The Journal of Experimental Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vjxe20

Measuring Relational Reasoning

Patricia A. Alexander[a], Denis Dumas[a], Emily M. Grossnickle[a], Alexandra List[a] & Carla M. Firetto[b]
[a] University of Maryland, College Park
[b] The Pennsylvania State University
Published online: 30 Jan 2015.

To cite this article: Patricia A. Alexander, Denis Dumas, Emily M. Grossnickle, Alexandra List & Carla M. Firetto (2015): Measuring Relational Reasoning, The Journal of Experimental Education, DOI: 10.1080/00220973.2014.963216

To link to this article: http://dx.doi.org/10.1080/00220973.2014.963216

THE JOURNAL OF EXPERIMENTAL EDUCATION, *00*(0), 1–33, 2015

ISSN: 0022-0973 print/1940-0683 online
DOI: 10.1080/00220973.2014.963216

Routledge
Taylor & Francis Group

LEARNING, INSTRUCTION, AND COGNITION

Measuring Relational Reasoning

Patricia A. Alexander, Denis Dumas, Emily M. Grossnickle, and Alexandra List
University of Maryland, College Park

Carla M. Firetto
The Pennsylvania State University

Relational reasoning is the foundational cognitive ability to discern meaningful patterns within an informational stream, but its reliable and valid measurement remains problematic. In this investigation, the measurement of relational reasoning unfolded in three stages. Stage 1 entailed the establishment of a research-based conceptualization of the construct and the development of a corresponding Test of Relational Reasoning (TORR). Stage 2 focused on the reliability and validity of data from the TORR. Analyses showed the data from the TORR to be reliable indicators of students' ability to reason relationally, and TORR performance predicted students' performance on SAT verbal and math problems. Stage 3 examined the underlying structure of the construct through Confirmatory Factor Analysis (CFA). Of the three CFA models tested, models with dedicated factors for analogical, anomalous, antinomous, and antithetical reasoning were deemed the best fit for the data.

RELATIONAL REASONING, which can be broadly conceptualized as the ability to recognize or derive meaningful relations between and among pieces of information that would otherwise appear unrelated (Alexander & the Disciplined Reading and Learning Research Laboratory, 2012) be that information linguistic, graphic, or numeric in nature (Bulloch & Opfer 2009; Crone et al., 2009; Holyoak, 2012), stands as one of the most foundational of human cognitive abilities. When individual words cohere to communicate an idea, when isolated images coalesce into a recognizable composition, or when separate data points come to signify a reportable trend or result, relational reasoning is involved. According to William James (1890), we would be imprisoned in a world of isolated stimuli if not for the ability to perceive relevant relations among the objects of our perception, even when they are separated by time and space. In an age where individuals live their lives awash in information, this ability to perceive and attend to patterns

Address correspondence to Patricia A. Alexander, University of Maryland, College Park, Department of Human Development and Quantitative Methodology, 3304 Benjamin Building, College Park, MD 20742, USA. E-mail: palexand@umd.edu

within the deluge of data is a particularly timely concern (Braasch, Bråten, Strømsø, Anmarkrud, & Ferguson, 2013).

Since the earliest research on perception (e.g., Wertheimer, 1900), ample evidence has been amassed suggesting that humans display an innate tendency to seek order within or to impose some structure on otherwise unrelated information (e.g., Chase & Simon, 1973; Hofstadter, 2001; Köhler, 1951; Spearman, 1927; Sternberg, 1977). The mechanisms broadly considered within those early and more contemporary literatures pertain to the concurrent identification of the similarities and dissimilarities within the informational flow. As James (1890) argued: "It is probable, also, that man's *superior association by similarity* [author's emphasis] has much to do with those discriminations of character on which his higher flights of reasoning are based" (p. 345).

Yet, acknowledging the foundational role played by relational reasoning in human thought does not afford sufficient evidence about its inherent nature. Potentially important questions remain concerning the characterization of the construct itself. For example, is relational reasoning a unitary construct, or is it multidimensional with distinct forms that merit individual attention in research and in practice? Do the distinctions between similarity- versus discrimination-mechanisms suggested by James (1890) remain evident in the process of relational reasoning or do they become indistinguishable when that process unfolds? Furthermore, it is unclear from the available empirical evidence whether relational reasoning ability can be accounted for by other associated cognitive factors such as prior knowledge or working memory (Cromley, Snyder-Hogan, & Luciw-Dubas, 2010; Krawczyk, 2012; Newcombe et al., 2009). In essence, are the patterns recognized within a body of information largely predicated on retrieval of similar phenomena from long-term memory (Goswami, 1992)? Or, are they a consequence of abilities to perceive, retain, and manipulate data (Baddeley, 2000, 2012)?

Moreover, as has been documented with other fundamental cognitive abilities pertaining to thinking or reasoning (Epstein, Pacini, Denes-Raj, & Heier, 1996; Flavell, 1979), the existence of an innate tendency to discern pattern does not apparently ensure that individuals are particularly competent at relational reasoning or capable of harnessing this foundational ability to purposefully or effortfully deepen or expand understanding (e.g., Alexander, Pate, Kulikowich, Farrell, & Wright, 1989; Dinsmore, Doyle, Baggetta, & Loughlin, 2012; Gick & Holyoak, 1980). In effect, reasoning relationally seemingly requires more than pedestrian perception or attention when the situation that is confronted demands more than a habituated response (e.g., Chinn & Brewer, 1993; Dunbar, 2001). To capture this distinction, some in the literature have sought to distinguish between that mode of patterning that is unconscious and routine and that which demands greater awareness and focused effort (e.g., Stanovich, 2009). The former has been termed *relational thinking*, whereas the latter would be classified as *relational reasoning* (Alexander & Baggetta, 2013). It is that more conscious and effortful form of human patterning that we sought to investigate in this study.

What also remains to be established empirically is the degree to which the construct of relational reasoning differentially affects students' academic outcomes. For example, would we expect relational reasoning ability to predict undergraduate students' performance on a demanding task considered indicative of academic potential, such as the SAT? Moreover, would differing profiles of relational reasoning ability emerge in undergraduate students in natural sciences courses as compared with those enrolled in the social science courses? These are among the questions we aimed to explore in this investigation—an exploration that was predicated on the reliable and valid measurement of relational reasoning.

Specifically, our study of relational reasoning unfolded in three stages. The purpose of Stage 1 was to determine whether there was a literature-based conception of relational reasoning that could guide our investigation and what, if any, manifestations of relational reasoning should be empirically tested. Also, in this initial stage we sought to ascertain what particular configuration of items and scales should form the basis for the resulting measure and to test those decisions through a series of pilot studies. With those conceptual and operational questions addressed, we then turned to a systematic analysis of the reliability and validity of scores from the relational reasoning measure. In this second stage, we investigated not only the stability of scores across alternate forms (paper and online) of the measure but also examined test–retest reliability. As with reliability, we tested the validity of resulting scores in several ways, including tests of convergent, discriminant, and predictive validity. Stage 3 of this study was intended to answer the overarching question as to the nature of relational reasoning and whether this construct was best conceived as unidimensional or multidimensional. Toward that end, three confirmatory factor analyses were conducted, each representing an alternative, and theoretically defensible, representation of the nature of relational reasoning.

STAGE 1: CONCEPTUALIZATION AND OPERATIONALIZATION OF RELATIONAL REASONING CONCEPTUAL FRAMEWORK

The conceptualization of relational reasoning began with an intensive and extensive perusal of relevant literatures from philosophy, psychology, neuroscience, and set theory that could shed light on the nature and form of relational reasoning. Initial theoretical work on this topic led to some tentative judgments about the nature and process of relational reasoning (e.g., Alexander & Baggetta, 2013; Alexander and the Disciplined Reading and Learning Research Laboratory, 2012) that were then systematically investigated through the empirical literature. The culmination of that yearlong endeavor was a recently published cross-disciplinary review of relevant empirical literatures (Dumas, Alexander, & Grossnickle, 2013). In seeking to build a psychometrically sound measure that would permit us to empirically test our hypotheses about the nature and form of relational reasoning, we drew on that literature review to: (a) establish the conceptual basis for the construct of relational reasoning that we set out to measure; (b) identify particular forms of relational reasoning that have populated the literature and that should therefore be considered; and (c) examine approaches taken to its assessment that can inform the explicit configuration of the measure for this investigation.

For one, Dumas and colleagues (2013) determined that the term *relational reasoning*, while more often implicitly defined within the psychology, neuroscience, and developmental literatures, generally and rather consistently dealt with the discernment of underlying structures between multiple stimuli that on the surface bear little similarity. Furthermore, the particular forms of relational reasoning presented in this cross-disciplinary review were defined by the nature of the relations forged. Specifically, when the association formed between mental representations was based on underlying similarities, the mode of relational reasoning was *analogical* (Bassok, Dunbar, & Holyoak, 2012; Nersessian & Chandrasekharan, 2009). Interestingly, Dumas et al. (2013) found that even when the general construct of relational reasoning was the stated focus of investigation, it was analogical relations that were most often empirically examined. This emphasis on analogical relations—long established as pivotal in human learning and development

(Alexander & Murphy, 1999; Goswami, Leevers, Pressley, & Wheelwright, 2011; Sternberg, 1977)—may complicate efforts to examine the construct of relational reasoning more holistically. This tendency within the psychological literatures seems consistent with James's (1890) argument that critical discrimination judgments foundational to reasoning may unfortunately be neglected in theory and research in favor of relations of similarity.

In this investigation, therefore, we set out to systematically investigate not only relations of similarity, but also those of contrast and even of paradox. Our initial conceptualizations in these areas were informed by relevant philosophical (Hofstadter, 2001; James, 1890) and psychological literatures (Köhler, 1951; Spearman, 1927; Sternberg, 1977), as well as writings in mathematics set theory (Russell, 1908; Russell & Lackey, 1973). From the outset, we recognized that there could be innumerable relations that might be forwarded. However, given that our intention was to explore whether multiple forms of relational reasoning merited consideration, over and above the typical analogical association, we elected to begin with a parsimonious list that was justified by the empirical literature and that also captured various forms of discriminating relations. Specifically, in addition to the more commonly studied similarity relation (analogy), we also focused on discriminations based on perceived discrepancies (anomaly), incompatibilities (antinomies) and dichotomies (antithesis).

For example, when the underlying relations concerned discrepant events or aberrant cases within a given set, the form of reasoning was termed *anomalous* (Chinn & Brewer, 1993). The ability to recognize discrepancies or irregularities has been linked to reading comprehension (Bohan & Sanford, 2008; Ivanova, Pickering, Branigan, McLean, & Costa, 2012) and to academic achievement within scientific domains (Chinn & Malhotra, 2002; Chinn & Samarapunga, 2009). Furthermore, the identification and treatment of anomalies in data has been shown to be a central ability for the professional scientist that continues to develop with expertise (Trickett, Trafton, & Schunn, 2009).

In contrast, other forms of relational reasoning that Dumas and colleagues (2013) described were based on conflicts between objects or phenomena. In one case, that conflict is paradoxical in nature (Russell & Lackey, 1973). That is, the ideas or concepts being related appear incompatible as when scientific rules, representations, or categories are determined to be ontologically distinct (Chi & Roscoe, 2002; Opfer & Gelman, 2011; Slotta & Chi, 2006). This particular form of relational reasoning, termed *antinomous*, requires the determination of rule-based or principle-based criteria that allow the thinker to describe what something is by ascertaining what it is not. It is important to note that in this conceptualization of relational reasoning, incompatibilities and contrasts, such as those required for antinomies, are considered relations to be mapped during the reasoning process. For example, as Cole and Wertsch (1996) pointed out, an antinomy can arise when concurrently considering the theoretical arguments of Piaget and Vygotsky. Specifically, the idea that human development can be simultaneously driven by the individual child and by social processes may involve an antinomy-based paradox. Antinomous reasoning has also been fruitfully used in theoretical discussions in a variety of fields of inquiry including reading (Mosenthal, 1988), psycholinguistics (Shaumyan, 2006), and intelligence (Gardner, 1995).

In another case, the conflict being conveyed is represented by opposing or polar positions (e.g., agree/disagree, in/out, or true/false) and termed *antithetical reasoning*. In effect, antithetical reasoning requires the analysis of contrasting positions of the same trait, characteristic, or contention (Bianchi, Savardi, & Burro, 2011; Kreezer & Dallenbach, 1929). Although antithetical reasoning has long been associated with the semantic organization of language (de Saussure, 1916/2011;

Kjeldergaard & Higa, 1962; Marková, 1987), it has more recently been identified within the literatures on persuasion or argumentation (Andiliou, Ramsay, Murphy, & Fast, 2012; Nussbaum & Kardash, 2005). For example, when opposing stances or arguments are juxtaposed and readers or listeners are called upon to weigh these contrasting positions to comprehend what is read or heard (Murphy, 2001). As such, antithetical thinking has also been linked to conceptual change in scientific domains through the use of refutation text (Broughton, Sinatra, & Reynolds, 2010).

On the basis of the general and particularized conceptions of relational reasoning that populate the cross-disciplinary literature, we chose to create a measure of relational reasoning addressing the aforementioned forms of this construct (i.e., analogy, anomaly, antinomy, and antithesis). In so doing, our goal was to discern whether the forms of relational reasoning would prove empirically distinct or whether relational reasoning was better understood as a unitary cognitive ability that simply manifests in diverse ways. It is important to acknowledge that we do not hold that there are only four manifestations of relational reasoning that exist or that could be the focus of measure development. To the contrary, we presume that there are innumerable types of associations that are conceivable whenever we seek to relate objects or phenomena. However, we determined that these four forms had been substantiated in the empirical literature and afforded us adequate diversity to address the question of unidimensionality or multidimensionality of the underlying construct. We hereafter refer to the resulting measure formed around these varied manifestations of relational reasoning as the Test of Relational Reasoning (TORR).

Test Conceptualization

Before beginning the development of any of the specific items on the TORR, we made several critical decisions about the structure and form of the measure. Those decisions pertained to the targeted age range, representational format (e.g., linguistic or graphic), suitable length, and delivery method (i.e., on paper or online).

Age Range

With regard to the desired age range, we determined that this initial measure should target older adolescents through adults. There were theoretical and practical rationales for this decision. Theoretically, there were developmental concerns that needed to be considered. For example, prior research has suggested that it is not until this stage of development that most individuals reach their full ability to reason relationally (e.g., Dumonthiel, Houlton, Christoff, & Blakemore, 2010). Although there have been studies where relational reasoning or its particular manifestations has been identified within young populations (Baker, Friedman, & Leslie, 2010; A. L. Brown, Kane, & Long, 1989; Goswami, 1992; Savage, Dealt, Daki, & Aouad, 2011; Stevenson, Touw, & Resing, 2011; White & Caropresco, 1989), those studies entailed a level of scaffolding and support that we did not want to introduce into the current measure. Furthermore, because we were interested in questions dealing with delivery system and academic foci, we wanted to ensure that participants were not only able to function independently within online environments, but would also be actively pursuing varied fields of study.

From a pragmatic standpoint, there were also age-related issues to be considered. For one, the adequate assessment of four forms of relational reasoning would likely require a commitment of time and cognitive effort that could not be assumed for younger populations. In addition,

the need to conduct intensive cognitive labs and pilot studies would be aided by the availability of large samples of undergraduate and graduate students from various disciplines and majors. Moreover, we determined that once the construct of relational reasoning could be investigated in older adolescents and adults using a psychometrically sound measure, that measure could then serve as the model for alternative forms more suitable for younger adolescents and children.

Representational Form

As noted, one of our goals was to attempt to measure relational reasoning while controlling for potentially confounding factors such as prior knowledge and language background. To deal directly with the issue of prior knowledge and to focus directly on fluid rather than more crystallized cognitive abilities (Ackerman, 1988; Snow, 1989), we elected to construct a measure composed entirely of figural problem sets. Not only were items entirely figural, but they were also designed so that all the necessary information to solve the given problem was contained within the problem space (outside of the vocabulary in the brief directions). This design method has been used by the producers of cognitive assessments for three-quarters of a century (Cattell, 1940; Raven, 1941), and is generally accepted today as a valid method for limiting the influence of prior knowledge and culturally relevant experiences on a given evaluation or test (Krigbaum, Amin, Virden, Baca, & Uribe, 2012).

Measure Structure

At present, no measure exists that is designed to examine the potential multidimensionality of relational reasoning (Dumas et al., 2013). In response to this identified gap in the literature, the TORR was conceptualized as having four scales, each corresponding to one of the identified forms of relational reasoning (i.e., analogy, anomaly, antinomy, and antithesis). Furthermore, scales were originally presented in a fixed order with the more familiar manifestation of relational reasoning, analogy, appearing first followed by anomaly, antinomy, and then antithesis.

Suitable Measure Length

Even with the age range, representational form, and structure decided, we confronted the issue of suitable length. We wanted adequate assessment of each form of relational reasoning to allow us to assess the unidimensionality or multidimensionality of the construct and to explore cross-scale performance patterns. However, we appreciated that fatigue could well affect the outcomes we received. Therefore, although the TORR was conceived as an untimed measure, in which participants would be permitted to take whatever time they required, we wanted to create a measure that could generally be completed within 45 min; thus avoiding undue cognitive fatigue and other motivational concerns. For this reason, we initially aimed to generate up to 10 items per scale and to pilot these scales to determine the overall completion time. On the basis of those initial pilots, we later determined that a slightly reduced count of eight items per scale fit our needs for adequate sampling without proving too taxing for participants. In addition to the resulting 32 scored test items, there were two sample items included per scale. These sample items were used to familiarize participants with the directions for each of the scales.

Delivery Method

Two general forms of the TORR, further explicated in the subsequent section, were developed: a paper and an online version. The existence of the two forms allowed us to subsequently test whether the mode of delivery resulted in differences in the reasoning performance of participants.

Test Development

The process of measure development for the TORR had three phases: item creation, cognitive labs, and pilot studies. Each of these stages will now be further explicated.

Item Creation

Each of the items on the TORR was originally created using a collaborative brainstorming process. Throughout the item creation stage, the correspondence between the items and the theoretical description of the forms of relational reasoning they were being designed to tap was especially important. Each item was designed so that a correct answer would demonstrate a participant's ability to use the corresponding relational reasoning process (e.g., similarity or discrepancy) and was structured to ensure that those underlying processes would differ across scales, as we detail in the subsequent scale descriptions. Furthermore, to avoid the need for participants to inhibit reasoning processes used on a previous scale, the patterns used in any one of the scales were not repeated in the other scales. What follows is a more detailed description of each scale with a sample item to illustrate the corresponding reasoning entailed.

Analogy

The items on the analogy scale of the TORR are designed using a matrix format. This format has long been used to design nonverbal tests of analogical reasoning (e.g., Krawczyk, McClelland, & Donovan, 2011). John Raven (1941) is generally credited with the creation of the first matrix analogy problems. Since Raven, other researchers have created instruments using the matrix analogy format, for clinical (Naglieri & Insko, 1986) and empirical (Krawczyk et al., 2011) purposes. Because of its familiarity and nonverbal nature, the matrix format was seen as the most potentially useful configuration for the items on the analogy scale.

Analogical reasoning is generally defined by the presence of a fundamental relation of similarity (e.g., Holyoak, 1985; Novick, 1988). The analogy items on the TORR consisted of a three by three matrix of figures, with the figure in the lower right unspecified. Participants are asked to discern the pattern underlying the given problem, and then decide which of the possible answer choices would complete that pattern. This underlying pattern could be identified by processing the problem components either vertically or horizontally. In effect, to solve matrix analogies such as those on the TORR, a participant must identify the answer choice that makes one row or one column of elements relationally similar to another row or column of elements. It is this basic requirement to establish the similarity across elements of the problem to reach the solution that defines these items as analogical in nature (Carpenter, Just, & Shell, 1990).

The directions and a sample item from the analogy scale are provided herein. It is important to note that these sample items are used to familiarize participants with the structure of the test,

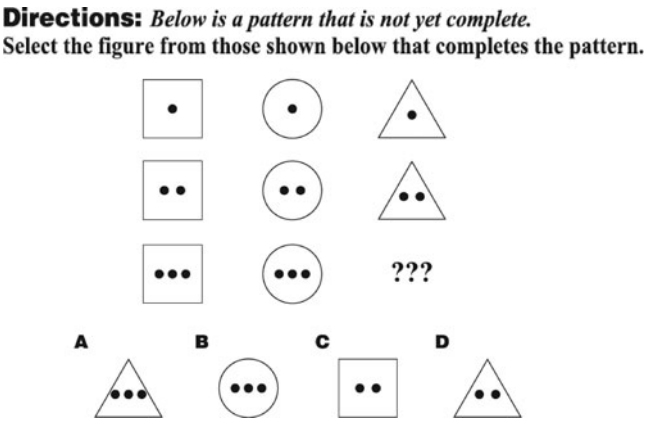

FIGURE 1 A sample analogy item from the Test of Relational Reasoning.

and as such are intended to be relatively easy compared with most of the items within the scale. In this sample item depicted in Figure 1, the correct answer is A, because that answer choice completes the pattern of changing shape (square, circle, triangle) across the rows and increasing number (1,2,3) of dots down the columns of the matrix.

Anomaly

Unlike the analogy scale, which was designed in the widely used matrix format, there were no graphical models for the measure of anomaly currently used by the field. So, a format for the items on this scale of the TORR was devised based upon the theoretical nature of an anomaly. Because an anomaly is defined as an unexpected deviation from a pattern or rule (Chinn & Brewer, 1993; Klahr & Dunbar, 1988), the items on the anomaly scale necessarily had to depict a discernable pattern that could then be broken by the anomaly itself. So, an odd-one-out format was used. Four figures were presented in a nonlinear array: three of the figures follow the same pattern, but one of them—despite being visually similar to the other three—does not. The respondent must attend to the features of each of four figures, recognize the pattern governing the array, and then select the figure that deviates from the pattern. A sample item used to familiarize participants with this scale of the TORR is shown in Figure 2. In this sample item, the correct answer choice is D because it is the only figure in the array that does not follow the pattern of having one fewer horizontal line than vertical lines. Although Figures A, B, and C each follow this pattern, D breaks the rule, and thus can be identified as the anomaly.

Antinomy

An antinomy can arise when two or more ideas appear incompatible (Sorensen, 2003). Antinomies have been described as occurring when mutually exclusive sets or categories are brought together (Chi & Slotta, 1993; Russell & Lackey, 1973). For example, Chi and Roscoe (2002)

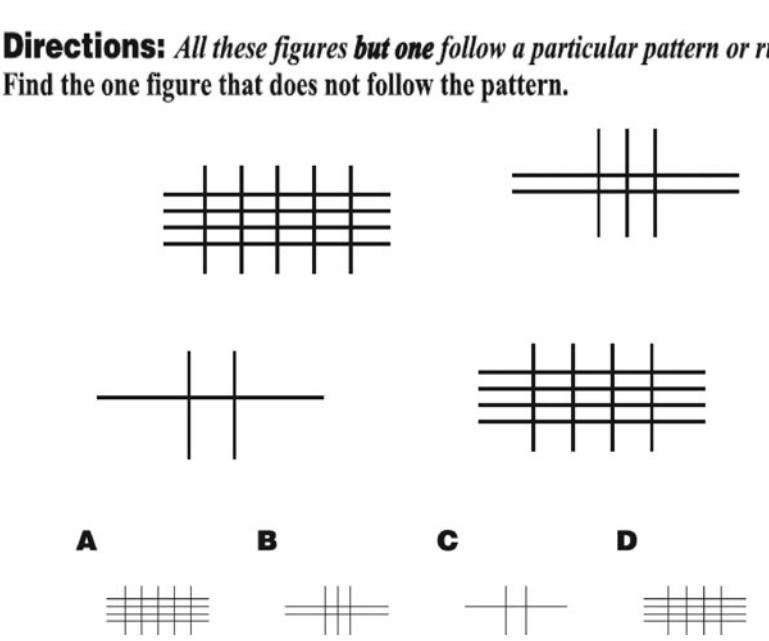

FIGURE 2 A sample anomaly item from the Test of Relational Reasoning.

described the necessity to reason with incompatible ontological categories during the process of conceptual change. To reflect this theoretical definition of antinomy, the items on the antinomy scale of the TORR were designed with a given set, governed by a rule for inclusion, and four answer choice sets, also governed by their own rules for inclusion. It is the respondent's task to decide which of the answer choice sets has a rule for inclusion that is antinomous, or incompatible, with the rule for inclusion in the given set. Simply stated, the correct answer choice is the set that could never have an item in common, or is mutually exclusive, with the given set. A sample item from the antinomy scale of the TORR, along with its directions, is displayed in Figure 3.

For this sample problem, the rule governing the given set allows differing shapes to be included, as long as they are of the same, designated color (i.e., gray). Options A, B, and C comprised one shape each (hexagon, circles, and diamonds, respectively), but those shapes are of varying fills. In this sample item, each of these options includes a gray shape corresponding to one in the given set. Thus, Options A, B, and C, while different from the given set, can have a member in common with the given and are thus not incompatible. Only Option D has a rule for inclusion that is incompatible with the given set. Because Option D can only include dotted shapes, it could never have a member in common with the given set, marking it as the correct choice.

Just as the analogy items were defined by the presence of a fundamental similarity association, the antinomous items were defined by a relational incompatibility. Thus, to identify the correct answer, the respondent needed to ascertain which of the options provided had *no* members potentially in common with the given set. In effect, membership in the selected set precluded membership in the given set.

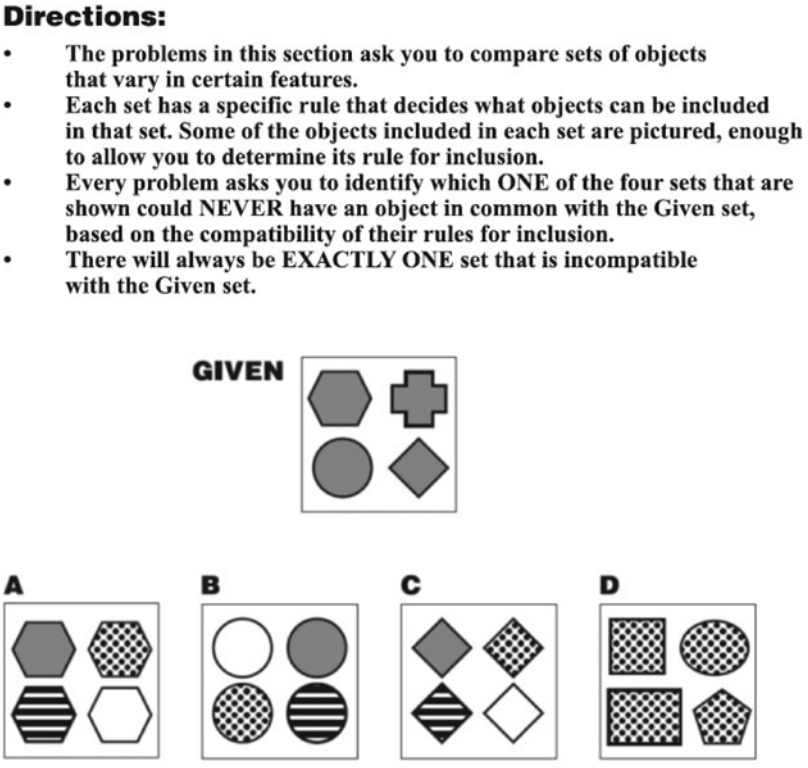

FIGURE 3 A sample antinomy item from the Test of Relational Reasoning.

Antithesis

An antithesis is theoretically described as directly oppositional or polar concepts (e.g., Bianchi, Savardi, & Kubovy, 2011). In the academic context, counterarguments, polar stances, or reversals of rules or processes—often associated with conceptual change—are instances of antithetical thought (Sinatra & Broughton, 2011). Therefore, the items on the antithesis scale of the TORR were designed so that the correct choice would have a relation of true opposition to the given figural array. To achieve this, the given array was created to depict a process where one figure (labeled "X") is changed into another figure (labeled "Y"). To correctly solve the item, a participant must decide what process has taken place in the given, and then select the option that depicts the opposite of that given process. A sample problem and its directions from the antithesis scale of the TORR is presented in Figure 4.

In this sample item, the process being depicted in the given array is a doubling of the number of squares and a changing of the color of the squares from white to black. Answer choice C depicts the antithesis of the given process because, in process C, the number of squares is being halved and the color is being changed from black to white. It is the fundamental relation of opposition that defines the items in this scale as antithetical. To select the correct answer to an antithesis

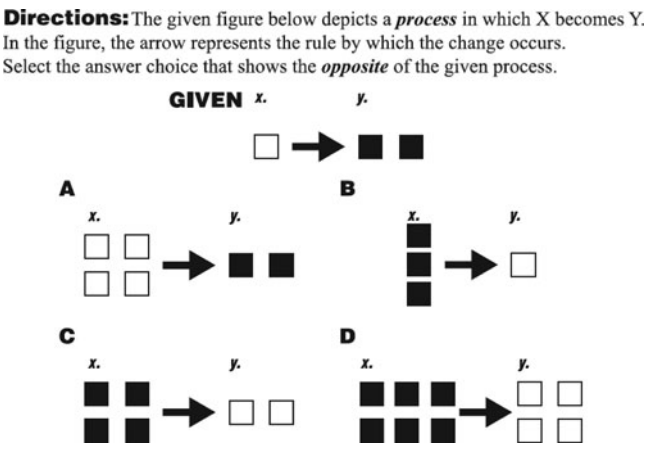

FIGURE 4 A sample antithesis item from the Test of Relational Reasoning.

item, a respondent had to reverse the process depicted in the given figure. This procedure closely parallels the process of refutation that has been implicated in conceptual change and persuasion literatures (Broughton et al., 2010; Murphy & Mason, 2006).

Cognitive Labs

After the initial item creation process, cognitive labs were used to gather feedback about the TORR items. The cognitive labs were in-depth interviews involving a small select sample ($N = 5$) of skilled thinkers that included advanced graduate students in science education ($n = 2$), mathematics education ($n = 2$), and a professor in human development ($n = 1$) from a university in the Mid-Atlantic region of the United States. These individuals were chosen because of their demonstrated capacity to think and reason with facility, as well as their ability to explain their reasoning processes.

Each cognitive lab participant first completed the TORR silently. Then, they verbally explained to a researcher how they had come to select their answer choice for each item. Where a participant's answer choice differed from what the researcher had expected, the researcher and participant discussed the discrepancy. Furthermore, any difficulties that the participant experienced in comprehending the scale directions, or understanding the format of the items or the test as a whole was noted and discussed. Cognitive lab participants responded to a series of questions concerning their perceptions of the TORR and its scales. For example, participants were asked to generate a possible title for each of the scales as well as for the test as a whole. Participants were also asked questions such as "If you were to give this test to a student, what do you think it would tell you about them?" and "How similar or different do you think these scales are?"

The resulting data were used to identify aspects of the draft TORR items that did not function in the way they were intended to function, ultimately leading to item revision. For example, in some cases, participants would argue that two of the available answer choices for an item were potentially correct. In these situations, the answer choices were revised to eliminate any ambiguity. In certain instances, specific items in the initial pool were deleted or replaced. Item

revision associated with the cognitive labs continued, and the pilot study phase of the investigation was not initiated, until there was agreement among those chosen as more competent reasoners that only one of the answer choices on the TORR was logically keyable.

Moreover, participants' answers to the questions posed were reviewed to determine whether the TORR and its scales appeared to measure the constructs of interest. Overall, participants described the TORR as comprising four distinct but related scales, each requiring participants to reason differently about problem elements and their associations. For example, one participant gave the title "test of complex patterns" to the TORR as a whole, while another named it the "fluid reasoning test." Labeling of the analogy scale generally centered on the construct of similarity and included such titles as "find the match" and "complete the pattern." For the anomaly scale, titles such as "one of these things is not like the other" and "odd-one-out" captured the discrepancy relation that was core to its conceptualization. Titles such as "mutually exclusive sets" and "shapes that do not fit" drew on the relation of incompatibility for the antinomy scale. Labels such as "opposites" and "reverse the rule" were offered for the antithesis scale.

Furthermore, cognitive lab participants described the scales of the TORR as being relatively discrete but complementary. One participant explained that they were like "different pieces of the puzzle" of the entire test, and went on to point out that a single scale could give information about "a student's ability to reason with a particular pattern," but the entire test would "see if they were able to reason with a variety of patterns." Only when we were satisfied that the scales and the items were functioning well for our select sample of respondents did we move forward with several pilot studies.

Pilot Studies

Following item refinements suggested by the cognitive labs, three pilot studies were conducted to (a) gather data about the item functioning and the corresponding item difficulty levels, (b) test the functionality of the online delivery system, and (c) gather information on the average time required to complete the measure. Each pilot study was conducted with undergraduate students enrolled in various human development courses at a large Mid-Atlantic university. For each pilot, participants completed the TORR as a take-home extra credit assignment. All of these pilot studies conformed to the requirements for human subjects including voluntary participation and confidentiality. Demographic data on pilot participants are provided in Table 1.

Item Functioning and Difficulty Level

To test the quality of each item that comprised the TORR, we examined data from each pilot with an eye toward the functioning of response options, including the item distractors and the item difficulty level. Specifically, we examined the response frequencies for the distractors for each item on the TORR to determine their plausibility and to ensure that no distractors were overly appealing or potentially keyable. Whenever distractors did not demonstrate plausibility, they were adjusted or replaced. In addition, we considered the overall item difficulty level (i.e., proportion of individuals responding correctly to that item). Although differences of opinion exist as to optimal item difficulty range (Allen & Yen, 1979; Downing, 2006), we decided to target difficulty levels between .30 and .70 for the TORR to increase variability for our target population of older adolescents and adults.

TABLE 1
Demographic Data on Pilot Study Participants

Pilot	*Number*	*Mean age (*SD*)*	*Gender*	*Demographics*
1	51	19.84 (1.20)	66.66% Female 33.34% Male	70.59% Caucasian 13.72% African American/Black 9.80% Asian 3.90% Hispanic
2	288	19.81 (1.88)	69.75% Female 31.25% Male	67.36% Caucasian 12.15% African American/Black 14.58% Asian 7.29% Hispanic
3	186	19.60 (2.81)	76.66% Female 23.34% Male	67.74% Caucasian 8.6% African American/Black 13.98% Asian 5.91% Hispanic

Each of the individual item difficulty levels was examined empirically during the pilot studies. When difficulty of a given item was outside of our target range (.3 to .7) that item was revised using a systematic procedure. Specifically, to make an item more difficult, the number of features or elements to which a respondent had to attend to arrive at a correct answer was increased. Conversely, to make an item less difficult, the number of features or elements to be considered was reduced. This method of altering item difficulty corresponds to the principles of relational complexity theory (e.g., Andrews, Birney, & Halford, 2006; Halford, Wilson, & Phillips, 1998), which posits that each salient feature or element of a problem (e.g., shape or size) adds to its relational complexity and, thus, its complexity.

After each pilot study, the item difficulties were examined and items that were deemed too easy or too difficult were revised or replaced. On the basis of the pilot studies, it was also determined that a second sample item should be added to each scale to afford increased familiarity with the intended problem-solving procedure. Following the completion of the pilot studies, the difficulties for the TORR items were judged acceptable with only 2 items falling somewhat outside the designated range (i.e., .28 and .81). The specific item difficulties for the final version of the TORR are displayed in Table 2.

Online System

The online version of the TORR was constructed and administered using the Qualtrics (2012) online service platform. During the pilot studies, the Qualtrics platform was determined to be a functional testing environment, with no malfunctions reported by any of our pilot participants. There were certain advantages associated with the online test. For example, participants were able to complete the TORR outside of class from any computer connected to the Internet, allowing a degree of flexibility. Furthermore, the online platform offered secure data management and was able to record time data easily. Following the pilot studies, it was determined that the scales of

TABLE 2
Difficulty Levels for Test of Relational Reasoning Items, by Scale

Scale	*Item*	*Difficulty*
Analogy	1	.49
	2	.32
	3	.65
	4	.31
	5	.67
	6	.40
	7	.57
	8	.28
Anomaly	1	.57
	2	.73
	3	.39
	4	.47
	5	.61
	6	.51
	7	.37
	8	.31
Antinomy	1	.76
	2	.64
	3	.44
	4	.61
	5	.55
	6	.44
	7	.36
	8	.53
Antithesis	1	.34
	2	.47
	3	.81
	4	.57
	5	.40
	6	.62
	7	.59
	8	.64

the TORR would be presented in random order within the online system, although the item order remained fixed within scales.

Timing

Although the TORR was conceived to be an untimed test, we endeavored to keep the completion time to 45 min or less to avoid undue fatigue. On the basis of the pilot data, this objective was met. Specifically, the mean amount of time needed for the pilot participants to complete the TORR was 24.71 min ($SD = 20.07$). As the magnitude of the standard deviation suggests, however, there was substantial variability in the time taken to task completion. This high level of variability in time was not unexpected for this novel and untimed test and provided substantiation that the total number of items per scale was suitable for the target population. However, this variability in the

time it took for participants to reason with the TORR items suggests that the examination of timing patterns may be a potential future direction for research on relational reasoning incorporating the TORR. At this point, revisions were complete, and the final form of the TORR, which appears as supplementary material in the journal's online archive, entered a new phase of development centered on ascertaining the reliability and validity of the measure.

STAGE 2: DETERMINING THE RELIABILITY AND VALIDITY OF TORR DATA TESTS OF RELIABILITY

To determine the consistency of scores on the instrument, three forms of reliability were investigated. First, the reliability of method was investigated by administering the TORR at two time points on paper as well as online. Cronbach's alpha was computed for the TORR at Time 1 and at Time 2. Furthermore, test–retest reliability was computed to determine stability of the measure over time.

Participants

Participants were 71 students at a large Mid-Atlantic university (46 female; 64.80%). Students were predominantly juniors ($n = 21$; 29.58%) and seniors ($n = 45$; 63.38%) enrolled in an upper-level human development course. Participants were offered extra credit for participation. These undergraduates ranged in age from 19 to 23 years old, with a mean age of 20.83 years old ($SD = 0.83$). The sample was majority White ($n = 39$; 54.93%), with 8.45% of students reporting their ethnicity as Black ($n = 6$); 15.49% students reporting their ethnicity as Hispanic/Latino ($n = 11$); and 14.08% reporting their ethnicity as Asian ($n = 10$). Also, 85.92% of the sample reported English as their first language ($n = 61$). The majority of participants ($n = 61$; 85.92%) represented majors in the social sciences/humanities, including psychology, kinesiology, and family sciences. In addition, 7.04% of students ($n = 5$) majored in the natural sciences. Participants reported a mean grade point average of 3.27 ($SD = 0.37$) on a four-point scale, with GPAs ranging from 2.40 to 4.00.

There was 19.72% attrition ($n = 14$) from Time 1 to Time 2. The sample at Time 2 consisted of 57 students (42 female; 73.68%). Those students who completed the TORR at Time 1 but not at Time 2 had a mean age of 21.2 years ($SD = 0.42$) and an average GPA of 3.39 ($SD = 0.49$). No systematic differences in participant demographics were found between students completing the TORR at Time 1 and Time 2 and participants completing the TORR at Time 1 only. Reliability analyses for Time 1 were conducted on the full sample ($n = 71$), while reliability analyses at Time 2 and test–retest reliability will be computed based participants with data for Time 1 and Time 2 ($n = 57$).

Stability Across Form

Participants were randomly assigned to complete the TORR either on paper or online. Students were asked to complete the TORR outside of class, at their convenience. Those in the paper condition were given a paper copy of the TORR, whereas those assigned to complete the TORR

electronically were given a recruitment letter with a link to the online version of the TORR. Otherwise, directions and items presented to participants were identical.

To check for a method effect a one-way analysis of variance was conducted to check for mean differences in performance on the paper versus electronic version of the TORR. There were no significant differences in performance between methods at Time 1, $F(1, 65) = 0.90$, $p = .35$, and Time 2, $F(1, 54) = 0.15$, $p = .70$. Furthermore, there were no significant differences found between methods of administration (i.e., on paper versus online) for the four relational reasoning scales. This was the case at Time 1, Fs$(1, 68) < 0.79$, ps $> .38$, and Time 2, Fs$(1, 51) < 0.54$, ps $> .47$. Thus, given these results, we determined that the mode of delivery was not a significant factor in the measurement of students' relational reasoning abilities for the TORR. As there were no significant differences found between methods of administration, scores for the on paper and online groups were collapsed for further reliability analyses.

Reliability at Time 1 and Time 2

At Time 1, the mean total score on the TORR was 16.37 ($SD = 6.45$). The mean scores on the individual scales at Time 1 were as follows: analogy ($M = 4.06$, $SD = 2.11$), anomaly ($M = 3.81$, $SD = 2.11$), antinomy ($M = 4.12$, $SD = 1.82$), and antithesis ($M = 4.35$, $SD = 2.04$). At Time 2, the mean total score on the TORR was 15.64 ($SD = 6.07$). The mean scores on the individual scales at Time 2 were as follows: analogy ($M = 3.53$, $SD = 2.13$), anomaly ($M = 3.85$, $SD = 2.12$), antinomy ($M = 4.21$, $SD = 1.91$), and antithesis ($M = 3.93$, $SD = 1.80$). Separate reliability coefficients were computed for the TORR at Time 1 and Time 2. At Time 1, Cronbach's alpha for the measure was $\alpha = .84$. At Time 2, the Cronbach's alpha for the TORR was $\alpha = .82$. On the basis of these analyses, we determined that scores on the TORR were a reliable indicator of relational reasoning at both testing points.

Test–retest Reliability

Test–retest reliability was established by asking participants to complete the TORR at Time 1 and subsequently to complete the same test in the same form, 3 weeks later, at Time 2. Consistent with prior research, this gap in administration was deemed appropriate to minimize any practice effects that may influence students' performance at Time 2 (Collie, Maruff, Darby, & McStephen, 2003). Performance on the TORR at Time 1 was significantly correlated with performance at Time 2, $r = 0.71$, $p < .001$. Furthermore, there were significant correlation coefficients between performance at Time 1 and Time 2 for each of the scales (analogy: $r = 0.61$, $p < .001$; anomaly: $r = 0.57$, $p < .001$; antinomy: $r = 0.44$, $p < .01$; and antithesis: $r = 0.48$, $p < .001$). Thus, as would be expected for a foundational cognitive ability in the absence of any direct intervention, there was sufficient stability in undergraduates' performance across a 3-week period.

Convergent and Discriminant Validity

With regard to relational reasoning as measured by the TORR, we investigated two forms of validity. First, the convergent validity between scores on the TORR and on a related test of relational reasoning, Raven's Advanced Progressive Matrices (RPM; Raven, 1941) was examined. Second, we investigated the degree to which undergraduate students' performance on the TORR

correlated with a measure of visual-spatial working memory to establish that relational reasoning could not largely be explained by this relevant individual difference ability.

Convergent Validity

Using the same group of undergraduate human development students (n = 71) described in the reliability section, the correlation between scores on the TORR and the RPM short form (RPM; Raven, 1941) was examined. It was our expectation that a positive, albeit moderate, correlation should arise between these two measures because the RPM, while widely used as a measure of fluid intelligence (e.g., Warren, Allen, Sommerfield, Deary, & Frier, 2004), comprised matrix analogy problems that tap one form of relational reasoning central to the TORR. In addition, this matrix format served as a general model for the development of the analogy scale in this study, although the specifics of item configuration on the RPM were not paralleled in the TORR. Moreover, researchers engaged in the study of relational reasoning have relied heavily upon the RPM (Dumas et al., 2013). Thus, the correlation between the RPM and the TORR was of particular interest.

Each student took the RPM under the mandated administration conditions and received extra credit for their participation, and the raw score was entered into the analysis. For these respondents, the mean score on the RPM short form was 9.33 ($SD = 2.21$). The two measures were determined to be significantly and moderately correlated ($r = .49$, $p < .001$). Thus, as hypothesized, performance on these two measures converged to a reasonable extent, although the strength of the correlation indicated that they were not capturing precisely the same cognitive processes. These findings seem consistent with the conceptual difference between the RPM and the TORR discussed. That is, while the TORR is designed to tap multiple forms of relational reasoning, the RPM is generally regarded as wholly analogical in nature (Bunge, Wendelken, Badre, & Wagner, 2005).

Furthermore, because the analogy scale of the TORR was designed using a similar matrix analogy format to the RPM, we examined the correlation between that particular scale and the RPM. This correlation was determined to be moderate ($r = .34$, $p = .004$), indicating an overlap in performance between these two indicators of relational reasoning, but also providing evidence that the two assessments are different in important ways. One factor that may have potentially contributed to the small degree of convergence between the RPM and the analogy scale of the TORR was item difficulty. For these participants, the items on the analogy scale of the TORR were seemingly more difficult than those on the RPM on the basis of mean performance. Specifically, the mean item difficulty level on the analogy scale of the TORR was .46 ($SD = .16$), while the mean item difficulty level on the RPM was .78 ($SD = .19$) for this group of participants.

Discriminant Validity

Because of the potential importance of visual-spatial working memory in the successful completion of figural reasoning items similar to those on the TORR (Logie, 2003), the correlation coefficients between the TORR and a measure of visual-spatial working memory was of interest. We wanted to ensure that the TORR was assessing something other than visual-spatial ability. For this analysis, the Shapebuilder task (Sprenger et al., 2013) was used as a measure of visual-spatial working memory. This task was preferable to other measures of visual-spatial working memory

TABLE 3
Correlations Between Measures and Scales

	1	*2*	*3*	*4*	*5*	*6*	*7*
1. Test of Relational Reasoning	—	.49**	.31*	.83**	.85**	.70**	.80**
2. Raven's Advanced Progressive Matrices		—	.22	.34**	.53**	.25*	.45**
3. Shapebuilder			—	.28*	.36**	.04	.25
4. Analogy				—	.67**	.45**	.50**
5. Anomaly					—	.40**	.61**
6. Antinomy						—	.44**
7. Antithesis							—

*$p < .05$. **$p < .001$.

such as complex span tasks, because it could be administered over the Internet without researcher supervision. As such, participants completed the Shapebuilder task outside of class and received extra credit for that participation.

The Shapebuilder task requires participants to maintain a mental representation of serially presented shapes (e.g., circle, square, or triangle) and recall those shapes in sequential order. In addition to order of presentation, the various shapes differed in number displayed, their color, and their location on a grid. Each participant has 15 min to be presented with varying serially presented strings of shapes. Each shape in a string correctly recalled earns a participant a certain number of points, calculated on the basis of the number of varying dimensions associated with that shape. The Shapebuilder task is then scored automatically.

In the present study, the mean number of points earned by the undergraduates was 1342.28 ($SD = 480.52$), and the correlation coefficient between the Shapebuilder task and the TORR for our participants was positive and low-moderate ($r = .31$, $p = .02$). This finding implies that while visual-spatial working memory plays a role in participants' ability to correctly respond to the items on the TORR, that construct does not account for an undue proportion of variance in scores. The correlation matrix for scores on the total TORR, individual scales, RPM, and the Shapebuilder task is displayed in Table 3.

Predictive Validity

Predictive validity for the TORR was established by examining the relation between the Test of Relational Reasoning and students' performance on two sets of released SAT items, one verbal and one math. Verbal and math SAT problems were selected as the predictive validity outcome measures for this study for a number of reasons. First, the SAT was developed as a measure of academic potential associated with college readiness and school success (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008). Second, the SAT is appropriate for use with a college-age sample and has been commonly used as a general scholastic assessment in empirical work (e.g., Frey & Detterman, 2004; Fuertes, Sedlacek, & Liu, 1994; Noftle & Robins, 2007). We further expected students to be familiar with SAT problems and instructions. It is important to note that from a predictive validity standpoint, the SAT problem sets are domain-specific in nature, whereas

the TORR was conceptualized as a domain-general assessment. Our intention in this predictive validity analysis was to determine whether students' performance on the TORR would predict their outcomes for the domain-specific verbal and math SAT problems.

Yet, another feature of the verbal and math SAT problems was highly pertinent to this predictive validity analysis. As with the TORR, relational reasoning was implicated by the general structure of the SAT problems. Specifically, the SAT problem sets comprised verbal analogies and mathematics problems. Together, these measures allowed us to capture how the TORR predicted performance on varied cognitive tasks, one with a strong linguistic component (i.e., verbal analogies) and one where relations were manifest in a nonlinguistic symbol system (i.e., mathematics).

For the prior convergent and discriminant analyses, the focus had been on measures that shared salient structural features with the measure of relational reasoning. For example, RPM (Raven, 1941) and Shapebuilder (Sprenger et al., 2013) are figural tasks that target fluid rather than crystallized cognitive abilities. Furthermore, both these measures are specialized assessments that bear little resemblance to everyday instructional tasks. For that reason, we elected to use outcome measures in the analysis of predictive validity that were familiar academic tasks to college students and that more explicitly assessed potentially more crystallized cognitive abilities.

Participants

Participants were 42 undergraduate students recruited from a human development course at a large Mid-Atlantic University, of which 30 participants completed all three measures and were included as the final sample. None of these participants had previously taken the TORR, or taken part in another aspect of this investigation. The sample consisted of 66.67% women ($n = 20$) with a mean age of 21.31 years old ($SD = 1.34$). These undergraduates were ethnically diverse: 46.67% identified as White ($n = 14$), 30.00% identified as Asian or Pacific Islander ($n = 9$), 13.33% identified as Hispanic/Latino ($n = 4$), 6.67% identified as Black ($n = 2$), and 3.33% identified as Middle Eastern ($n = 1$). The sample also consisted of 73.33% native English speakers ($n = 22$) and was primarily seniors (90%, $n = 27$). Participants reported having completed an average of 100.64 credits ($SD = 14.48$), and included ndividuals majoring in the social sciences and the humanities (66.66%, $n = 20$) and natural sciences (33.33%, $n = 10$). These undergraduates had an average self-reported GPA of 3.38 ($SD = 0.39$) on a 4-point scale.

Measures

In addition to completing the TORR, students were asked to complete two sets of items, one verbal and one math, from the SAT. All questions were taken from previously administered and publically released SAT problems.

TORR. The TORR, as previously described was administered.

Verbal SAT section. Students were asked to complete 13-items from a single verbal section of the SAT. Items were multiple-choice two-term verbal analogies with five answer choices. The directions for the verbal analogies section were identical to those provided on the SAT. Specifically, students were instructed: "Each question below consists of a related pair of words or phrases, followed by five pairs of words or phrases. Select the pair that best expresses a

relationship similar to that expressed in the original pair." The sample item provided to students gave them the target analogy of: *crumb: bread* and asked them to select an analogous relation from five choices: (a) ounce:unit; (b) splinter:wood; (c) water:bucket; (d) twine:rope; and (e) cream:butter; the correct answer was indicated as (b) splinter:wood. Students were asked to complete the 13 items in 15 min. Reliability for the 42 participants completing the 13-item verbal section was $\alpha = .69$.

Math SAT section. As with the verbal section, the math SAT section consisted of 13 multiple-choice items selected from previously administered and publically released SAT problems. For this test, the directions read as follows: "In this section, solve each problem. Then decide which is the best of the choices given (A through E). The use of a calculator is permitted." A sample problem is "If 7.5 is x percent of 75, what is x percent of 10?" Students were then asked to choose from five possible answer choices: (a) 10, (b) 1, (c) 0.75, (d) 0.1, and (e) 0.075. Instructions to participants were identical to those given on the SAT. Reliability for 13 items based on all 42 students completing the math section was $\alpha = .74$.

Procedures. Students completed the three measures during two sessions 2 weeks apart. In Session 1, students completed the verbal and math SAT measures in class. In Session 2, students completed the TORR online outside of class. There was no time limit set for students' completion of the TORR.

Results

Students' average score on the verbal section of the SAT was 8.10 ($SD = 2.40$). Participants answered 3 to 12 items correctly. The average score on the math section of the SAT was 9.67 ($SD = 2.39$), and students answered 4 to 13 items correctly. Participants' average score on the TORR was 15.57 ($SD = 3.98$), ranging from students answering 7 to 24 items correctly.

Two simple linear regressions were conducted. In the first, the TORR was found to be a significant predictor of performance on the verbal section of the SAT, $F(1, 28) = 16.13$, $p < .001$ with $\beta = 0.36$, $t = 4.02$, $p < .001$. Scores on the TORR explained 36.6% of variance in performance on the verbal section of the SAT ($R^2 = 0.37$). The TORR was also a significant predictor of students' performance on the math section of the SAT, $F(1, 28) = 4.34$, $p < .05$ with $\beta = 0.23$, $t = 2.08$, $p < .05$. Performance on the TORR explained 13.4% of variance in scores on the math section of the SAT ($R^2 = 0.13$).

On the basis of these results, we determined that the TORR demonstrated good predictive validity with regard to a well-established measure of academic potential, whether that measure comprised verbal analogies or mathematical problems.

STAGE 3: DIMENSIONALITY OF RELATIONAL REASONING

The purposes of this third stage of analysis were twofold. First, we wanted to address whether relational reasoning manifests differently within undergraduates enrolled in a natural science versus a social science course. Specifically, we examined whether mean differences in relational reasoning performance existed between individuals in two different courses geared toward students in different major areas of study. Second, we wanted to address whether relational reasoning

is best represented as a unidimensional or multidimensional construct. To this end, confirmatory factor analyses were conducted to compare the fit of three theoretically defensible models. Given a priori identification of theoretically derived models and the interest in testing these competing models relative to one another, it was determined that a confirmatory factor analysis, rather than an exploratory factor analysis, would provide the flexible framework needed to address the question of dimensionality. Moreover, an examination of the relation among latent factors of interest for this investigation is possible only through confirmatory factor analyses rather than through exploratory techniques.

The three models are depicted in Figure 5. Model A was a one-factor model, with a single factor, relational reasoning, regarded as the underlying latent construct driving performance across items on each of the measure's scales. This model is consistent with writings of James (1890) and the recent neuroscience research (e.g., Crone et al., 2009) for which relational reasoning is conceived as an undifferentiated construct. Should this model prove most salient, then it would be argued the four scales of the TORR, although formulated according to distinct rules, capture the singular construct of relational reasoning. By comparison, Model B depicted relational reasoning as consisting of four separate, yet related, latent constructs corresponding to each of the identified manifestations of relational reasoning measured (i.e., analogy, anomaly, antinomy, and antithesis). For this analysis, the items were only allowed to load on the corresponding dimension of relational reasoning to which they pertained and the latent factors were allowed to correlate. This model is consistent with theoretical depictions of Alexander and colleagues (Alexander & the Disciplined Reading and Learning Research Laboratory, 2013; Dumas et al., 2013), who have described the construct of relational reasoning as consisting of distinct, although associated, forms.

The third model (Model C) tested in this analysis was different from the previous two in that it not only allowed for discernible manifestations of relational reasoning, but also tested simultaneously for the existence of a higher order factor. Specifically, rather than specify a correlation between latent factors as was done for Model B, Model C included a higher order factor, relational reasoning, which was regarded as influencing the latent factors of analogy, anomaly, antinomy, and antithesis. Parallels for such a model can be found within the intelligence literature where it has been argued that a general *g* factor underlies mental tasks that share certain cognitive abilities or processes (Spearman, 1927; Sternberg, 2013). For the present study, this particular model would suggest that there are varied forms of relational reasoning that can be measured reliably and validly. Moreover, these forms are particular iterations of the more general and encompassing ability to reason relationally.

Participants

Participants included 71 participants recruited from an upper-level human development course at a large Mid-Atlantic university (see Stage 2 for participant demographics), and 549 students recruited from one large lecture section of a biology course at a large northeastern university. For the sample of 549, participants ranged in age from 18 to 35, ($M = 19.26$, $SD = 1.54$) with the majority of participants being between the ages of 18 and 21 ($n = 520$; 94.72%). Few participants were 22 and older ($n = 29$; 5.28%). While the majority of participants self-reported as White ($n = 442$; 80.51%), also present were Asian/Pacific Islander ($n = 36$; 6.56%), Black ($n = 30$; 5.46%), Hispanic ($n = 23$; 4.19%) and Native American ($n = 3$; 0.55%) students. A small percentage of respondents ($n = 15$; 2.73%) chose not to identify their race. Participants represented a range

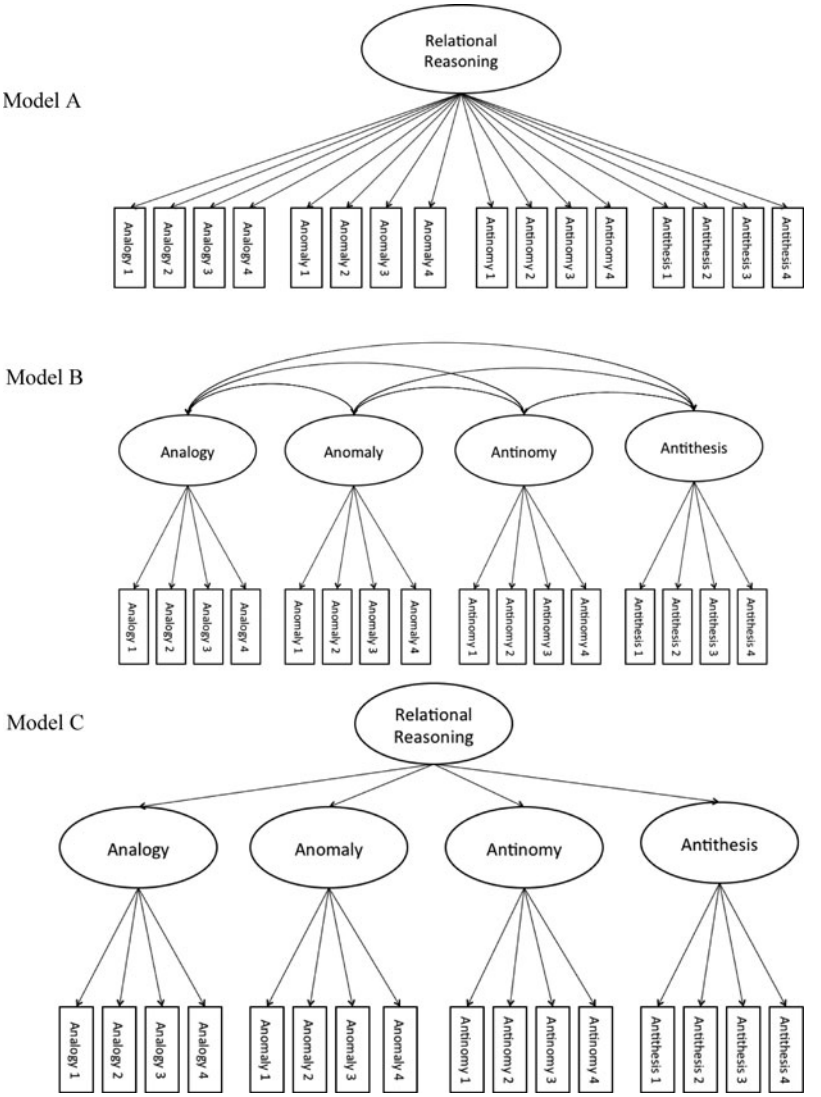

FIGURE 5 Three theoretically viable models of relational reasoning tested by means of confirmatory factor analyses. Only four items per scale are depicted; however, all eight items per scale were used in the analyses.

of semester standing: freshman ($n = 225$; 40.98%), sophomore ($n = 180$; 32.79%), junior ($n = 84$; 15.30%), and senior or greater ($n = 60$; 10.93%). A moderate proportion of the participants in this entry-level course reported having declared a biology-specific major ($n = 211$; 38.43%), and a large proportion had identified majors in the natural or physical sciences or were pursuing premed or health-related degrees ($n = 467$; 85.06%). Most participants reported English as their first language ($n = 509$; 92.71%).

Cross-Course Comparison

To address the question of whether students in different courses manifest different levels of relational reasoning, we compared the participants enrolled in the biology course at a large northeastern university with the participants enrolled in a human development course at a large Mid-Atlantic university. Students in the biology course primarily consisted of natural and physical science majors, while students in the human development course consisted primarily of social science majors. A comparison of mean differences for the overall test and for each of the scales of the TORR was conducted using independent-samples *t* tests for unequal variances (to account for differences in sample size across groups). No significant mean differences were identified between groups for the total score of the TORR, $t(76.74) = -.04$, $p = .97$, or for each of the individual scales: analogy scale, $t(86.45) = -1.41$, $p = .16$; anomaly scale, $t(83.52) = .46$, $p = .65$; antinomy scale, $t(86.54) = .85$, $p = .40$; or antithesis scale, $t(85.73) = .24$, $p = .81$.

On the basis of these results for the total score and for each of the four scales, it was determined that the course of study for undergraduates in our investigation (i.e., natural science vs. social science) did not manifest statistically distinct patterns of performance in relational reasoning. Given the comparability of these groups in terms of relational reasoning, the samples were combined for further analyses. Before conducting the factor analyses, internal reliability of the TORR was calculated for this larger sample ($\alpha = .78$).

Confirmatory Factor Analysis

The confirmatory factor analysis models were run in Mplus 6 (Muthén & Muthén, 2010) using a robust weighted least squares estimation. Weighted least squares estimation has been identified as an appropriate estimation procedure for binary data with samples greater than 200 (Flora & Curran, 2004). It has also been shown to adequately control for Type I error and to provide more accurate parameter estimates and test statistics than weighted least squares (T. A. Brown, 2006; Flora & Curran, 2004) or maximum likelihood estimate (Beauducel & Yorck Herzberg, 2006) when analyzing binary data. In this case, data are considered binary because each item on the TORR is coded as correct or incorrect. A combination of fit statistics to capture the absolute, incremental, and parsimonious fit were examined to determine the fit of the data to each of the three models. In addition to the relative chi-square values, comparative fit index values $\geq .96$, Tucker-Lewis Index values $\geq .90$, and root mean square error of approximation values $\leq .05$ (Yu, 2002) were identified as a priori guidelines for determining appropriate fit of the tested models. These guidelines for binary data with $N > 250$ are slightly more stringent than Hu and Bentler's (1999) well-cited criteria. As recommended values to indicate appropriate fit for models with categorical data have not been well established (Beauducel & Yorck Herzberg, 2006), and prior guidelines were based on large factor loadings (Yu, 2002), these served as general recommendations. Fit statistics for the three models are reported in Table 4.

On the basis of the reported fit statistics, it was determined that Model A did not adequately fit the data. The comparative fit index and Tucker-Lewis Index were lower than the recommended values. In addition, the chi-square value for Model A was the largest of the three models, indicating that the other models were a better fit to the data. The fit of Model B and Model C were similar, and both models fit the data well on the basis of the established criteria. Moreover, Models B and C support the conceptualization of relational reasoning as composed of four independent but

TABLE 4
Confirmatory Factor Analysis Model Fit Statistics

Model	χ^2	df	*Comparative fit index*	*Tucker-Lewis Index*	*Root mean square error of approximation*
Model A	883.321	464	0.861	0.852	0.038
Model B	583.869	458	0.958	0.955	0.021
Model C	584.181	460	0.959	0.956	0.021

related components, the premise put forward as the guiding framework for this study (Dumas et al., 2013).

Models B and C correspond to the literature-based conceptualization of relational reasoning that drove the development of the TORR, a conceptualization that proposed four distinct, yet related, forms of relational reasoning. While Model C is the more statistically parsimonious model, Models B and C provide an explanation of the relation among the latent factors. Specifically, Model B posits that the forms of relational reasoning are directly related, while Model C indicates that the forms are related through an overarching factor. Compared with Model C, Model B does not require the inclusion of an overarching factor to explain the relation among the four forms of relational reasoning. The presence of an overarching factor, as represented in Model C, has often been posited in models of intelligence (Sternberg, 2013), yet Model B is potentially more useful to guide future investigations and interventions aimed at promoting students' relational reasoning. Given the fit of these two models, both were acknowledged in the present study as viable models. The standardized model for Model B with path coefficients and latent factor correlations is presented in Figure 6. For this four-factor model, the correlations between the factors ranged from .41 to .78. The standardized model for Model C is presented in Figure 7. For this model, the paths between the overarching factor, relational reasoning, and each of the four forms ranged from .46 to .92. For Models B and C, all of the factor loadings were significant at $\alpha < .05$, and for each model, 29 out of the 32 loadings were greater than .3.

After the confirmatory factor analyses empirically identified the four scales as potentially representing latent factors, the internal reliability for each of these scales or factors was calculated (analogy: $\alpha = .65$; anomaly: $\alpha = .54$; antinomy: $\alpha = .51$; antithesis: $\alpha = .59$). While these

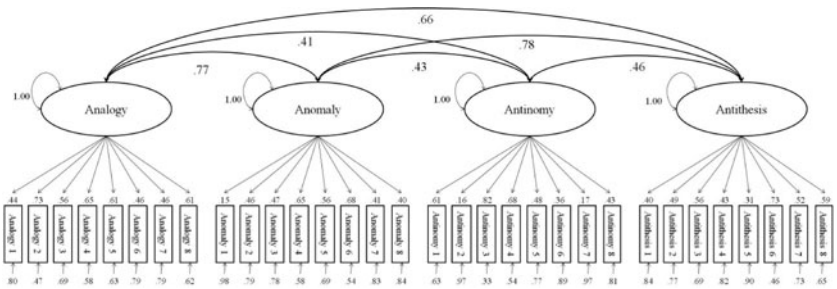

FIGURE 6 Standardized model with path coefficients and latent factor correlations for the four-factor model of relational reasoning, Model B.

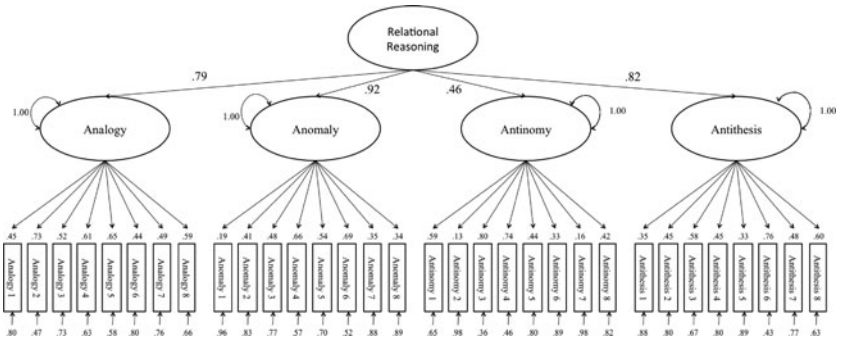

FIGURE 7 Standardized model with path coefficients and latent factor correlations for the two-level model of relational reasoning with an overarching factor, Model C.

reliability coefficients are lower than those associated with the TORR as a whole, the coefficients calculated for the scales were likely attenuated by the small number of items per scale (i.e., 8), and should be cautiously interpreted as a potential underestimate of scale score reliabilities (Cronbach, 1951; Sijtsma, 2009).

Conclusions and Implications

Despite the essential and pervasive role of relational reasoning in everyday cognition and in academic development (e.g., Ehri, Satlow, & Gaskins, 2009; Trey & Khan, 2008; Tzuriel & George, 2009), its measurement has remained elusive. Thus, the overarching purpose of this investigation was to explore this foundational but elusive construct by means of a figural measure of relational reasoning. In part, this measurement conundrum has persisted because the conceptualization of relational reasoning has itself proven problematic. For example, some scholars have described the construct as incorporating nearly every human cognitive function (Hofstadter & Sanders, 2013), whereas others have described it so as to entail only a particular form of conceptual patterning (e.g., Farrington-Flint, Canobi, Woor, & Faulkner, 2007). In this study, the definition that guided measurement was derived from an extensive review of the theoretical and empirical literature (Dumas et al., 2013). That resulting definition positioned relational reasoning as a broadly applicable and foundational construct supporting the discernment of meaningful patterns within any stream of information, be it linguistic, graphic, or numeric in nature (Alexander & the Disciplined Reading and Learning Research Laboratory, 2012; Krawczyk, 2012). However, despite this broad conceptualization of relational reasoning in general, specificity of definition and operationalization was offered for four particular forms of the construct that were targeted in measurement. These forms (i.e., analogical, anomalous, antinomous, and antithetical reasoning), each correspond to a particular mapped relation within a body of information (i.e., similarity, discrepancy, incompatibility, and polarity).

Lingering issues in the measurement of relational reasoning are also reflective of the significant challenges that come in the operationalization of this foundational construct. For example, it is imperative to filter out the variance in human performance that might be explained by task familiarity, background knowledge, or visual-spatial working memory to the degree possible, if

there is hope of developing a clear picture of respondents' relational reasoning ability. Because no multifaceted measure of this latent construct was to be found, we had to develop one. Specifically, we devised the Test of Relational Reasoning as a figural measure, which we judged would tap more of the fluid than crystallized dimensions of human cognition (Cattell, 1961; Roberts & Lipnevich, 2012). In addition, we targeted the measure for older adolescents and adults whom we presumed would have reached the level of cognitive development to undertake such an abstract and challenging task (Goswami et al., 2011). Through a series of pilot studies, we also sought to ensure that the items were within a suitable range of difficulty and that the time required for completion of the entire measure did not prove overly taxing. Interestingly, the rather large standard deviation for completion time (more than 20 min) opens the door to questions about speed of processing as a factor in relational reasoning. It would appear that among the undergraduates we tested there were a number who were able to discern patterns quickly and accurately, whereas others required more deliberate reflection to reach the same outcome. Thus, the role that speed of processing may play in relational reasoning remains an empirical question worthy of examination.

With these conditions met in Stage 1 of the study, and with a workable measure in hand, we engaged in two subsequent stages of data analysis. In Stage 2, we set out to establish the reliability and validity of the data for the relational reasoning measure for undergraduate students enrolled in social science and natural science courses. The tests of reliability demonstrated that the overall measure of this latent construct was performing quite well and that the scales were functioning adequately. We also determined that students' performance on the TORR remained stable over time and that it did not matter whether the test was delivered on paper or online. In terms of validity, we found that there was a reasonable, albeit moderate, association between students' relational reasoning performance and their scores on the RPM short form (Raven, 1941), a test that shared certain format and content features. It was also determined that students' Shapebuilder (Sprenger et al., 2013) performance was only moderately correlated with TORR performance, suggesting that there is more to reasoning relationally with figural items than can be accounted for by visual-spatial ability.

Because of the foundational character ascribed to relational reasoning, it was important for us to investigate the degree to which such abilities carried forward into other cognitive tasks. In the present study, we were able to examine the predictive validity of the TORR in terms of a task that on the surface bore little resemblance to that measure. We chose a typical undergraduate-level academic task, verbal and math SAT problems, and we tested whether students' performance on these measures could be predicted by their relational reasoning performance, and it did. We were encouraged by this outcome and will begin to delve into other ecologically valid tasks for which relational reasoning, as measured by a domain-general figural task, may be predictive. For example, including reading tasks in which the texts expressly incorporate analogical, anomalous, antinomous, or antithetical references may afford a more direct test of predictive validity in the future. Moreover, think-aloud protocols investigating students' ability to reason relationally in ill-structured tasks may be an important future direction. An examination of the verbal interactions between an attending physician and residents provided insights into the role that relational reasoning plays in medical diagnosis and treatment (Dumas, Alexander, Baker, Jablansky, & Dunbar, 2014).

In Stage 3, we tested three theoretically viable models to discern whether relational reasoning is best understood as a unidimensional or multidimensional construct for the targeted population. Before conducting this analysis, we compared performance across the social science and natural

science courses and determined that the performance for these two classes did not differ by overall score or for the four specific scales. This cross-class comparison addressed one of our guiding questions about relational reasoning; that is, whether there would be differences in performance attributable to courses of study. For this investigation, the answer was no.

Of the three models we tested, the one that represented relational reasoning as a unidimensional construct was deemed unacceptable. This means that it is important to consider the form of relational reasoning that is being assessed. The remaining two models depicted relational reasoning as multidimensional. What ultimately differentiated the two models was the inclusion of a higher-order factor over and above the factors representing the four manifestations of relational reasoning. As both models had acceptable fit statistics, future research is needed to examine the comparative viability of these models, particularly as it may reflect developmental differences.

Despite these promising findings, the investigation of individuals' ability to discern meaningful patterns within any informational stream (i.e., relational reasoning) initiated here must be extended and deepened in various systematic ways to better understand its nature and its role in human learning and development. For example, there are a series of developmental questions that warrant exploration. One of those questions pertains to the onset of relational reasoning abilities. Determining when and how relational reasoning emerges necessitates the measurement of this construct in populations younger than the older adolescents and young adults studied here. Would we expect that some variation of the TORR would function with middle school or high school students, for example, or that the multidimensionality identified here would similarly manifest in these younger populations?

Questions about the underlying structure of relational reasoning could likewise be examined developmentally. For example, is it reasonable to hypothesize that relational reasoning is more unidimensional in its underlying structure with cognitively less advanced populations and only begins to differentiate as individuals cognitively mature? Conversely, it is conceivable that the model of relational reasoning encompassing a higher-order factor may be more viable for cognitively advanced populations; that is, those who have learned to harness the power of this foundational ability.

Also from a developmental standpoint, it seems worthwhile to examine changes in relational reasoning that occur over time or to gauge the degree to which established developmental trajectories can be influenced by external factors such as explicit training. Such questions require longitudinal studies that track relational reasoning patterns over substantial time frames and that experimentally seek to intervene in the normal course of its development. Relatedly, it would be informative to position studies of relational reasoning within the context of expertise development. For example, would we expect shifts in relational reasoning performance among older adolescents and adults as they move toward expertise in their chosen fields of study? Addressing these questions would minimally call for cross-sectional studies that include individuals who are identified novices and experts in select domains, if not longitudinal studies that focus on individuals at particularly critical junctures in their academic development.

Understanding the role of relational reasoning in human learning and development would also be enhanced by additional studies of its predictive nature. For example, is success in specific academic domains associated with relational reasoning ability in general or does the ability to discern relational patterns of a certain type (e.g., anomalous or antinomous) play a more central role in select fields of study? In the present study, we conducted a cursory analysis of undergraduates enrolled in social science and natural science coursework and found no differences

in overall performance or in scale profiles. However, more detailed and focused analyses are required to better address such domain-specific questions. For example, would we expect that individuals successfully pursuing degrees in sciences reliant on classification processes (e.g., biology) demonstrate greater facility with antinomous reasoning than those pursuing degrees in more mechanistic sciences (e.g., mechanical engineering)?

Our purpose in developing a psychometrically sound measure of relational reasoning was never solely focused on empirical investigation. There were also strong practical aims underlying our efforts. As with others (Stanovich & Stanovich 2003; Sternberg, 1988), we hold that having the means to uncover foundational abilities that may be otherwise overlooked in traditional intelligence measures or within the educational system is critical, especially when those abilities seem core to continued human learning and development. Being equipped to gauge students' abilities to reason analogically, anomalously, antinomously, and antithetically without the added complications of language differences, academic history, or cultural background seems invaluable.

As this brief discussion suggests, there is much left unanswered about relational reasoning and its measurement. The present study hopefully offers evidence that such further pursuits are not only feasible but also desirable. In general, we concur with William James (1890) that without the ability to perceive relevant relations among the objects of our perception, even when they are separated by time and space, we are potentially imprisoned in a world of isolated stimuli. Furthermore, without the means to assess and measure relational reasoning, we are equally hampered in our ability to identify and nurture this most basic of cognitive abilities.

AUTHOR NOTES

Patricia A. Alexander is the Jean Mullan Professor of Literacy and Distinguished Scholar-Teacher in the Department of Human Development at the University of Maryland, College Park. She is the senior editor of Contemporary Educational Psychology and director of the Disciplined Reading and Learning Research Lab. She has published over 280 articles, books, or chapters in the area of learning and instruction. **Denis Dumas** is a doctoral candidate of educational psychology in the department of Human Development and Quantitative Methodology at the University of Maryland, College Park. His research focuses on cognitive abilities such as reasoning, intelligence, and creativity and their predictive relations to success in academic and professional settings. **Emily M. Grossnickle** is a doctoral candidate in educational psychology in the Department of Human Development and Quantitative Methodology at the University of Maryland, College Park. Her research focuses on how learner characteristics such as epistemic beliefs, curiosity, interest, and relational reasoning impact learning processes and academic development. **Alexandra List** is a doctoral candidate in educational psychology in the Department of Human Development and Quantitative Methodology at the University of Maryland, College Park. Her research interests include multiple source use and evaluation, the impact of task features on multiple source use, and process measures of online source use. **Carla M. Firetto** is postdoctoral research fellow on an Institute of Education Sciences funded grant at The Pennsylvania State University. Her current research pertains to the influence of small group, Quality Talk discussions on elementary school students' high-level comprehension of text, as well as how students learn when reading multiple texts.

REFERENCES

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*, 288–318.

Alexander, P. A., & Baggetta, P. (2013). Percept-concept coupling and human error. In D. N. Rapp & J. L. G. Baasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 297–329). Boston, MA: MIT Press.

Alexander, P. A., & Murphy, P. K. (1999). Nurturing the seeds of transfer: A domain-specific perspective. *International Journal of Educational Research*, *31*, 561–576.

Alexander, P. A., & the Disciplined Reading and Learning Research Laboratory. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist*, *47*, 259–280. doi:10.1080/00461520.2012.722511

Alexander, P. A., Pate, P. E., Kulikowich, J. M., Farrell, D. M., & Wright, N. L. (1989). Domain-specific and strategic knowledge: Effects of training on students of differing ages or competence levels. *Learning and Individual Differences*, *1*, 283–325. doi:10.1016/1041–6080(89)90014–9

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Andiliou, A., Ramsay, C., Murphy, P. K., & Fast, J. (2012). Weighing opposing positions: Examining the effects of intratextual persuasive messages on students' knowledge and beliefs. *Contemporary Educational Psychology*, *37*, 113–127. doi:10.1016/j.cedpsych.2011.10.001

Andrews, G., Birney, D., & Halford, G. S. (2006). Relational processing and working memory capacity in comprehension of relative clause sentences. *Memory & Cognition*, *34*, 1325–1340. doi:10.3758/BF03193275

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. doi:10.1146/annurev-psych-120710-100422

Baker, S. T., Friedman, O., & Leslie, A. M. (2010). The opposites task: Using general rules to test cognitive flexibility in preschoolers. *Journal of Cognition and Development*, *11*, 240–254. doi:10.1080/15248371003699944

Bassok, M., Dunbar, K. N., & Holyoak, K. J. (2012). Introduction to the special section on the neural substrate of analogical reasoning and metaphor comprehension. *Journal of Experimental Psychology*, *38*, 261–263.

Beauducel, A., & Yorck Herzberg, P. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203. doi:10.1207/s15328007sem1302_2

Bianchi, I., Savardi, U., & Burro, R. (2011). Perceptual ratings of opposite spatial properties: Do they lie on the same dimension? *Acta Psychologica*, *138*, 405–418. doi:10.1016/j.actpsy.2011.08.003

Bianchi, I., Savardi, U., & Kubovy, M. (2011). Dimensions and their poles: A metric and topological approach to opposites. *Language and Cognitive Processes*, *26*, 1232–1265. doi:10.1080/01690965.2010.520943

Bohan, J., & Sanford, A. (2008). Semantic anomalies at the borderline of consciousness: An eye-tracking investigation. *The Quarterly Journal of Experimental Psychology*, *61*, 232–239. doi:10.1080/17470210701617219

Braasch, J. L. G., Bråten, I., Strømsø, H. I., Anmarkrud, Ø., & Ferguson, L. E. (2013). Promoting secondary students' evaluation of source features of multiple documents. *Contemporary Educational Psychology*, *38*, 180–195. doi:10.1016/j.cedpsych.2013.03.003

Broughton, S. H., Sinatra, G. M., & Reynolds, R. E. (2010). The nature of the refutation text effect: An investigation of attention allocation. *The Journal of Educational Research*, *103*, 407–423. doi:10.1080/00220670903383101

Brown, A. L., Kane, M. J., & Long, C. (1989). Analogical transfer in young children: Analogies as tools for communication and exposition. *Applied Cognitive Psychology*, *3*, 275–293. doi:10.1002/acp.2350030402

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Bulloch, M. J., & Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science*, *12*, 114–122. doi:10.1111/j.1467-7687.2008.00738.x

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex*, *15*, 239–249. doi:10.1093/cercor/bhh126

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology*, *31*, 161–179. doi:10.1037/h0059043

Cattell, R. B. (1961). Fluid and crystallized intelligence. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 738–746). East Norwalk, CT: Appleton-Century-Crofts.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81. doi:10.1016/0010-0285(73)90004-2

Chi, M. T. H., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 3–27). Amsterdam, The Netherlands: Kluwer.

Chi, M. T. H., & Slotta, J. D. (1993). The ontological coherence of intuitive physics. *Cognition and Instruction*, *10*, 249–260. doi:10.1207/s1532690xci1002&3_5

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*, 1–49. doi:10.2307/1170558

Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology*, *94*, 327–343. doi:10.1037/0022–0663.94.2.327

Chinn, C. A., & Samarapungavan, A. (2009). Conceptual change—Multiple routes, multiple mechanisms: A commentary on Ohlsson (2009). *Educational Psychologist*, *44*, 48–57. doi:10.1080/00461520802616291

Cole, M., & Wertsch, J. V. (1996). Beyond the individual-social antinomy in discussions of Piaget and Vygotsky. *Human Development*, *39*, 250–256. doi:10.1159/000278475

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the test performance of neurologically normal individuals assessed at brief test–retest intervals. *Journal of the International Neuropsychological Society*, *9*, 419–428.

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Cognitive activities in complex science text and diagrams. *Contemporary Educational Psychology*, *35*, 59–74. doi:10.1016/j.cedpsych.2009.10.002

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–34. doi:10.1007/BF02310555

Crone, E. A., Wendelken, C., van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science*, *12*, 55–66. doi:10.1111/j.1467–7687.2008.00743.x

de Saussure, F. (2011). *Course in general linguistics* (W. Baskin, Trans.). New York, NY: Columbia University Press. (Original work published 1916)

Dinsmore, D. L., Doyle, S., Baggetta, P., & Loughlin, S. M. (2012, April). *Not all transfer is created equal: Making the case for different types of transfer*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, Canada.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Mahwah, NJ: Erlbaum.

Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning and its manifestations in the educational context: A systematic review of the literature. *Educational Psychology Review*, *25*, 391–427. doi:10.1007/s10648-013-9224-4.

Dumas, D., Alexander, P. A., Baker, L. M., Jablansky, S., & Dunbar, K. N. (2014). Relational reasoning in medical education: Patterns in discourse and diagnosis. *Journal of Educational Psychology*. doi:10.1037/a003677

Dumontheil, I., Houlton, R., Christoff, K., & Blakemore, S. J. (2010). Development of relational reasoning during adolescence. *Developmental Science*, *13*, 15–24. doi:10.1111/j.1467–7687.2010.01014.x

Dunbar, K. (2001). The analogical paradox: Why analogy is so easy in naturalistic settings yet so difficult in the psychological laboratory. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 313–334). Cambridge, MA: MIT Press.

Ehri, L. C., Satlow, E., & Gaskins, I. (2009). Grapho-phonemic enrichment strengthens keyword analogy instruction for struggling young readers. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, *25*, 162–191. doi:10.1080/10573560802683549

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive experiential and analytic-rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390–405.

Farrington-Flint, L., Canobi, K. H., Woor, C., & Faulkner, D. (2007). The role of relational reasoning in children's addition concepts. *British Journal of Developmental Psychology*, *25*, 227–246. doi:10.1348/026151006X108406

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive developmental inquiry. *American Psychologist*, *34*, 906–911. doi:10.1037/0003-066X.34.10.906

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analyses with ordinal data. *Psychological Methods*, *9*, 466–491. doi:10.1037/1082-989X.9.4.466

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between scholastic assessment test and general cognitive ability. *Psychological Science*, *15*, 373–378. doi:10.1111/j.0956-7976.2004.00687.x

Fuertes, J. N., Sedlacek, W. E., & Liu, W. M. (1994). Using the SAT and noncognitive variables to predict the grades and retention of Asian American university students. *Measurement and Evaluation in Counseling and Development*, *27*, 74–84.

Gardner, H. (1995). Perennial antinomies and perpetual redrawings: Is there progress in the study of mind? In R. Solso & D. Massaro (Eds.), *The science of the mind: 2001 and beyond* (pp. 65–78). New York, NY: Oxford University Press.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355. doi:10.1016/0010-0285(80)90013-4

Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.

Goswami, U., Leevers, H., Pressley, S., & Wheelwright, S. (1998). Causal reasoning about pairs of relations and analogical reasoning in young children. *British Journal of Developmental Psychology*, *16*, 553–569.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–831.

Hofstadter, D. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 499–538). Cambridge, MA: MIT Press.

Hofstadter, D., & Sander, E. (2013). *Surfaces and essences: Analogy as the fuel and fire of thinking*. New York, NY: Basic Books.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 19, pp. 59–87). New York, NY: Academic Press.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). New York, NY: Oxford University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118

Ivanova, I., Pickering, M. J., Branigan, H. P., McLean, J. F., & Costa, A. (2012). The comprehension of anomalous sentences: Evidence from structural priming. *Cognition*, *122*, 193–209. doi:10.1016/j.cognition.2011.10.013

James, W. (1890). *The principles of psychology*. New York, NY: Holt.

Kjeldergaard, P. M., & Higa, M. (1962). Degree of polarization and the recognition value of words selected from the semantic atlas. *Psychological Reports*, *11*, 629–630. doi:10.2466/pr0.1962.11.3.629

Klahr, D., & Dunbar, K. (1988). The psychology of scientific discovery: Search in two problem spaces. *Cognitive Science*, *12*, 1–48.

Köhler, W. (1951). Relational determination in perception. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon Symposium* (pp. 200–243). Oxford, UK: Wiley.

Krawczyk, D. C. (2012). The cognition and neuroscience of relational reasoning. *Brain Research*, *1428*, 13–23. doi:10.1016/j.brainres.2010.11.080

Krawczyk, D. C., McClelland, M. M., & Donovan, C. M. (2011). A hierarchy for relational reasoning in the prefrontal cortex. *Cortex*, *47*, 588–597. doi:10.1016/j.cortex.2010.04.008

Kreezer, G., & Dallenbach, K. M. (1929). Learning the relation of opposition. *The American Journal of Psychology*, *41*, 432–441. doi:10.2307/1414683

Krigbaum, G., Amin, K., Virden, T. B., Baca, L., & Uribe, A. (2012). A pilot study of the sensitivity and specificity analysis of the standard-Spanish version of the culture-fair assessment of neurocognitive abilities and the Examen Cognoscitivo Mini-Mental in the Dominican Republic. *Applied Neuropsychology*, *19*, 53–60. doi:10.1080/09084282.2011.643938

Logie, R. H. (2003). Spatial and visual working memory: A mental workspace. In D. E. Irwin & B. H. Ross (Eds.), *Cognitive vision: The psychology of learning and motivation* (Vol. 42, pp. 37–78). San Diego, CA: Academic Press.

Marková, I. (1987). On the interaction of opposites in psychological processes. *Journal for the Theory of Social Behaviour*, *17*, 279–299. doi:10.1111/j.1468-5914.1987.tb00100.x

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT* (College Board Research Report No. 2008–4). New York, NY: College Board.

Mosenthal, P. B. (1988). Anopheles and antinomies in reading research (Research views). *Reading Teacher*, *42*, 234–235.

Murphy, P. K. (2001). What makes a text persuasive? Comparing students' and experts' conceptions of persuasiveness. *International Journal of Educational Research*, *35*, 675–698.

Murphy, P. K., & Mason, L. (2006). Changing knowledge and beliefs. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 305–324). Mahwah, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

Naglieri, J. A., & Insko, W. R. (1986). Construct validity of the Matrix Analogies Test—Expanded Form. *Journal of Psychoeducational Assessment*, *4*, 243–255. doi:10.1177/073428298600400308

Nersessian, N. J., & Chandrasekharan, S. (2009). Hybrid analogies in conceptual innovation in science. *Cognitive Systems Research*, *10*, 178–188. doi:10.1016/j.cogsys.2008.09.009

Newcombe, N. S., Ambady, N., Eccles, J., Gomez, L., Klahr, D., Linn, M., ... Mix, K. (2009). Psychology's role in mathematics and science education. *American Psychologist*, *64*, 538–550. doi:10.1037/a0014813

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, *93*, 116–130. doi:10.1037/0022-3514.93.1.116

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 510–520.

Nussbaum, E. M., & Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, *97*, 157–169. doi:10.1037/0022-0663.97.2.157

Opfer, J. E., & Gelman, S. A. (2011). Development of the animate-inanimate distinction. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 213–238). Oxford, UK: Wiley-Blackwell.

Qualtrics (2012). *Qualtrics software version 37.892*. Provo, UT, USA. Available at http://www.qualtrics.com

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology*, *19*, 137–150. doi:10.1111/j.2044-8341.1941.tb00316.x

Roberts, R. D., & Lipnevich, A. A. (2012). From general intelligence to multiple intelligences: Meanings, models, and measures. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook, Vol 2: Individual differences and cultural and contextual factors* (pp. 33–57). Washington, DC: American Psychological Association.

Russell, B. (1908). Mathematical logic as based on the theory of types. *American Journal of Mathematics*, *30*, 222–262.

Russell, B., & Lackey, D. (1973). *Essays in analysis*. New York, NY: Allen & Unwin.

Savage, R. S., Deault, L., Daki, J., & Aouad, J. (2011). Orthographic analogies and early reading: Evidence from a multiple clue word paradigm. *Journal of Educational Psychology*, *103*, 190–205. doi:10.1037/a0021621

Shaumyan, S. (2006). Antinomies of language and language operations of the mind. In H. R. Arabnia, E. B. Kozerenk, & S. Shaumyan (Eds.), *Proceedings of the 2006 International Conference on Machine Learning; Models, Technologies and Applications* (pp. 3–9). Las Vegas, NV: CSREA Press.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. doi:10.1007/s11336-008-9101-0

Sinatra, G. M., & Broughton, S. H. (2011). Bridging reading comprehension and conceptual change in science education: The promise of refutation text. *Reading Research Quarterly*, *46*(4), 374–393. doi:10.1002/RRQ.005

Slotta, J. D., & Chi, M. T. H. (2006). Helping students understand challenging topics in science through ontology training. *Cognition and Instruction*, *24*, 261–289. doi:10.1207/s1532690xci2402_3

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. *Educational Researcher*, *18*, 8–14.

Sorensen, R. A. (2003). *A brief history of the paradox: Philosophy and the labyrinths of the mind*. New York, NY: Oxford University Press.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence*, *41*, 638–663. doi:10.1016/j.intell.2013.07.013

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

Stanovich, P. J., & Stanovich, K. E. (2003). *Using research and reason in education: How teachers can use scientifically based research to make curricular and instructional decisions*. Washington, DC: National Institute for Literacy.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Mahwah, NJ: Erlbaum.

Sternberg, R. J. (1988). Applying cognitive theory to the testing and teaching of intelligence. *Applied Cognitive Psychology*, *2*, 231–255. doi:10.1002/acp.2350020402

Sternberg, R. J. (2013). Intelligence. In D. K. Freedheim & I. B. Weiner (Eds.), *Handbook of psychology, Vol. 1: History of psychology* (2nd ed., pp. 155–176). Hoboken, NJ: Wiley.

Stevenson, C. E., Touw, K. W. J., & Resing, W. C. M. (2011). Computer or paper analogy puzzles: Does assessment mode influence young children's strategy progression? *Educational and Child Psychology*, *28*, 67–84.

Trey, L., & Khan, S. (2008). How science students can learn about unobservable phenomena using computer-based analogies. *Computers and Education*, *51*, 519–529. doi:10.1016/j.compedu.2007.05.019

Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2009). How do scientists respond to anomalies? Different strategies used in basic and applied science. *Topics in Cognitive Science*, *1*, 711–729. doi:10.1111/j.1756–8765.2009.01036.x

Tzuriel, D., & George, T. (2009). Improvement of analogical reasoning and academic achievement by the Analogical Reasoning Programme (ARP). *Educational and Child Psychology*, *26*, 71–94.

Warren, R. E., Allen, K. V., Sommerfield, A. J., Deary, I. J., & Frier, B. M. (2004). Acute hypoglycemia impairs nonverbal intelligence: Importance of avoiding ceiling effects in cognitive function testing. *Diabetes Care*, *27*, 1447–1448. doi:10.2337/diacare.27.6.1447

Wertheimer, M. (1900). *Gestalt theory*. Raleigh, NC: Hayes Barton Press.

White, C. S., & Caropreso, E. J. (1989). Training in analogical reasoning processes: Effects on low socioeconomic status preschool children. *The Journal of Educational Research*, *83*, 112–118.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished doctoral dissertation, University of California, Los Angeles.