problem solving.

In terms of our **SPM error classification**, while we obtained high inter-rater reliability, this reliability reflects the post-negotiation agreement of just two raters. Additional confidence in the reliability of our coding system could be obtained by conducting studies with new independent raters, and with groups of more than two raters. It would also be valuable to apply the current coding system to other RPM tests, such as the CPM and APM, in order to systematically study how the error classification obtained using our new coding system compares to the previously published error classifications for these tests.

In terms of **methods for analyzing RPM error patterns**, the method we have proposed does not currently include a way to evaluate statistical significance of the results. Also, it only applies to looking at group differences and cannot be used to study individual performance. Further analyses should include overall ability (IQ and/or total SPM score) as additional variables, ideally in a way such that individual error patterns can be correlated with ability as a continuous variable. More fine-grained analyses could also look for possible variations in error patterns shown by a particular group on different subsets of the test. All of these are important directions for future work on RPM error analyses. In addition, having large normative samples of RPM error patterns would be helpful in evaluating the magnitude and implications of particular study results. Finally, as mentioned in Section 1.2, one key observation about RPM errors that calls for further investigation is how properties of the complete set of distracters might influence response patterns. For example, items that have only one or two error types might influence test-takers in a different way than items that have several error types available, for the same total number of distracters. Relatedly, our method does not address the possibility that the correct answer is selected by chance (or that it is selected erroneously, for instance if it also happens to represent a repetition error). Our two-stage coding method does provide a potential route for accounting for these possibilities, since we have a classification of the correct answer in terms of what type of distracter it *would* represent, were it not the correct answer. Finding ways to incorporate these kinds of information about the entire field of available distracters will represent an important refinement of current RPM error analysis methods.

In terms of research on **computational models of RPM problem solving**, looking at the errors made by the ASTI model has led us to propose two steps for future research. First, the model should be able to lessen its direct reliance on matrix entries when generating answer predictions in order to produce more creative, and less visually repetitive, answers. Second, the model should be susceptible to the visual salience of distracters. Neither of these observations would have been possible by looking at the model's total score alone, or even at the pattern of correct vs. incorrect answers. Future work on test-taking by computational models should continue to look at error patterns in order to fully understand model performance and implications for human cognition. In particular, looking at error patterns fosters our understanding not just of whether a particular model can solve each problem, but also *how* it solves a problem and *why* it might be successful or unsuccessful.

Acknowledgements

We are grateful to the US National Science Foundation for its support through NSF (RI) Grant #1116541, "Addressing visual analogy problems on the Raven's intelligence test," and through a GRFP fellowship, and to the US Office of Naval Research for support through an NDSEG fellowship. We also thank Isabelle Simard, Patricia Jelenic, Bryan Wiltgen, and Keith McGreggor for their assistance.

Appendix A. Detailed summary of literature on RPM error patterns

This appendix contains more detailed descriptions of the procedures and results used for each study listed in Section 1.2. Studies are listed here in chronological order.

Miller and Raven (1939) looked at the performance of two groups of children: one group of girls of unspecified school age, and another group of younger children between 5 1/2 and 7 1/2 years of age. Using variations of matrix problems, they established that there are at least two influences on which wrong answers participants choose, in terms of there being non-random effects on the distribution of answers that are chosen. One influence is the absolute position of the answer choice with respect to the matrix. When alternatives are all listed horizontally to the right of the matrix, the position effect is very marked, and participants tend to choose the left-most choices that are closer to the matrix. When alternatives are listed in rows underneath the matrix, position effects are less marked though still present, and participants tend to choose answers from the top row and those towards the middle-right of any particular row (i.e. closer to the empty space in the matrix). The other influence on answer choices is the conceptual type of error represented by the entry given in each respective answer choice, and in particular, for difficult problems, participants tend to make errors of repetition. These two influences are not independent, however; they do interact in complex ways. If a correct answer happens to be in the preferred position, it will be chosen more often than otherwise. Likewise, if an obviously implausible answer is put into this preferred position, participants will go on to examine more alternative choices, whereas if a "familiar" but still incorrect answer is in the preferred position, e.g. a repetition error, participants tend to stick with that answer.

Raven (1939) examined the performance of several different groups of children and adults, including those of typical ability levels as well as those classified as "mentally defective" (p. 16–17). Children under the age of 8 and the mentally defective group were given the CPM in its board form, while all other participants were given the SPM. Some SPM test administrations used mounted displays of SPM problems presented individually by the examiner; group or self-administered sessions used the SPM with a paper scoring form. Participants commonly chose distracters near the blank space in the matrix, especially if these were repetition or incomplete correlate errors, across all ages and ability levels. Raven also noted that when the distracters in this central position were "obviously wrong" (p. 32) and if the correct answer were towards the left, then participants seemed to answer more quickly. Raven mentions error types that correspond to wrong principle and difference errors, e.g., "combining all the characters shown in a matrix" (p. 17). Raven also mentions the choice of distracters based on solution methods that applied to previous matrix problems, which is an intriguing inter-problem phenomena that is rarely, if ever, mentioned in the

rest of the RPM error literature. Finally, Raven notes that most of the errors made by low-ability children were repetition errors, and that these participants often would not provide any answer on the more difficult problems.

Halstead (1943) compared results on the SPM from a clinical group of individuals diagnosed with neuroses with those of healthy controls. In order to examine group differences at a finer level of detail than overall score, Halstead created subgroups individually matched on raw scores. The groups did not differ on measures of "unevenness," i.e. score consistency across sets when compared to norms, or "reversals," i.e. scoring higher on a later set than on an earlier one. He also examined test variables as a function of age, time (for taking the test), attitude, etc. Finally, Halstead looked at the most frequent errors made by a very large number of control participants ($n = 2790$). He broadly classifies these errors according to conceptual type and observed that low ability participants tended to make "perceptual" errors like repetition, whereas high ability participants tended to make "inadequate reasoning" errors. He also observed that: "Mathematically minded subjects seem to do as well as any on the test, and indeed some items in Set E can only be solved logically. High scores have, however, been obtained by artistic people who have an eye for form (Gestalt), symmetry, etc." (p. 211).

Eysenck (1945) looked at the performance of elderly adults with senile dementia, compared to typical adults, on sets A and B of the CPM. She looked at errors in terms of the most frequent distracters chosen and found that in both groups, both the absolute position of distracters as well as distracters that repeat entries from the matrix influence the incorrect answer choices made by participants. In particular, distracters in positions 1, 2, and 6 were more frequently chosen than those in the other positions, but the only group difference was for position 2, which was chosen more frequently by the senile group. For both groups, matching the entry above the empty space accounted for a significant proportion of errors, and matching the entry to the left of the empty space accounted for a smaller, but still significant, proportion of errors.

Sigel (1963) presented a focused argument on the importance of obtaining and using more detailed measures from intelligence tests than just total score, in order to capture variations in strategy as well as ability. He used RPM error analyses as one of two primary examples. He analyzed errors made on individual problems A7 and B7 in groups of boys and girls at 9 and 10 years of age and found that participants often made errors of repetition as well as "perceptual discrimination" (p. 50), though it is not totally clear what he meant by this term. He found some evidence of gender differences in error types, though without conclusive interpretations of these differences, and also observed that error types made did not appear to be related to total score on Set B. In a different group of children consisting of boys between the ages of 7 and 11, Sigel analyzed repetition and position-based errors to find that younger children tended to make more errors of both types, especially when a repetition distracter was located close to the problem matrix.

Bromley (1953) wrote an interesting paper in which he studied performance on the SPM by a group of older individuals with various psychiatric disorders, whom he described as evincing "primitive" forms of thinking. The individuals were instructed to explain their reasoning as they took the test, and Bromley provided qualitative analyses of their responses. His main observation was that there seemed to be two ways to approach the SPM: one the intended way, with abstract, relational, analogical thinking, and another involving more global, holistic, concrete thinking (including mental imagery). He observed that the primitive thinking shown by the test participants fell more into the latter category and could explain many of the errors made on the test. With respect to errors, Bromley observed that the participants seemed susceptible to effects of both the absolute position of answer choices as well as other features of incorrect answer choices (e.g. repetition of a part of the matrix, etc.). He characterized the answer choice error types as: "part of the matrix, simple or distorted figure like the correct one, relatively unrelated figure, global figure, similar to part of the matrix reversed or distorted" (p. 384). The highest proportion of errors was for "part of the matrix" answer choices, followed rather distantly by "simple or distorted figure like the correct one" and "global figure." Bromley also listed the types of thinking that he supposed gave rise to errors on the test, and he emphasized that types of thinking differed significantly on an individual differences level. He also surmised that many of these forms of "primitive" thinking might have developed in an individual (and thus used on a test like the SPM) as a compensatory mechanism, to make up for difficulties with other forms of thinking (e.g. abstract, analogical, etc.).

1) Global responses are those that involve global/Gestalt solutions.
2) Concrete responses are those that fail to adequately abstract from the directly perceived features of the problem.
3) Mechanization of response involves the inability to switch set from initially successful strategies. Bromley points out that the test itself encourages this sort of mechanization (foreshadowing the strategy findings of Kirby and Lawson (1983)); he observes that early problems on the test influence the strategy chosen by participants for later problems. This seems very akin to perseveration.
4) Inability to explain refers to failures in verbalizing a strategy (for successful problems) or the difficulties with a particular problem (for unsuccessful problems).
5) Sensori-motor responses refer to the tendency of participants to point and trace their answer on the matrix and on the answer choices. Bromley observed that on occasion, participants would trace the correct answer but be unable to choose that answer choice.
6) Physiognomic responses.
7) Subjective responses occurred when participants seemed to think there was ambiguity in the answer choices, and the correct answer was a matter of personal preference.
8) Fluid responses were those in which participants seemed to use arbitrary selection criteria, including just picking the "odd man out" among the answer choices.
9) Avoidance of reality referred to participants who picked answer choices and described how they should be different, or to participants who tried to evade the problem by trying to match the frame shapes instead of the entry content.

Forbes (1964), as part of an item analysis to revise the problems found on the APM, analyzed the types of errors made by participants as a function of their ability level (i.e. total APM score). He classified errors as being of four types: incomplete correlate, wrong principle, incomplete individuation, and repetition. His analysis looked at each third of the test with respect to a single ability level: low for the first third, average for the second third, and high for the third third. He noted that the incomplete correlate was the most frequent error type overall, but represented a smaller proportion of errors for the low ability group, for whom wrong principle was the most frequent error. Individuation and repetition errors were the least frequent in any group. He also looked at overall selection of answer choices as a function of position, and found that positions 6 and 7 tended to gain fewer responses than the others, and positions 1 and 4 were the most frequently chosen. He surmised that perhaps 1 is favored by typical scanning patterns, and 4 is closest to the empty space in the matrix.

Weatherick (1966) looked at the errors made by healthy adult subjects on the SPM to directly compare to Bromley's (1953) results with senile psychiatric patients. The subjects were overall high scoring, and he found "very close agreement between our sample of n = 236 and Bromley's sample of n = 35." As a result, Weatherick contends that the specific errors identified by Bromley do not indicate "primitive thought processes." Weatherick does observe that, in instances where his control results did differ from Bromley's results, the senile patients tended to prefer (instead of the most frequent control error) a repetition error, of an answer that repeats a part of the matrix adjacent to the empty space.

Vejleskov (1968) looked at performance on the SPM among Danish children. He gave results on error frequencies for only a few problems, and observed that for these problems (all from Set B), girls tended to fail by choosing the distracter that was the same as the correct response except rotated or flipped.

Carter (1970) supposed that solving SPM problems involved processes of induction as well as evaluating similarity (i.e. similarity of the induced answer to the given answer choices). He gave subjects five tests: regular problems from the SPM (induction + similarity), problems from the SPM in which the answer choices were omitted completely and the answer had to be described (pure induction), tests to rank the similarity of answer choices and matrix entries, according to shared features in a propositional encoding (pure similarity), and problems from other, non-visual tests of induction (pure induction). The similarity rankings of answer choices and matrix items by subjects might have given interesting insight into how they might be viewing the different distracters, but the study only scored them as correct or incorrect in their rankings. Further, the problems that they ranked were not the same as the ones in the first two tests, so it was not possible to see how their perceived rankings might have affected their actual performance on the problem. In fact, while the author designed the two ranking tasks to be different from inductive reasoning, it does seem as though evaluating similarities on a feature-by-feature basis would share a lot in common with solving a matrix task, even one without the answer choices, inasmuch as both tasks involve evaluating differences between entries in a systematic way.

Jacobs and Vandeventer (1970) looked at error patterns on the CPM. In particular, for a given 2 × 2 CPM problem, they classified the answer choices based on whether the answer choice followed a horizontal rule only, a vertical rule only, both (which would be the correct answer), or neither. They assumed that an incorrect answer choice was "superior" if it followed at least the horizontal or vertical rule (as opposed to answer choices that followed neither). They found that 18 of the 36 problems on the CPM contained both "superior" answer choices, and they restricted their analysis to these 18 problems. Then, for each participant, they calculated a proportion Ps that was the number of superior answer choices chosen divided by the total number of wrong answers. (Participants who answered fewer than five problems incorrectly were excludEd.) Looking at data from American children in the first and third grades, Eskimo adults and young adults (from Canada), and Temne adults and young adults (from Sierra Leone), Ps appeared to be more strongly correlated with total number of correct answers in lower-ability groups of participants (i.e. those with lower average scores). In addition, Ps appeared to be higher in the more able groups (i.e. Ps for Eskimos was higher than Ps for Temne). One difficulty in this study is that Ps data from more able participants becomes less valid, because fewer errors have been made to contribute to the Ps score. In addition, Ps was defined solely based on answer choices that followed row or column rules in the matrix; it is a very strong assumption to say that these answer choices were "superior" to the other answer choices. A stronger methodology would require coding classes of distracters for all the answer choices, and then looking at types of errors of the various classes. It could be that the classes of distracters could still be ranked according to their "correctness" level, but that would depend on how the distracter classes were defined.

Guttman (1974) looked at familial correlations in SPM scores among children and their parents. Guttman observed that for each item, two or three of the incorrect answer choices seemed to be chosen with greater frequency. However, she did not observe any inter-family differences in these frequency distributions.

Thissen (1976) characterized incorrect response choices on the SPM according to frequency: the first most-chosen, second most-chosen, and then all other incorrect answer choices. He used this information to calculate a latent trait model for each test item that gave the probability of choosing a particular answer choice as a function of ability (the latent, unobserved trait). He found that for different problems, the answer choices behaved differently for different levels of ability, but the analysis was purely done along this unidimensional notion of ability; no explanations were offered for why certain answer choices might be more or less chosen than others.

Horner and Nailling (1980) adapted a listing of error types from Raven (1965) and present a listing of the error type for each answer choice in the CPM. In a study of left-, right-, and non-brain-damaged patients, they found that each group showed nearly identical patterns of error types across the four error types. In particular, only one type of error, "repetition of a pattern," seemed to be made with any considerable frequency, other than the correct answer.

Vodegel Matzen, van der Molen, and Dudink (1994) looked at types of errors made on the SPM by typically developing children. They adopted error categories from the APM to categorize SPM errors as: incomplete correlate, wrong principle, repetition, or additional elements, and they only analyzed sets C through E. Inter-rater reliability for coding the error types was only around 72%. They found that the incomplete correlate was most frequent overall, followed by wrong principle, repetition, and additional elements. When they divided the subjects by ability level, they found that low ability subjects made relatively more errors of the incomplete correlate and repetition types. The authors then devised an "experimental progressive matrices" test, in which all errors were of the "incomplete correlate" type, but they varied according to how many rules were omitted to generate that answer choice. The problems varied according to rule type and number, following Carpenter et al. (1990). The EPM was similar in difficulty to the SPM, and they found that most error choices (for any ability level) were made due to the omission of a single rule. Furthermore, the rules increased in difficulty (as measured by number of errors) in this order, for all ability levels: constant in a row, quantitative pairwise progression, distribution of three values, addition/subtraction, distribution of two values. This is the same rule ordering that was chosen by Carpenter et al. (1990) for inducing rules.

Van der Ven and Ellis (2000) looked at the most frequent incorrect answer choice for the SPM in sets B, C, and E, in order to determine what factors these problems might load upon. They identified different types of errors, including: "lack of completeness of analogical reasoning," "freedom from perceptual distracters," and "coping." They also present data from sets C and E giving the frequencies of each answer choice for each problem, using their sample of several hundred Dutch schoolchildren.

Babcock (2002) classified each answer choice from the APM as being one of four different error types: incomplete correlate, wrong principle, confluence of ideas, or repetition. She studies the error responses made by adults of varying age and ability, according to whether their frequency of making a particular type of error was above or below chance levels. She found that adults of varying ages tended to make similar types of errors, but adults of high ability made different errors than those of low ability. In particular, high ability adults tended to make more incomplete correlate errors, and few errors of other types. Lower ability adults tended to make each type of error at chance levels. Also, she studied errors as a function of rule type, based on Carpenter et al. (1990), and found some differences between subjects of varying abilities.

Gunn & Jarrold (2004) looked at types of errors made by TD children, children with moderate learning disabilities (MLD), and children with Down syndrome (DS) on the CPM. They classified error choices as being of one of four types, following the CPM manual: difference, repetition of a figure, inadequate individuation, and incomplete correlates. They found that, even after controlling for total number of errors, the DS group made different types of errors than the other two groups. In particular, the DS group produced fewer repetition of a figure errors and more inadequate individuation errors and difference errors (which is choosing an unrelated answer choice). Furthermore, the pattern of errors produced by the DS group is similar to that shown by younger TD children, even in cases where the DS group shows better performance than younger TD children. The authors surmise that

individuals with DS may have either difficulty in combining features to produce the target pattern, difficulty in visual discrimination, or less rigor in choosing their final response, in the case of incomplete or partial solutions.

Fajgelj, Bala, and Katic (2010) as part of a factor analysis of the CPM, looked at types of errors made by their sample of Serbian children. They found that for younger children, more CPM problems had certain distracters that were chosen by significant portions of subjects (i.e. > 20%). They observed that the most common distracters involved choosing the answer identical to the entry to the left of the empty space or above the empty space in the matrix. They also note, interestingly, that number 2 was chosen more frequently than other answer choices, and especially so for younger children, possibly because this choice is spatially closest to the empty spot in the matrix.

Matzen et al. (2010) performed an analysis of errors on the SPM and on artificial SPM-like items. On the artificial items, errors were classified systematically according to how each distracter was related to content in the problem matrix. In particular, error types were classified as (for a single relation in the problem): match to diagonal, match to top left, match to adjacent, flanker, and unclassified. They were able to categorize some, but not all, SPM errors using the same scheme (i.e. the one-relation SPM problems but not the two-relation problems). For certain problems, they found that the error type seemed to have a relationship to the direction of the relation in the problem; for instance, problems that were diagonal in one direction tended to have more "match to adjacent" errors, whereas problems that were diagonal in the other direction tended to have more "flanker" errors. Though the authors do not draw this connection, it seems as though participants might have been distracted by Gestalt properties of the overall matrix in making such errors.

Van Herwegen, Farran, and Annaz (2011) looked at error types on the CPM between TD children and individuals with Williams syndrome (WS). They classified errors on the CPM following the CPM manual into four categories: difference, inadequate individuation, repetition, and incomplete correlation. They looked at proportion of each error type out of total error for each participant. Participants were matched on CPM raw score, and the WS group had a much higher mean chronological age than did the TD group. Their results were very similar to those in Gunn & Jarrold (2004), in the proportions of each type of error made, on average, though they found no group differences in this study between the WS and TD individuals. They also studied developmental effects on error type, and again found similar results to Gunn & Jarrold (2004), in that the difference and inadequate individuation errors decreased and repetition errors increased; however, incomplete correlation errors did not increase with age. They also did an item analysis, following Facon and Nuchadee (2010), to look at whether items differed in difficulty between the two groups. Only 3 of the 36 items differed. They close with speculating that one might expect to see different patterns autism, since autism has perceptual atypicalities more so than WS and the RPM is a perceptual task.

Appendix B. SPM error coding protocol

This appendix contains a copy of the protocol provided to the two human coders to classify distracters on the SPM according to error type.

(Protocol Page 1) Overview: Error type classification on the SPM:

There are four basic conceptual types of errors on the Raven's Standard Progressive Matrices Test: 1) incomplete correlate, 2) repetition, 3) difference, and 4) wrong principle.

Incomplete correlate (IC) errors are those in which the distracter is almost, but not quite, correct. For example, some IC distracters represent a rotation or reflection of the correct answer. Other IC distracters differ from the correct answer in a single feature dimension, e.g. they might have four elements instead of three, or straight elements instead of curvy ones, or have the correct shape but the wrong texture. Alternately, an IC distracter might be only missing an element from the correct answer. Oftentimes, an IC distracter might be correct in terms of a single row or column in the matrix, e.g. looking just at the right-most column or just at the bottom-most row, but when both rows and columns are taken into account, it no longer fits the matrix pattern. These kinds of errors are made when a test-taker more or less "gets" the problem, in terms of identifying and understanding the relevant matrix relationships, but then fails to fully account for all of the problem details when selecting an answer.

Repetition (RP) errors are those in which the distracter is a copy of one of the matrix entries that is adjacent to the blank space. Choosing an RP distracter may represent a sort of cognitive bias or fixation on the matrix entries, in which an answer is selected based on simple perceptual matching between the answer choices and the matrix entries closest to the blank space. These entries may be privileged because of their proximity to the blank space. Alternately, assuming a top-left to bottom-right visual scanning pattern, these adjacent entries may be the last viewed before the test-taker moves on to look at the answer choices, assuming a sequential inspection of the problem in a matrix-first, answers-second ordering.

Difference (DF) errors are those in which the distracter is somehow qualitatively different in appearance from the other distracters. DF distracters include those that are completely blank, as well as those that have extraneous shapes that are not found anywhere in the problem matrix. In addition, a DF distracter is often the most complex-seeming answer choice, either combining all of the matrix entries together into a single agglomeration of matrix elements or taking some feature from the matrix and increasing its value until it surpasses all the other entries and answer choices. A DF distracter might be chosen because it visually "pops" from among the other answer choices.

Wrong principle (WP) errors are those in which the distracter is a copy of or composition of various elements from various matrix entries (with the exception of copies of adjacent entries, which would still fall under the "repetition" error type). A WP distracter might be chosen if the test-taker does not successively educe the correct relationship from the matrix entries and instead combines the entries according to some other rule or relationship to produce an answer choice.

(Protocol Page 2) The following table gives specific criteria that can be used to distinguish among the various conceptual types of errors found on the SPM.

Error type taxonomy and classification criteria for the SPM

Error type		Criteria
Incomplete correlate	1	Negative (color-inversion) of correct answer
	2	Change only in fill, texture, or style
	3	Rotation/reflection of correct answer
	4	Change only in spatial layout of elements
	5	Change only in size or scale
	6	Change only in number of discrete elements
	7	Incomplete, with missing element or portion
Repetition	8	Repetition of matrix entry to left of blank space
	9	Repetition of matrix entry above blank space
	10	Repetition of matrix entry to top-left of blank space
Difference	11	Filled completely white or black
	12	Union or agglomeration of all or most matrix entries
	13	Maximizes some feature value
	14	Differs qualitatively from matrix and other answers, or contains information not found anywhere in matrix
Wrong principle	15	Repetition of matrix entry not adjacent to blank space
	16	Rotation/reflection of matrix entry
	17	Transformation/combination of matrix entries

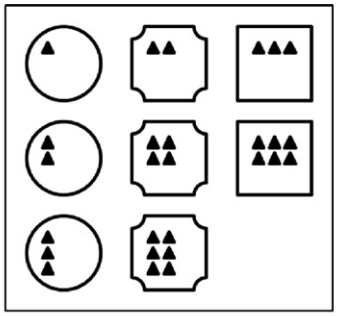
Note that these error type criteria can be broadly divided into two categories:

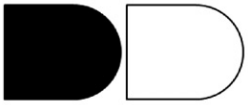
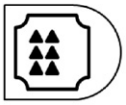
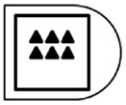
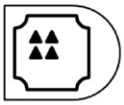
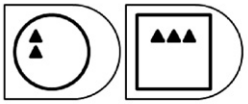
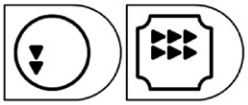
1) If an individual does not know or guess the correct answer, even partially, then they may make repetition, difference, or wrong principle errors. The distracters that represent these error types can be identified based on how the distracter is related to information in the matrix.
2) If an individual does partially guess the correct answer, then they may make incomplete correlate errors. The distracters that represent these error types can be identified based on how the distracter is related to the correct answer choice.

Therefore, the scheme for coding error types on the SPM actually has two parts:

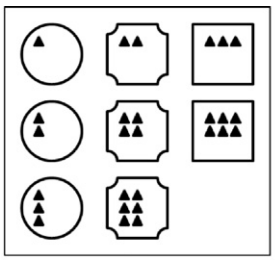
1) Without consideration of the correct answer, first code each answer choice in terms of criteria #8–17, which represent how each answer is related to information in the matrix.
2) Then, without consideration of the matrix, using only knowledge of the correct answer, code each answer choice in terms of criteria #1–7, which represent how each answer is related to the correct answer.

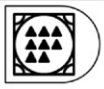
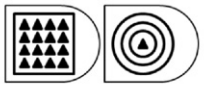
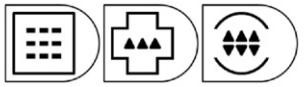
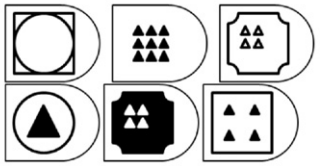
(Protocol Page 3) Here is an example problem, along with examples of the kinds of distracters that fall under each error type criterion.



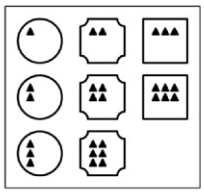
F	Mark any answer choice that is filled completely white or completely black.	
L	Mark any answer choice that is a repetition of the matrix entry directly to the left of the blank space.	
T	Mark any answer choice that is a repetition of the matrix entry directly to the top of the blank space.	
D	Mark any answer choice that is a repetition of the matrix entry directly to the diagonal top-left of the blank space.	
C	Mark any answer choice that is a copy of any matrix entry that is not directly adjacent to the blank space.	
↻	Mark any answer choice that is a rotation or reflection of any matrix entry.	

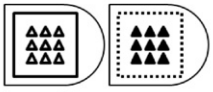
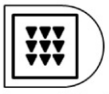
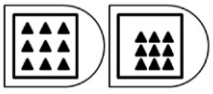
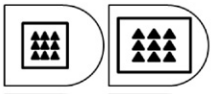
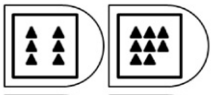
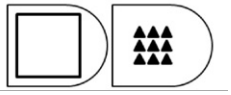
(Protocol Page 4)



U	Mark any answer choice that is a union or agglomeration of all or most of the matrix entries (or aspects of them).	
+	Mark any answer choice in which some particular feature found in the matrix is maximized or made more complex.	
X	Mark any answer choice that contains new content not found in the matrix or other answer choices or is otherwise qualitatively different from the other answer choices.	
M	Mark any answer choice that represents any other transformation or combination of matrix entries.	

(Protocol Page 5)



④	Mark the correct answers by circling the number choice.	
N	Mark any answer choice that is a negative (color-inverted) image of the correct answer.	
F	Mark any answer choice that is the same as the correct answer except with a change only in fill, texture, or style.	
↻	Mark any answer choice that is a rotation or reflection of the correct answer.	
L	Mark any answer choice that is the same as the correct answer except with a change only in spatial layout of elements.	
S	Mark any answer choice that is the same as the correct answer except with a change only in size or scale.	
#	Mark any answer choice that is the same as the correct answer except with a change only in number of elements.	
–	Mark any answer choice that is the same as the correct answer but with a missing element or portion.	

(Protocol Page 6) Part 1: How answer choices are related to content in the problem matrix:

Instructions: Part 1 uses Test Booklet A, which contains the complete matrix and answers for each problem.

First, using the six codes found in Step 1 in the table below, go through each problem in the test and mark any answers that fit these six criteria. If more than one criterion fits a particular answer choice, just mark the first one that applies using the order specified in the table below.

Then, go through the test once more, this time marking answers that fit the codes and criteria listed in Step 2. If an answer choice has already been marked during Step 1, skip it. At the end, each answer choice should have exactly one code assigned to it.

Step	Code	Criteria
1)	F	Mark any answer choice that is filled completely white or completely black.
	L	Mark any answer choice that is a repetition of the matrix entry directly to the left of the blank space.
	T	Mark any answer choice that is a repetition of the matrix entry directly to the top of the blank space.
	D	Mark any answer choice that is a repetition of the matrix entry directly to the diagonal top-left of the blank space.
	C	Mark any answer choice that is a copy of any matrix entry that is not directly adjacent to the blank space.
	↻	Mark any answer choice that is a rotation or reflection of any matrix entry.
2)	U	Mark any answer choice that is a union or agglomeration of all or most of the matrix entries (or aspects of them).
	+	Mark any answer choice in which some particular feature found in the matrix is maximized or made more complex.
	X	Mark any answer choice that contains new content not found in the matrix or other answer choices or is otherwise qualitatively different from the other answer choices.
	M	Mark any answer choice that represents any other transformation or combination of matrix entries.

(Protocol Page 7) Part 2: How answer choices are related to correct answer:

Instructions: Part 1 uses Test Booklet B, which contains only the answers for each problem.

Step 0 has already been completed; the correct answers have been marked by circling the number of the appropriate choice.

Using the seven codes found in Step 1 in the table below, go through each problem in the test and, for each answer choice other than the correct one, mark any answers that fit these seven criteria. If more than one criterion fits a particular answer choice, just mark the first one that applies using the order specified in the table below. Not all answer choices need to be marked; if an answer choice fits none of these seven criteria, then just leave it blank. At the end, each answer choice (excluding the correct answer) should have zero or one codes assigned to it.

Step	Code	Task
0)	④	Mark the correct answers by circling the number choice.
1)	N	Mark any answer choice that is a negative (color-inverted) image of the correct answer.
	F	Mark any answer choice that is the same as the correct answer except with a change only in fill, texture, or style. Mark any answer choice that is a rotation or reflection of the correct answer.
	L	Mark any answer choice that is the same as the correct answer except with a change only in spatial layout of elements.
	S	Mark any answer choice that is the same as the correct answer except with a change only in size or scale.
	#	Mark any answer choice that is the same as the correct answer except with a change only in number of discrete elements.
	I	Mark any answer choice that is the same as the correct answer but is incomplete, with a missing element or portion.

References

Babcock, R. L. (2002). Analysis of age differences in types of errors on the Raven's advanced progressive matrices. *Intelligence, 30*(6), 485–503.

Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*(3), 205–238.

Bölte, S., Dziobek, I., & Poustka, F. (2009). Brief report: The level and nature of autistic intelligence revisited. *Journal of Autism and Developmental Disorders, 39*(4), 678–682.

Bromley, D. B. (1953). Primitive forms of response to the matrices test. *The British Journal of Psychiatry, 99*(416), 374–393.

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*(3), 404.

Carter, D. E. (1970). *An exploratory investigation of the processes involved in the completion of Raven's Progressive Matrices.* M.A. Thesis: Department of Psychology and Sociology, University of Canterbury.

Cohen, P. R., & Howe, A. E. (1988). How evaluation guides AI research: The message still counts more than the medium. *AI Magazine, 9*(4), 35.

Dawson, M., Soulières, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science, 18*(8), 657–662.

Embretson, S. (2004). Measuring human intelligence with artificial intelligence: Adaptive item generation. In R. J. Sternberg, & J. E. Pretz (Eds.), *Cognition and intelligence: Identifying the mechanisms of the mind* (pp. 251–267). New York, NY: Cambridge University Press.

Eysenck, M. D. (1945). A study of certain qualitative aspects of problem-solving in senile-dementia patients. *Journal of Mental Science, 91*, 337–345.

Facon, B., & Nuchadee, M. L. (2010). An item analysis of Raven's colored progressive matrices among participants with down syndrome. *Research in Developmental Disabilities, 31*, 243–249.

Fajgelj, S., Bala, G., & Katic, R. (2010). Latent structure of Raven's colored progressive matrices. *Collegium Anthropologicum, 34*, 1015–1026.

Fancher, R. E. (1985). *The intelligence men.* New York: Norton & Company.

Forbes, A. R. (1964). An item analysis of the advanced matrices. *British Journal of Educational Psychology, 34*, 223–236.

Goddard, H. H. (1920). *Human efficiency and levels of intelligence.* Princeton University Press.

Green, K. E., & Kluever, R. C. (1992). Components of item difficulty of Raven's matrices. *Journal of General Psychology, 119*(2), 189–199.

Gunn, D. M., & Jarrold, C. (2004). Raven's matrices performance in down syndrome: Evidence of unusual errors. *Research in Developmental Disabilities, 25*(5), 443–457.

Guttman, R. (1974). Genetic analysis of analytical spatial ability: Raven's Progressive Matrices. *Behavior Genetics, 4*, 274–284.

Halstead, H. (1943). An analysis of matrix test results. *Journal of Mental Science, 89*, 202–215.

Horner, J., & Nailling, K. (1980). *Raven's colored progressive matrices: Interpreting results through analysis of problem-type and error-type.* Clinical Aphasiology Conference: BRK Publishers, Minneapolis, MN.

Hunt, E. (1974). Quote the raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and cognition.* Erlbaum Associates: Hillsdale, NJ.

Hunt, E. (1980). Intelligence as an information processing concept. *British Journal of Psychology, 71*, 449–474.

Jacobs, P. I., & Vandeventer, M. (1970). Information in wrong responses. *Psychological Reports, 26*, 311–315.

Keating, D. P. (1984). The Emperor's new clothes: The new look in intelligence research. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence. Vol. 2.* (pp. 1–46). Lawrence Erlbaum Associates: Hillsdale.

Keating, D. P., & Bobbitt, B. L. (1978). Individual and developmental differences in cognitive-processing components of mental ability. *Child Development, 49*(1), 155–167. http://dx.doi.org/10.2307/1128604.

Kirby, J. R., & Lawson, M. J. (1983). Effects of strategy training on progressive matrices performance. *Contemporary Educational Psychology, 8*(2), 127–140.

Kunda, M. (2013). *Visual problem solving in autism, psychometrics, and AI: The case of the Raven's Progressive Matrices intelligence test (doctoral dissertation).* Atlanta, GA: Georgia Institute of Technology.

Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research, 22-23*, 47–66.

Kunda, M., & Goel, A. K. (2011). Thinking in pictures as a cognitive account of autism. *Journal of Autism and Developmental Disorders, 41*(9), 1157–1177.

Le Couteur, A., Lord, C., & Rutter, M. (2003). *The autism diagnostic interview-revised (ADI-R).* Los Angeles, CA: Western Psychological Services.

Little, D. R., Lewandowsky, S., & Griffiths, T. (2012). A Bayesian model of rule induction in Raven's Progressive Matrices. *In Proceedings of the 34th annual conference of the cognitive science society*, 1918–1923.

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *Autism diagnostic observation schedule-WPS edition.* Los Angeles, CA: Western Psychological Services.

Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. *In Proceedings of CogSci, 10*, 2761–2766.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's standard progressive matrices. *Intelligence, 32*(4), 411–424.

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*(2), 525–541.

Miller, F. M., & Raven, J. C. (1939). The influence of positional factors on the choice of answers to perceptual intelligence tests. *British Journal of Medical Psychology, 18*, 35–39.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.
Mottron, L. (2004). Matching strategies in cognitive research with individuals with high-functioning autism: Current practices, instrument biases, and recommendations. *Journal of Autism and Developmental Disorders, 34*(1), 19–27.
Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*(2), 252–284.
Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence, 30*(1), 41–70.
Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science, 3*(1), 140–153.
Raven, J. C. (1936). *Mental tests used in genetic studies: The performance of related individuals on tests mainly educative and mainly reproductive.* University of London: Unpublished Master's Thesis.
Raven, J. C. (1939). The R. E. C. I. series of perceptual tests: An experimental survey. *British Journal of Medical Psychology, 18*(1), 16–34.
Raven, J. C. (1965). *Guide to Using the Coloured Progressive Matrices.* London: H.K. Lewis & Co., Ltd.
Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and vocabulary scales.* Pearson.
Schenck, C., Sinapov, J., Johnston, D., & Stoytchev, A. (2014). Which object fits best? Solving matrix completion tasks with a humanoid robot. *IEEE Transactions on Autonomous Mental Development, 6*(3), 226–240.
Sigel, I. E. (1963). How intelligence tests limit understanding of intelligence. *Merrill-Palmer Quarterly of Behavior and Development, 9*(1), 39–56.
Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. *Advances in the Psychology of Human Intelligence, 2*, 47–103.
Soulières, I., Dawson, M., Samson, F., Barbeau, E. B., Sahyoun, C. P., Strangman, G. E., ... Mottron, L. (2009). Enhanced visual processing contributes to matrix reasoning in autism. *Human Brain Mapping, 30*(12), 4082–4107.
Thissen, D. M. (1976). Information in wrong responses to the Raven's Progressive Matrices. *Journal of Educational Measurement, 13*, 201–214.
Van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences, 29*(1), 45–64.
Van Herwegen, J., Farran, E., & Annaz, D. (2011). Item and error analysis on Raven's Coloured progressive matrices in Williams syndrome. *Research in Developmental Disabilities, 32*(1), 93–99.
Vejleskov, H. (1968). An analysis of raven matrix responses in fifth grade children. *Scandinavian Journal of Psychology, 9*(1), 177–186.
Vodegel Matzen, L. B. L., van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of raven test performance. *Personality and Individual Differences, 16*(3), 433–445.
Weatherick, N. E. (1966). The responses of normal adult subjects to the matrices test. *British Journal of Psychology, 57*(3–4), 297–300.
White, A. P., & Zammarelli, J. E. (1981). Convergence principles: Information in the answer sets of some multiple-choice intelligence tests. *Applied Psychological Measurement, 5*(1), 21–27.