Methods for Classifying Errors on the Raven's Standard Progressive Matrices Test

Maithilee Kunda (mkunda@gatech.edu)[1]
Isabelle Soulières (soulieres.isabelle@uqam.ca)[2]
Agata Rozga (agata@gatech.edu)[1]
Ashok K. Goel (goel@cc.gatech.edu)[1]

[1]School of Interactive Computing, Georgia Tech,
85 Fifth Street NW, Atlanta, GA 30308 USA
[2]Département de Psychologie, Université du Québec à Montréal
C.P. 8888 succursale Centre-ville, Montréal (Québec) H3C 3P8 Canada

Abstract

Although many psychometric tests, like Raven's Progressive Matrices, are commonly evaluated according to total score, additional variables can lend insight into the underlying cognitive processes. We examine conceptual errors on the Raven's Standard Progressive Matrices (SPM) test. We present a complete classification of error types on the SPM using a two-kind coding scheme, yielding ≥ 95% inter-rater reliability. We also examine how to extract error data from a computational model, and we present a method for measuring errors through systematic ablation to create a "population" of models whose performance can be examined as a group. We present a preliminary analysis of error patterns on the SPM from typically developing individuals, individuals diagnosed with autism, and a computational model called ASTI. We discuss what the error patterns suggest regarding cognition on the SPM and routes towards improving the ASTI model.

Keywords: ablation experiments; computational modeling; error patterns; mental imagery; psychometrics; Raven's Progressive Matrices; visual representations.

Introduction

Raven's Progressive Matrices (RPM) is a widely used series of intelligence tests that consist of multiple choice visual analogy problems, as in Fig. 1. Each problem contains a matrix of geometric figures with one figure missing; the correct missing figure that completes the matrix pattern must be selected from a set of answer choices.

Performance is generally measured in terms of overall score, i.e. number correct, which can then be used as an index into normative test data to determine an IQ score or percentile ranking for that individual. While total score is certainly an important variable, serving as a coarse measure of an individual's overall ability, there are alternative dimensions of performance that may provide a finer-grained view of an individual's cognitive processing:

1) Per-item accuracy, e.g. differential item functioning, takes into account potential variation even when individuals may obtain the same total score (Facon & Nuchadee, 2010; Lynn, Alik, & Irwing, 2004; Van Herwegen, Farran, and Annaz, 2011).
2) Reaction time can be used to understand the stages of processing in solving a single item (Bethell-Fox, Lohman, & Snow, 1984) or to compare performance across individuals or groups (Soulières et al., 2009).
3) Patterns of errors—for a problem answered incorrectly, which of the given distracters is selected?—have been studied as a window into cognitive strategy (Bromley, 1953; Gunn & Jarrold, 2004; Miller & Raven, 1939; Van Herwegen, Farran, and Annaz, 2011; Vodegel Matzen et al., 1994).

All of these dimensions represent measurable aspects of the "output" of a human cognitive system taking the RPM test. The "input" to such a system, in addition to the test itself, can be conceptualized as the set of cognitive functions drawn upon while solving the test. Unlike the output measures, it is difficult to directly measure cognitive functioning. Some studies have used eye-tracking as a measure of visual attention (Bethell-Fox, et al., 1984; Carpenter, Just, & Shell, 1990), and some have used verbal reporting protocols (Carpenter et al., 1990) though verbal report may bias the cognitive strategies used by participants (DeShon, Chan, & Weissbein, 1995).

Another way to elucidate these invisible cognitive mechanisms is to construct computational models of various aspects of RPM problem solving and then inspect these models in relation to human behavioral data. Aspects of RPM (or RPM-like) problem solving that have been investigated using computational models include:

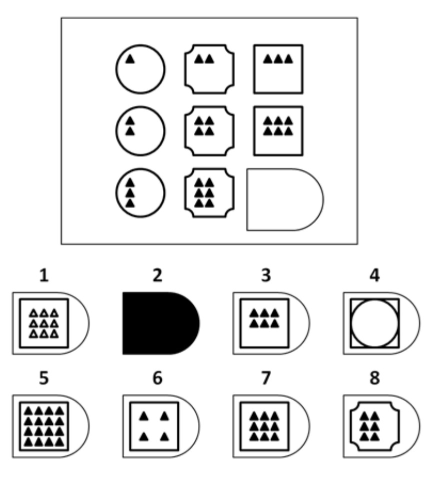

Figure 1: Example of an RPM-like problem.

1) Knowledge representation, i.e. visual versus verbal representations of problem content (Hunt, 1974; Kunda, Goel, & McGreggor, 2013; McGreggor, Kunda, & Goel, 2011).
2) Goal-subgoal maintenance (Carpenter et al., 1990).
3) Problem-solving process, i.e. constructive matching (mentally constructing the answer and then selecting an answer choice) versus response elimination (inspecting each answer choice to find the best fit) (Bethell-Fox et al., 1984; Lovett & Forbus, 2012).
4) Answer selection process in terms of confidence (McGreggor & Goel, 2012) or probability (Little, Lewandowsky, & Griffiths, 2012).

However, in the extant literature on computational models of the RPM, many models tend to focus on only one measure of output performance: total score. We believe it is not only valuable but critical that models examine the other dimensions of "output" that we have mentioned, in order to investigate how models relate to human cognition at increasingly fine-grained levels of resolution.

In this paper, we focus on one such "output" measure—error patterns—and one computational model—the ASTI model, described in detail in a previous publication (Kunda et al., 2013). We first present an operationalization of error patterns on the Raven's Standard Progressive Matrices (SPM) test, in the form of a two-kind classification of conceptual error types. Then, we briefly summarize the algorithms and performance of the ASTI model. Finally, we present a method for analyzing the errors made by a computational model, and we give preliminary results based on a comparison of the errors made by the ASTI model against human error data from typically developing individuals and individuals diagnosed with autism, along with an evaluation of what these differences in error patterns can tell us about cognitive processing on the RPM.

Types of Conceptual Errors on the SPM

One way to examine errors on an RPM test is to look at which distracter is chosen in comparison to those most frequently chosen (Thissen, 1976; van der Ven & Ellis, 2000). However, many studies have shown that errors can also be classified according to conceptual type, which may provide additional insight into what it means when a certain error is made (Forbes, 1964; Horner & Nailling, 1980).

However, there is currently one significant barrier to the widespread analysis of error patterns on the SPM test; while the published manuals for two of the RPM tests, the Colored Progressive Matrices (CPM) and the Advanced Progressive Matrices (APM), include taxonomies of conceptual error types, the manual for the Standard Progressive Matrices (SPM) does not (Raven, Raven, & Court, 2003). Vodegel Matzen et al. (1994) attempted to use the APM error type classifications on a portion of the SPM, but inter-coder reliability reached only about 70%. The authors concluded that classification of SPM distracters seemed "problematic" in that there did not seem to be a systematic methodology used for constructing distracters.

The taxonomies given in the CPM and APM manuals (Raven et al., 2003), although having different labels, seem to represent the same four notions of error types. We now present a synthesized description of these four error types which, along with criteria used to classify a particular distracter, are also summarized in Table 1.

1) **Incomplete correlate (IC) errors** occur when the chosen distracter is almost, but not quite, correct. For example, some IC distracters have the correct shape but the wrong texture, as exemplified by distracter #1 in Fig. 1. These kinds of errors are made when a test-taker more or less "gets" the problem, in terms of identifying the relevant matrix relationships, but then fails to fully account for all of the details when selecting an answer.
2) **Repetition (R) errors** occur when the chosen distracter copies a matrix entry adjacent to the blank space, as shown by distracters #3 and #8 in Fig. 1. The choice of an R distracter may represent perseveration or fixation on the matrix entries, in which an answer is selected via perceptual matching between the answer choices and the matrix entries closest to the blank space.
3) **Difference (D) errors** occur when the chosen distracter is qualitatively different in appearance from the other choices. D distracters include completely blank entries, as exemplified by distracter #2 in Fig. 1, as well as those that have extraneous or complex shapes not found in the matrix. A D distracter might be chosen because it visually "pops" from among the other choices.
4) **Wrong principle (WP) errors** occur when the chosen distracter is a copy or composition of elements from various matrix entries, as exemplified by distracters #4 and #6 in Fig. 1. A WP distracter might be chosen if the test-taker fails to educe the correct relationship from the matrix and combines the entries according to some other rule or relationship to produce an answer.

Two-Kind Taxonomy and Coding Results

The main difficulty we observed in coding SPM distracters is that the same distracter often seems to fall under multiple categories, e.g. it might represent a repetition as well as an incomplete correlate; this difficulty was shared by Vodegel-Matzen, et al. (1994). From this observation, we realized that the four error types listed above actually represent *two* orthogonal classifications of distracters:

Kind I: Relationship of distracter to matrix entries: Repetition, difference, and wrong principle errors all have to do with how a distracter is related to information in the matrix and in the other answer choices, without any regard to the content of the correct answer choice. In particular, errors of the first kind assume the participant is attending to irrelevant or erroneous aspects of the problem, and that they are not able to discover even a partial solution.

Kind II: Relationship of distracter to correct answer: Incomplete correlate errors have to do with how a particular distracter is related to the correct answer choice. These errors assume the participant correctly guesses some part of the solution but does not quite attain the correct answer.

Table 1: Criteria for classifying distracters on the SPM.

Error type	#	Criteria
Kind I: Repetition	1	Repetition of matrix entry to left of blank space
	2	Repetition of matrix entry above blank space
	3	Repetition of matrix entry to top-left of blank space
Kind I: Difference	4	Filled completely white or black
	5	Union of matrix entries or aspects of them, so that union has more components than any single matrix entry
	6	Maximizes some feature value or makes it more complex
	7	Differs qualitatively from matrix and other answers, or contains information not found anywhere in matrix
Kind I: Wrong Principle	8	Copy of matrix entry not adjacent to blank space
	9	Rotation/reflection of matrix entry
	10	Other transformations or combinations of matrix entries or aspects of them, including negative images
Kind II: Incomplete Correlate	11	Negative (color-inversion) of correct answer
	12	Change only in fill, texture, or style
	13	Rotation/reflection of correct answer
	14	Change only in spatial layout of elements
	15	Change only in size or scale, in either or both dimensions (allowing for feature-wise scaling)
	16	Change only in number of discrete elements (allowing for slight changes in layout)
	17	Incomplete, with missing element or portion

Using this two-kind taxonomy, two raters independently coded all 432 distracters on the SPM[1] in two separate passes, first for Kind I and then for Kind II. Kind I classification used a copy of the test booklet in which no answers had been marked, and raters assigned every distracter to one of categories #1-10 in Table 1. Kind II classification used another test booklet 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; raters assigned each distracter to one of categories #11-17 in Table 1, or left it uncategorized.

Initial agreement between the two raters was 82% for Kind I errors and (coincidentally) 82% for Kind II errors. Kappa coefficients were calculated to test for independence between raters. The kappa values were 0.79 for Kind I errors and 0.67 for Kind II errors.

Discrepancies were resolved during a negotiation phase between the two raters. Each discrepancy was discussed, and each rater presented a rationale for the classification. It was found that there were several systematic discrepancies easily resolved by making the coding criteria more specific. For example, Criterion #5 in Table 1 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. Table 1 shows the final coding criteria, after these changes had been incorporated.

After the negotiation phase, rater agreement was recalculated. Post-negotiation agreement was 95% for Kind I errors and 98% for Kind II errors. Remaining differences were resolved by the primary rater based on consideration of the conceptual error type intended to be captured.

Fig. 2 shows the overall proportions of error types across all distracters of the SPM. Interestingly, there is roughly the same proportion of incomplete correlate distracters as correct answers, and all remaining distracters are divided nearly evenly among the three remaining error types.

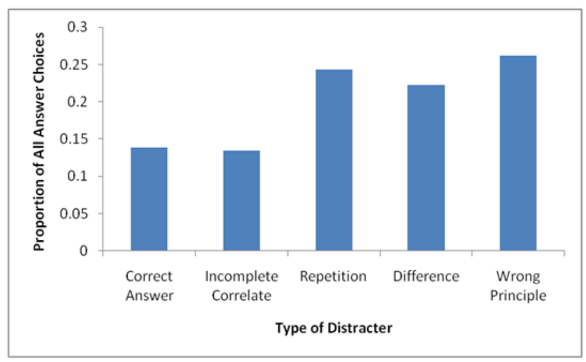

Figure 2: Proportions of each error type on the SPM.

The ASTI Model

In previous work (Kunda et al., 2013), we presented a computational model of problem solving on the RPM, 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 extant computational RPM models had previously 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).

[1] Each of the first 24 problems on the SPM has 6 answer choices, and each of the latter 36 problems has 8 answer choices.

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; 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.

```
Initialization
1   Read matrix entries into list of images M
2   Read answer choices into list of images A
3   For any two images a and b, define a
    similarity metric S(a, b) → z ∈ [0, 1]
4   Define set of base transforms T
5   Define set of analogies I_0 → I_1, where I_0
    contains image sequences representing
    complete row, column, or diagonal lines in
    the matrix, and for each i_0 ∈ I_0, I_1 has
    the corresponding images i_1 representing
    the parallel partial line in the matrix
Transformation Induction
1   For each image sequence i_0 ∈ I_0, induce the
    best-fit composite transform t_C:
2       For each base transform t ∈ T:
3           Apply t to the first image(s) in i_0
            to produce image i_t
4           Search all possible translation
            offsets (x, y) between i_0 and i_t to
            find the best match, as calculated
            by S(i_0(x,y), i_t)
5           Select the best-fit base transform
            t_B as per S, as calculated above
6       t_C is then a composition of t_B and the
        translation offset (x, y)
7   Obtain a final transform t_F by selecting
    that t_C which produces the best average
    fit, across each subset of parallel i_0 ∈ I_0
Candidate Prediction and Answer Selection
1   Choose image sequence i_0 that results in
    the best-fit t_F, according to S as
    calculated in the previous step
2   Apply t_F to corresponding partial image
    sequence i_1 ∈ I_1 to produce candidate
    answer image i_C
3   For each answer choice i_A ∈ A, compute
    similarity S(i_C, i_A)
4   Select the best-fit answer choice iA as
    per S, as calculated above
```

Figure 3: Algorithm used by the ASTI model.

Obtaining Error Data from the ASTI Model

The current version of the ASTI model correctly answers 50 out of 60 problems on the SPM. 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). 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 removing access to subsets of these mechanisms, we can observe the errors made by general classes of ASTI configurations.

Table 2 lists mechanisms used for 2x2 matrices (found in Sets A and B of the SPM) and 3x3 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 correct.

Table 2: Mechanisms for Ablation in the ASTI Model

Type	Image sets	Transforms
2x2 matrices	1. Rows 2. Columns	1. Identity 2. Rotation/reflection 3. Addition/subtraction
3x3 matrices	1. Rows 2. Columns 3. Diagonals	1. Identity 2. Rotation/reflection 3. Addition/subtraction 4. Composition

Analysis of Error Patterns

Using the new classification of error types on the SPM that we described above, we conducted an analysis to compare the error patterns of typically developing individuals, individuals diagnosed with autism, and the ASTI model.

Human 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 (Lord et al., 1999; Rutter et al, 2003).

Using a cutoff of 17 years, participants were grouped into children and adults. Data included answer choices given for each SPM problem, including a few instances in which no answer was given. (One participant in the autism group was excluded from analyses, as he had selected answer choice "1" for more than half of the problems.)

Table 2 summarizes total SPM score, age, and Wechsler full-scale IQ information for these groups. While total SPM scores between TD and AUT groups are not significantly different, the ASTI SPM scores are significantly lower. This introduces a potential confound, if error types are dependent on overall ability. To address this issue, we conducted an analysis using three subgroups (TD children, AUT children, and the ASTI model) individually matched on total SPM score. Table 3 gives data on these subgroups.

We looked at the proportions of each error type that were made on the entire SPM test, averaged across participants in each group. Fig. 3 presents the results of these comparisons for the score-matched subgroups. Results for the full groups of children and adults were similar, and so we present detailed results of this first analysis only.

There is no significant difference in overall error distributions between the TD and AUT groups, $\chi^2(N = 826)$

= 1.89, p = 0.60, whereas the error distribution from the ASTI model differs significantly from each of the human groups, $\chi^2(N = 826) = 91.62$, $p < 0.001$ for TD, and $\chi^2(N = 826) = 98.69$, $p < 0.001$ for AUT.

A one-way ANOVA was used to test for differences in error proportions among the three groups. Proportions differed significantly for repetition, $F\ (2, 111) = 6.20$, p = 0.003, and difference errors, $F = 32.03$, $p < 0.001$, but did not differ significantly for incomplete correlate, $F = 0.14$, p = 0.87, or wrong principle errors $F = 1.61$, $p = 0.20$.

Table 2: Demographic data for full participant groups. Values as shown as: mean (standard deviation).

	Children		Adults		Model
	TD	AUT	TD	AUT	ASTI
N	54	108	52	44	96
SPM score	42.61 (9.79)	37.43 (12.17)	50.69 (5.38)	48.43 (9.64)	32.57 (9.74)
Age in years	11.96 (3.40)	11.02 (2.99)	22.98 (4.28)	26.80 (6.72)	n/a
IQ	109.82 (10.35)	84.38 (20.03)	106.91 (11.76)	97.61 (16.40)	n/a

Note: IQ data was not available for all participants.

Table 3: Demographic data for score-matched subgroups.

	Children		Model
	TD	AUT	ASTI
N	38	38	38
SPM score	38.26 (8.07)	38.26 (8.09)	38.29 (8.07)
Age in years	11.11 (3.30)	10.76 (2.71)	n/a
IQ	106.08 (9.08)	88.83 (18.79)	n/a

Note: IQ data was not available for all participants.

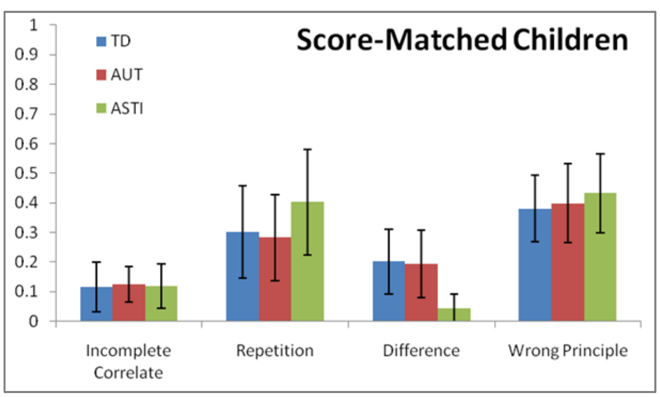

Figure 3: Proportions of each error type made on the SPM by typically developing (TD) individuals, individuals diagnosed with autism (AUT), and the ASTI model. (Error bars represent one standard deviation.)

Discussion

We discuss results from two perspectives. First, what does this analysis tell us about the error patterns shown by the TD versus AUT groups? Second, what does this analysis tell us about the error patterns shown by the ASTI model?

First, we see that the distribution of conceptual errors made on the SPM does not seem to differ significantly between the TD and AUT groups. Following a prior study suggesting that individuals with autism tend to use visual strategies to solve these kinds of problems (Soulières et al., 2009), one interpretation may be that looking at error types of this kind does not by itself indicate potential differences in problem solving modality (i.e. visual/verbal). However, as TD individuals most likely use a combination of visual as well as verbal strategies on the SPM, another, currently unexplored, hypothesis is that differences in error types may only surface for problems solved verbally by the TD group and visually by the autism group. If this is the case, then detecting such differences would require a finer-grained analysis of error types on various subsets of SPM problems instead of across the entire test as a whole.

To address the latter question, comparisons of errors between human participants and the ASTI model show agreement on two types of errors (incomplete correlate and wrong principle) and discrepancies on the other two types (repetition and difference). Looking at these differences in error patterns lends valuable insight into how specific aspects of the ASTI model affect its overall behavior and simultaneously suggests concrete avenues for improving the cognitive fidelity of the ASTI model.

First, with regard to the relative increase in repetition errors, the ASTI model predicts answers based on the matrix entries adjacent to the blank space. Thus, it is likely that its prediction is visually similar to an adjacent matrix entry, leading to an error of repetition. While humans do often make repetition errors, they also likely draw upon more aspects of the matrix when selecting an answer, which the ASTI model could also be modified to do.

Second, regarding the relative scarcity of difference errors made by the ASTI model, recall that these errors are made according to how a particular distracter might seem different or more complex than the other answer choices. Making difference errors thus should only affect test-takers using a response elimination strategy, i.e. looking at the answer choices as a set at the start of or during problem-solving. Test-takers using a constructive matching strategy already have an answer in mind before moving to inspect the answer choices, and if this answer is constructed by examining and combining matrix entries, it would likely be similar to these entries and thus not be likely to lead to a difference error.

Difference errors may thus be considered a result of test-takers fixating on the visual salience of one particular answer choice over another. The ASTI model currently does not contain mechanisms to detect salience or perform response elimination; the addition of these mechanisms will improve the fidelity with which problem-solving strategies used by the ASTI model mirror those of humans.

Conclusion

The main motivation for this work stems from the view that conceptual types of errors made on the Raven's 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-kind approach that yielded ≥95% inter-rater reliability. This classification should have considerable utility for further studies of human or machine SPM performance, and it adds a significant new component of information for the RPM family of tests, as both the CPM and APM tests already have such error classifications, but the SPM previously did not. One area of future work is to examine the error patterns made by humans on different subsets of test problems, instead of across the test as a whole, to achieve a finer-grained analysis of what kinds of errors people make on certain problems.

The second major contribution is the methodology presented for measuring the conceptual errors made by a computational model on the RPM. Looking at the errors made by the ASTI model has led us to propose two modifications to improve its cognitive fidelity: first, the model should consider additional aspects of the matrix when generating answer predictions, in addition to just the adjacent entries, and the model should be able to adopt a response elimination strategy and also be susceptible to the visual salience of particular answer choices.

Neither of these observations would have been possible by looking at total score alone, or even at the pattern of correct vs. incorrect answers. Future work on test-taking by humans and computational models should continue to look at multiple performance measures, beyond just total score, to fully understand performance and cognitive implications.

Acknowledgments

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.

References

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

Bromley, D. (1953). Primitive forms of response to the Matrices test. *Journal of Mental Sciences*, *99*, 374-393.

Carpenter, P., Just, M., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the RPM test. *Psychol. Rev*., *97*, 404-431.

Cohen, P., & Howe, A. (1988). How evaluation guides AI research. *AI Magazine, 9* (4), 35–43.

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

DeShon, R. P, Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence*, *21* (2), 135-155.

Facon, B., & Nuchadee, M. (2010). An item analysis of Raven's CPM among participants with Down syndrome. *Research in Developmental Disabilities*, *31*, 243-249.

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

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

Horner, J. & Nailling, K. (1980). Raven's Colored Progressive Matrices: Interpreting results through analysis of problem-type and error-type. *Clinical Aphasiology Conference*, Minneapolis, MN. 226-239.

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.

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

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *The ADOS-G*. Los Angeles, CA: Western Psychol. Services.

Lovett, A., & Forbus, K. (2012). Modeling multiple strategies for solving geometric analogy problems. *Proc. 34th Annual Conference of the Cognitive Science Society*.

McGreggor, K., & Goel, A. K. (2012). Fractal analogies for general intelligence. *Artificial General Intelligence, LNAI 7716*, 177-188.

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

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

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

Soulières, I., Dawson, M., Samson, F., Barbeau, E., Sahyoun, C., et al. (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 (3), 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 CPM in Williams Syndrome. *Research in Developmental Disabilities*, *32*, 93-99.

Vodegel-Matzen, L., van der Molen, M., & Dudink, A. (1994). Error analysis of Raven test performance. *Personality and Individual Differences*, *16* (3), 433-445.