Investigating a Novel Approach to Assessing Vocabulary Knowledge

Hongyang Zhao[1] & Patricia A. Alexander[1]
[1] College of Education, University of Maryland College Park, USA.

Author Information

Hongyang Zhao, Ph.D.
Correspondence author, first author
School of Education, University of California-Irvine, USA.
Email: zhy02163@gmail.com
Address: 401E Peltason Drive, CA 92617 (primary affiliation)
ORCID: 0000-0003-4711-0314

Patricia A. Alexander, Ph.D.
Second author
College of Education, University of Maryland College Park, USA
Email: palexand@umd.edu
Address: 3304F Benjamin Building, College Park, MD 20742
ORCID: 0000-0001-7060-2582

Author Note

Correspondence concerning this paper should be addressed to Dr. Hongyang Zhao, School of Education, University of California-Irvine, USA. 401E Peltason Drive, CA 92617 (primary affiliation). Phone: 617-893-2064. Email: zhy02163@gmail.com. Secondary affiliation: College of Education, University of Illinois Urbana-Champaign.

Data Availability Statement

The study materials, data, and analytical codes are available by contacting the corresponding author of this study.

Ethics Approval Statement

The current study was reviewed and approved by the Institutional Review Board at the University of Maryland College Park (ID: 1674423).

Funding Information

This work was supported by the SPARC: Supporting Program for Advancing Research and Collaboration from College of Education at the University of Maryland College Park. The writing of this work was also supported by the National Science Foundation under the grant number: 2327398. The opinions expressed are those of the authors and do not necessarily represent the views of the funders.

Permission to Reproduce Materials

The authors have permission from other sources to reproduce the materials used in this study.

Disclosure Statement

The authors report there are no competing interests to declare.

Abstract

This study explored an alternative approach to assessing individuals' word knowledge by gauging the ability to recognize subtle similarities and differences among associated terms. Informed by the theoretical and empirical work on relational reasoning, the Measure of Vocabulary Knowledge through Relational Reasoning (MVKR^2) was developed and validated. Participants were 338 college students who completed the MVKR^2, the Test of Relational Reasoning (TORR), and released items from the SAT Verbal and Math tests. The TORR and SAT tests were administered to examine the convergent and concurrent validities of the MVKR^2. Findings from item confirmatory analyses and correlations demonstrated that the MVKR^2 is a reliable and valid measure of vocabulary knowledge for college-age students. In addition, fluid relational reasoning ability was associated with the performance on this novel measure, but the association with vocabulary knowledge was stronger. When examined on the scale and item levels, the contribution of fluid relational reasoning varied across scales and items within each scale. This study offered an alternative way to examine vocabulary knowledge that has implications for future empirical research and instructional practice.

Keywords: vocabulary knowledge, vocabulary assessment, relational reasoning, college students, measurement

INVESTIGATING A NOVEL APPROACH TO ASSESSING VOCABULARY KNOWLEDGE

“Words are central to listening, speaking, reading, and writing, and are therefore an essential component of almost every aspect of our lives” (Webb & Nation, 2017, pp. 8). The words that individuals know and use expressively or productively are commonly referred to as their *vocabulary* (Butler et al., 2010). Further, how much individuals know about the words that comprise their vocabulary is described as their *vocabulary knowledge* (Anderson & Freebody, 1981). Vocabulary has also been identified as one of five major components of reading by the National Reading Panel (National Institute of Child Health and Human Development, 2000).

Individuals’ vocabulary knowledge is considered a significant predictor of academic outcomes, such as reading comprehension (Muter et al., 2004; Roth et al., 2002; Snow et al., 1995), writing (Olinghouse & Wilson, 2013; Staehr, 2008), mathematics (Akbasli et al., 2016; Carlson et al., 2011), and science (Cohen, 2012; Taboada, 2012). Despite its educational significance, it has been argued that the assessment of vocabulary “is grossly undernourished, both in its theoretical and practical aspects—that it has been driven by tradition, convenience, psychometric standards, and a quest for economy of effort rather than a clear conceptualization of its nature” (Pearson et al., 2007, p. 282). Regrettably, Pearson et al.’s impassioned plea has not resulted in a cache of new measures that afford richer and deeper understandings of individuals’ vocabulary knowledge. In this study, we introduce a novel measure of vocabulary knowledge that was constructed to address the shortcomings that Pearson and others (Pearson et al., 2007) have identified.

A Brief Review of the Current Vocabulary Assessment

Vocabulary assessment is an area of study that has a long history within the domain of literacy (Beck & McKeown, 1985). Since the 1910s, researchers have developed various measures to gauge individuals' knowledge of words. Although existing measures differ along multiple dimensions, our focus in this study pertains to the assessment of words whose meanings must be derived through their relation to other words rather than in isolation. Vocabulary assessments also differ in whether the focus is on common or less common word meanings that are derived from a given context. For the tests that require analysis of multiple words, respondents may be assessed on whether they are able to identify two or more words that are meaningfully associated. For example, respondents may be asked to identify two words that are habitually used together (e.g., *edit* - *text*). As we will discuss, the novel measure we proffer expressly differs even from such multiple-word assessments. Here we briefly overview several commonly used measures that fall into the individual and multiple word categories.

Individual Word Knowledge Assessments

Vocabulary assessments that tap individuals' knowledge of individual words can be differentiated based on the level of context they provide for determining the meaning of target words (Pearson et al., 2007; Read, 2000). Word meanings can be assessed in a completely isolated fashion (decontextualized) in which respondents select a definition, synonym, or perhaps an antonym for a target word. For example, the Peabody Picture Vocabulary Test (Figure 1; Dunn & Dunn, 1997) asks individuals to select a picture that maps onto the meaning of a spoken word. The Iowa Test of Basic Skills (Hoover et al., 2001) requires respondents to select a definition or synonym for a target word (*To sink in the water – play, rest, wash, go down)*. The Vocabulary Level Test (Nation, 1983, 1990) asks examinees to match a subgroup of given words to their definitions (Figure 2a). In the Antonym section of the Woodcock Reading Mastery Test (Woodcock, 1998), respondents are required to read a target word aloud (e.g., *near*) and then provide a word that means the opposite (e.g., *far*).

Another decontextualized way of assessing vocabulary knowledge is to have examinees self-report their perceived level of knowledge. For example, the Eurocentres Vocabulary Size Test (Meara, 1990; Meara & Buxton, 1987) requires examinees to indicate whether they know each given word or not. For this particular measure, a number of nonwords are inserted within the word list as a validity check. Another decontextualized measure that relies on the self-report format is the Vocabulary Knowledge Scale (Figure 2b; Paribakht & Wesche, 1993; Wesche & Paribakht, 1996). Responses at the (a) to (b) level were characterized as *perceived* knowledge, whereas responses at levels (c) to (e) were coded as *demonstrated* knowledge. Self-report measures have the advantage of assessing vocabulary knowledge for a large number of words within a short period of time. However, the shortcoming is that the validity of the assessment results relies on participants' honest answers and careful evaluation of their knowledge for every word presented to them. Although there is the built-in mechanism to prevent cheating or overestimating one's vocabulary size (i.e., inclusion of nonwords), it is impossible for individuals to *demonstrate* their knowledge for every word, which might affect the test validity. Moreover, it is almost impossible to control the easiness for rejection of the nonwords. Some combinations of the base and affix in some nonwords may seem less likely than other, hence making them easier to reject than other nonwords. For L2 learners who have different first languages, it might be easier for some L2 learners to reject certain nonwords than other L2 learners due to the different levels of resemblance of the nonwords with those real words in their first language (Read, 2000).

For context-dependent vocabulary assessments, such as the National Assessment of Educational Progress, the meaning of a target is determined based on the text that is read. Respondents read the text and then select the most appropriate word or definition. Items can vary in their level of context-dependency. For example, in the following two items,

a) The people *consumed* their dinner.
A. ate or drank B. prepared C. brought D. enjoyed

b) The citizens *consumed* their supply of gravel through wanton development.
A. ate or drank B. used up C. spent wastefully D. destroyed

Although textual content is provided for both items, the first one taps into the common meaning of the target word, *consume* (i.e., ate and drank), whereas respondents would need to draw on the context for the second item that targets a less common meaning (i.e., used up).

Multiple Word Assessments

A few existing measures assess word knowledge in association with other related words, including a free association task (Meara, 1984; Schmitt, 1998), the Word Associate Test for English language learners (WAT; Read, 1993, 1998), and a similar test to WAT but for middle school students (Deane et al., 2014). In the free association task (Meara, 1984; Schmitt, 1998), respondents are asked to name the words that first came to their minds upon hearing a given word (e.g., *abandon*). Schmitt (1998a) developed a four-level descriptive system of association behavior in determining whether English learners' word association network was native-like. Three associated words for the target word *abandon* (i.e., *leave*, *desert*, and *alone*) provided by an English language learner were classified as highly native-like if they were also frequently mentioned by native speakers. One shortcoming of the free association task is that it is difficult to assess its validity and reliability, rendering it a less used tool for assessing vocabulary knowledge. Read (1993, 1998) developed the Word Associate Test (WAT) that assessed college

English language learners' knowledge of the semantic network for academic words of high frequency. In the item shown, respondents are asked to select all the words from the eight options that are semantically associated with the target word, *edit*:

arithmetic	film	pole	publishing
revise	risk	surface	text

Each item was developed to have four correct associate words and four unrelated distractors. The correct choices represented one of the three possible relations with the target word: a) *paradigmatic*: the two words are synonyms or share a general meaning, such as *edit – revise*; b) *syntagmatic*: the two words are collocates, which are words often used together, such as *edit – text*; c) *analytic*: the associated word represents one aspect of the frequently-used meaning of the target word, such as *edit – publishing* (Read, 1993).

Similar to the WAT, the Educational Testing Service developed a test for middle school students (Deane et al., 2014). This variation was devised to identify three types of word relations: (a) typical co-occurrence patterns of multiple-word use (similar to syntagmatic relation in WAT; Figure 2c); (b) general word associations to a single topic or concept without necessarily having a deeper conceptual understanding of the target (similar to analytic relation in WAT; Figure 2c); (c) broad or categorical meaning to which the target word belongs (similar to paradigmatic relation in WAT; Figure 2c).

Limitations of the Current Vocabulary Assessments

After reviewing the current vocabulary assessments, one limitation was particularly salient to this investigation. No measure we identified allows researchers to assess the individuals' nuanced understanding of a word's meaning. By *nuanced understanding*, we are referring to respondents' ability to demonstrate broad and fluid knowledge of words that manifests according to distinct types of relations. Decontextualized vocabulary tests assess participants' knowledge of the commonly used meaning of individual target words, without explicit reference to other words. By comparison, context-dependent vocabulary tests, a format that many educational reading assessments adopt, have certain benefits. Specifically, the textual content provided serves to guide respondents' selection of a suitable meaning for that particular context (Pearson et al., 2007). Nonetheless, such context-dependent measures are limited in their ability to capture subtle differences in meaning between the target words and similar words.

The key problem with context-dependent measures is the extent to which items distinguish various levels of vocabulary knowledge is often dependent on the semantic distance between the intended target word and the distractors. If the distractors are semantically distant from the target, the item will not be effective in revealing the nuanced understanding of the target word. However, if the distractors are semantically close to the intended target, they may well become viable options. For example, in the following item, respondents are directed to choose the option closest in meaning to the italicized word.

> In a *democratic* society, we presume that individuals are innocent until and unless proven guilty. *Establishing* guilt is *daunting*. The major question is whether the prosecution can overcome the presumption of reasonable doubt about whether the suspect committed the alleged crime.

establishing

a. attributing
b. monitoring
c. creating
d. absolving

daunting

a. exciting
b. challenging
c. intentional
d. delightful

The options *attributing* and *challenging* have the closest meaning to the target words *establishing* and *daunting* for the given context. On one hand, other synonyms to *establishing* and *daunting*, such as *confirming* or *determining* and *difficult* or *alarming,* may also fit the context, making them eligible candidates for the target words. However, the meaning of these words subtly differs from the intended targets, and these subtle differences cannot be captured by the current assessment format. Thus, respondents' nuanced understanding of the words *establishing* and *daunting* are not thoroughly probed. Yet, the given distractors in the items are quite semantically distant from the target word, making them more readily rejected as plausible answers. Therefore, the popular context-dependent approach to assessing vocabulary knowledge does not seem optimal for capturing the subtlety or depth of individuals' word knowledge.

For the measures that capture aspects of word relations, such as WAT (Read 1993, 1998), there are also limitations to consider. One shortcoming of WAT is that individuals with a cursory understanding of the target word and the various options would likely be able to identify associated terms (e.g., *edit* - *revise*). Moreover, these tests usually incorporate only a few types of relations between words, overlooking other informative relations that could exist, such as opposition or contrast. By involving more words for comparison and by introducing more complex word relations in the measure, researchers should be able to better ascertain individuals' understanding of similarities and differences in word meanings.

Relational Reasoning in Assessing Vocabulary Knowledge

In order to address the limitations in vocabulary knowledge assessments, we developed a measure that should: (a) assess nuanced understanding of vocabulary knowledge from a relational perspective; (b) examine word relations among sets of words instead of word pairs allowing space for assessing more complex word relations; (c) incorporate less typical forms of relations; and (d) cover a wide range of words that vary in frequency and complexity. Specifically, the current assessment embedded vocabulary knowledge assessment within a relational reasoning framework. This novel measure, the *Measure of Vocabulary Knowledge through Relational Reasoning* (MVKR^2), was developed for older adolescents and adults. *Relational reasoning* refers to the intentional, conscious, and effortful identification of meaningful patterns within a stream of information that appears unrelated through the analysis of similarities and dissimilarities (Alexander & DRLRL, 2012a). Relational reasoning is regarded as foundational to complex problem-solving in a variety of fields, including medicine (Dumas et al., 2014), engineering (Dumas & Schmidt, 2015), reading (Hattan, 2019), and mathematics (Zhao et al., 2021). Four forms of relational reasoning have been identified in the literature (Alexander & DRLRL, 2012a): analogy, anomaly, antinomy, and antithesis.

Specifically, *analogical reasoning* involves recognizing similarities among seemingly unrelated objects, ideas, or events (Gentner & Maravilla, 2018). *Anomalous reasoning* requires the detection of a deviation from the general pattern shared by an informational set (Chinn & Brewer, 1993). *Antinomous reasoning* entails recognizing the mutual exclusivity of two sets of entities that form ontological categories (Sorsensen, 2003). *Antithetical reasoning* calls for the discernment of relational opposites along a specified continuum (Alexander & DRLRL, 2012a). Relational reasoning thus provides a framework for assessing fine-grained vocabulary knowledge in accordance with varied associations. By juxtaposing multiple semantically associated words representing different relational forms, researchers should be better positioned to gauge the depth of respondents' vocabulary knowledge.

Further, this relational reasoning framework provides an appropriate tool for uncovering the degree to which individuals can determine the level of semantic similarity or dissimilarity for multiple words based on their core features. A written word usually entails rich semantic information that could be broken down into multiple attributes or features. For example, the word, *peninsula*, stands for an area of *land almost* completely surrounded by *water* except for an *isthmus* connecting it with the *mainland*. One way to reveal a deeper understanding of given words is to present multiple semantically related words to individuals and require them to compare their meanings. This approach would require respondents to determine the features that words share in common (i.e., semantic similarity) and the features upon which they differ (i.e., semantic dissimilarity). For example, *peninsula* vs. *continent*: both are *land areas* and *surrounded by water*, but they differ in precisely how they are surrounded by water.

Also, when individuals try to identify the outlier word that does not fit in a group, they need to recognize the shared feature among most of the words before deciding on the anomalous word that deviates from the others. For example, to identify that *omit* does not belong to the word group, *conflict*, *omit*, *oppose*, and *dispute*, individuals need to analyze the meaning features of the four words and realize that *conflict*, *oppose*, and *dispute* are related because they all share the feature of *disagreement* between two parties, whereas *omit* deviates from them on that feature, meaning to *leave out* or *exclude*.

As these examples illustrate, individuals' knowledge of word relations can be analyzed on a finer level involving specific features that constitute the richness of the word meanings examined relationally. It is reasonable to expect that when engaged in reasoning relationally with multiple words, individuals need to draw upon their semantic network knowledge and constantly analyze the shared meaning features across these words, while at the same time, identifying the distinct feature that makes each word unique. It is through these deliberate comparisons and contrasts that one's understanding of the similarities and differences of a word's meaning in comparison to multiple other words can be revealed.

Purposes of the Study

The purposes of this study were twofold. First, we set out to develop and validate a novel measure that focuses on assessing individuals' vocabulary knowledge from a relational perspective (MVKR^2). Second, we wanted to determine the unique contributions of fluid relational reasoning ability and vocabulary knowledge to college students' performance on the novel measure (Figure 1). Fluid relational reasoning ability is expected to be a contributing factor to performance on MVKR^2, along with respondents' vocabulary knowledge. In effect, individuals' inability to identify the relevant semantic relations in an item could be either due to limits in their relational reasoning capability to abstract and apply the higher-order relations or their vocabulary. Therefore, we wanted to measure the contributions of both fluid relational reasoning and vocabulary knowledge. To achieve these ends, the study was carried out in three phases: Phase I, measure development; Phase II, psychometric validation; and Phase III, examination of underlying contributors. The specific research questions guiding the analysis were as follows:

1. Based on analyses of its structure and content at the test, scale, and item levels, to what extent is the Measure of Vocabulary Knowledge through Relational Reasoning or MVKR^2 a psychometrically sound measure for older adolescents and adults?

Hypothesis: MVKR^2 was conceptualized as an assessment to measure individual differences in subtle vocabulary knowledge. Considering a set of procedures were carefully followed during the development and revision processes of the MVKR^2 (i.e., expert panels, cognitive labs, pilot tests), we hypothesized that the psychometric properties of the MVKR^2 would be supported by the empirical data. Specifically, we hypothesized that the difficulty and discrimination index for most of the MVKR^2 items should fall under an acceptable range as for other norm-referenced assessments (item difficulty: .30 to .80 [Wainer & Thissen, 2001]; item discrimination: equal to or above .10 [University of Washington, 2021]). Reliability and validity based on relations to other variables (i.e., convergent and concurrent validity) should be acceptable as well. Given MVKR^2 was developed following the same theoretical framework of another test of fluid relational reasoning (i.e., Test of Relational Reasoning or TORR [Alexander & DRLRL, 2012b]), TORR provides an appropriate reference point for predicting the reliability and validity of the MVKR^2. Thus, we hypothesized that the internal consistency reliability of the MVKR^2 should be comparable to that of TORR, which was ω = 0.82. The validity of the MVKR^2 based on its relations to vocabulary knowledge and math performance should be close to or beyond r = .60 (for verbal) and r = .36 (for math; Alexander et al., 2016). We hypothesized that the validity of the MVKR^2 based on its relation to TORR should be comparable to the correlation between TORR and another verbal test of relational reasoning (vTORR; Alexander, Singer, et al., 2016), which was r = .52.

2. What are the unique contributions of fluid relational reasoning ability and vocabulary knowledge to college students' performance on the MVKR^2 at the test, scale, and item level, respectively?

Hypothesis: As described in the framing of this investigation, the MVKR^2 was intended to reveal vocabulary knowledge by requiring individuals to compare the meanings of a set of carefully chosen words through relational reasoning. Performance on this novel measure would largely depend on a deep understanding of the meaning for an extensive range of words. At the same time, individuals would need to possess a certain level of fluid relational reasoning ability to navigate through those complex relations among the presented words. Therefore, we hypothesized both vocabulary knowledge and relational reasoning ability would significantly contribute to college students' performance on the MVKR^2 with word knowledge being the driving source of impact. However, whether this pattern would be observed across all the levels of examination (i.e., test, scale, item) remains unclear.

Phase I: Measure Development

Overall Structure

The Measure of Vocabulary Knowledge through Relational Reasoning (MVKR^2) aimed to assess older adolescents' and adults' vocabulary knowledge from a relational perspective by inviting them to reason analogically, anomalously, antinomously, and antithetically with words. Despite the novelty of its focus, the structure of this measure was built on the prior measures of relational reasoning in several important ways (Alexander & DRLRL, 2012b; Alexander et al., 2016). Specifically, we sought to parallel existing measures of relational reasoning (i.e., Verbal Test of Relational Reasoning [Alexander & DRLRL, 2014]; Test of Relational Reasoning [Alexander & DRLRL, 2012b]; Test of Relational Reasoning-Junior [Alexander & DRLRL, 2018]) in terms of the specific numbers of items per scale, reasoning processes, and patterns represented in each scale. This decision was predicated on the fact that this overall structure has proven effective across both figural and linguistic measures created for both younger and older populations (Alexander, Singer, et al., 2016; Dumas & Alexander, 2016).

Thus, the final version of the MVKR^2 consisted of 32 selected-response and constructed-response test items organized in four 8-item scales, each mapping onto one identified form of relational reasoning. Each scale set began with two sample items that familiarized participants with the specific procedure to be followed for that scale. These sample items are not scored, and no explanation or additional feedback was provided beyond the correct answer. Moreover, as we will illustrate, graduated response options (Alexander & Kulikowich, 1991) were generated systematically that reflected gradual deviations from the expected response, so as to assess various levels of understanding of word meaning.

Scale Development

The construct validity of each scale, or domain description inference under Kane's validity framework (Kane, 2013), was determined by the degree to which the item set accurately represented the nature of that identified relational form. In the following sections, we explicate how items and distractors were created for each scale and how each scale assesses subtle vocabulary knowledge uniquely through each form of relational reasoning.

Analogy

The analogy items took the form of classic A:B::C:__ problems (Figure 3a. Goswami, 1992; Klix, 1992; Sternberg et al., 1981). For a scale meant to assess deep vocabulary knowledge by identifying similarities, it is vital to assess respondents' ability to identify precise attributional similarities across two-word pairing in which the second pairing is incomplete. Like a verbal ratio problem, the correct response to the D term would result in a C:D pairing that is as equivalent as possible to what is represented in the A:B relation. The given pair of words, WHISPER and SHOUT, both entail the action of *speaking* but at different *volume* levels. Thus, the two salient attributes to be considered when searching for the correct term are the nature of the utterance and its loudness. A WHIMPER, like WHISPER, is a weak utterance but one associated with pain or discomfort and not communication. The correct response, WAIL also expresses pain or discomfort but at greater volume. The distractors were generated that address only one of the targeted attributes (D and A) or neither (C).

Anomaly

The anomaly items assessed the ability to identify semantic outliers. A sample item is presented in Figure 3b. It is essential for the anomaly scale to assess respondents' ability to extract the core or typical features signified by the given word set (A, B, C, D), and then identify the one word with a meaning that sets it apart from the others in some manner. Key to the anomaly items were groups of words that cohered in some way but that included one term that was aberrant based on an attribute such as direction, magnitude, or intensity. In the sample item, all terms capture some emotional state. However, GRIEVED, REMORSEFUL, and SORROWFUL represent various degrees of sadness, whereas INQUISITIVE conveys a state of being curious, which makes it an outlier in this word group.

Antinomy

The antinomy items measured individuals' ability to recognize ontologically distinct clusters of words. A sample item is presented in Figure 3c. The goal of the antinomy scale is to capture respondents' ability to discern the semantic similarities of the terms in each set of words. For the sample item, the words in Set 1 are related to *birds*, whereas words in Set 2 pertain to *buildings*. With this understanding, respondents search the set of given options for a target word that *only* fits in the bird set. Therefore, the correct answer is B. Three distractors that reflected a patterned deviation from the expected response: a) a word that does not belong to either category (C); b) a word that only fits in the building category (A); c) a word with multiple meanings that fits in both categories (D).

Antithesis

The antithesis items assessed knowledge of words with opposite meanings. A sample item is presented in Figure 3d. It is crucial for the antithesis scale to evaluate respondents' ability to accurately identify the underlying continuum and then find a word that can be appropriately placed on that continuum between the polar terms (Grossnickle et al., 2016). As illustrated in the sample item, the two given words, DIMINUTIVE and COLOSSAL, represented an opposite relation on the continuum of *size*. A scoring key that included all the acceptable answers for this question was determined a priori (Table 1). The scoring key was initially developed by the authors, and then evaluated, revised, and finalized by the expert panel. The response options for antithesis items included a correct response and three distractors that reflected a patterned deviation from the expected response: a) a word that is unrelated to the continuum (B); b) a word that falls out of the continuum set by the two polar words (D); c) a word represents a meaning related to the continuum in some way but not accurately reflecting the continuum (C).

Scoring Protocol

For items in the Analogy, Anomaly, and Antinomy scales, 1 point was awarded for each correct answer. For the Antithesis scale, 0.5 point was awarded if a response provided was identified as an acceptable answer based on the scoring key (Table 1). Another 0.5 point was awarded if participant chose the correct answer for the multiple-choice question. Four total scores were calculated for the four scales as well as one grand total score was calculated summing them up.

Word Selection

We followed the three-word-tier model proposed by Beck and McKeown (1985) when selecting the words to include in the MVKR^2. Tier 1 words represent the most basic words that rarely required instructional attention to their meanings in elementary or middle school, such as *clock*, *baby*, *happy*, and *walk*. Tier 2 words, in contrast, represent high-frequency words that are widely used across a variety of domains in both oral and written language. Some typical Tier 2 words include *coincidence*, *absurd*, *industrious*, *fortunate*, *benevolent*, and *perform*. Words at this level play a significant role in readers' repertoire and have been the primary focus of vocabulary instruction in formal schooling (Beck et al., 2013). Finally, Tier 3 words are those that arise within specific domains, but are rarely used or encountered in every life, such as *isotope*, *lathe*, and *refinery*. Words included in the MVKR^2 were mostly Tier 2 words and a few

Tier 1 and Tier 3 words. Older adolescents and adults are presumed to have some exposure to Tier 2 words, which makes them suitable for assessing vocabulary knowledge for these populations of English-speaking students.

Test Revision

The initial scales for the MVKR^2 each consisted of 16 multiple-choice items, which resulted in a total of 64 items for the entire measure. The number was more than the desired number of items (*n*=40) for the final version of the MVKR^2 in case some of the initial items do not have proper item properties (i.e., item discrimination and difficulty) when tested in the pilot stage, thus should be removed from the final pool. Items were initially developed and then revised iteratively based on the feedback from a group of experts (*n*=8) in relational reasoning and scale development at a large, mid-Atlantic university. The experts provided critique and feedback on the format of the four scales as well as on the item quality in open discussions. Although no quantitative metric was used in panel-based decisions, the revision plan following each critique was evaluated, openly discussed, and reached consensus among all panel members. Then, the initial items were submitted to cognitive labs (*n*=6) and two pilot tests to examine their functionality, and necessary revisions were made accordingly. We made revision decisions based on the problems that we identified during the expert panels, cognitive labs, and pilot tests. For example, we identified that the initial format of the Antithesis scale was not appropriate because test takers could rely on the meaning of the option words to find the correct answer without identifying the underlying continuum represented by the two given opposite words. Full details of the expert panels, cognitive labs, and pilot tests can be found in Supplemental Materials.

Phase II: Measure Validation

Participants

The recruited participants for the validation phase were 338 undergraduate students from 4 universities. Participation was voluntary, and participants could withdraw freely from the study at any time. Participants were offered an opportunity to enter their email addresses into a raffle for one of three $100 Amazon gift cards. Further, at the discretion of the individual course instructor who assisted with advertising this study, students who completed the study might receive 1%-2% of extra credits for research participation in their registered course. The current study was reviewed and approved by the Institutional Review Board (ID: 1674423).

Among the recruited participants, 246 (72.8%) identified as female, 69 (20.4%) as male, 8 (2.4%) reported gender non-binary/non-conforming/self-describe or "prefer not to answer," and 15 students did not report gender. Their age ranged from 18 to 55 years (*M* = 21.90, *SD* = 4.98). In terms of race/ethnicity, 41 (12.1%) identified as African American, 35 (10.4%) as Asian/Asian American, 32 (9.5%) as Hispanic, 180 (53.3%) as White, 4 (1.2%) as Other, and 31 (9.2%) identified as biracial or multiracial. 15 students (4.4%) did not report their race/ethnicity. Regarding class standing, 59 (17.5%) were freshmen, 69 (20.4%) were sophomores, 97 (28.7%) were juniors, 61 (18.0%) were seniors, and 52 (15.3%) had missing values. On the question of whether English is their first language, 282 (83.4%) reported "yes," 41 (12.2%) reported "no," and 15 (4.4%) cases had missing values.

Measures

Measure of Vocabulary Knowledge through Relational Reasoning

The final version of the MVKR^2 consisted of 8 sample items and 32 test items arranged in four scales selected from the initial version. More details of item screening can be found in Supplemental Materials. Following the Standards for Educational and Psychological Testing (American Education Research Association et al., 2014), we aimed to gather empirical evidence for the validity of the MVKR^2 based on its internal structure and relations to other variables (i.e., convergent and concurrent validity).

SAT Verbal

The final version of the SAT Verbal (SATV) measures vocabulary knowledge, which was used to establish convergent validity of the MVKR^2. It consisted of 12 multiple-choice items that were selected from the SAT Verbal – Initial based on the first pilot data. SATV was compiled by previously administered and publicly released SAT sentence completion items. Only the sentence completion items with one target word were included because they measured participants' knowledge of contextual use for individual words, which aligned with the definition of vocabulary knowledge in the current study. Items with multiple target words were excluded because they usually comprised complex syntactic structures and semantic relations, thus participants' performance on those items would involve knowledge or skills other than vocabulary knowledge. See the sample item and item screening in Supplemental Materials.

Test of Relational Reasoning

Test of Relational Reasoning (TORR), assessing fluid relational reasoning, was used as another measure to establish convergent validity for the MVKR^2. TORR is a 32-item fluid measure that takes a visual-spatial format to assess four distinct forms of relational reasoning ability–analogy, anomaly, antinomy, and antithesis. Each scale begins with two sample items intended to familiarize participants with the specific directions and procedures of the eight test items in that scale. For each item, the written directions were presented at the top followed by a figure problem with four response options displayed below. Figure 4a-4d presents one sample item from each scale. The sample items were not scored, and participants were provided with the correct answer option after they selected a response. No additional feedback or explanation was presented for any sample or test item beyond the correct answer option.

The reliability and validity of the TORR were established in Alexander et al. (2016). The internal consistency reliability of TORR as measured by Cronbach's alpha was .84 and .82 at two times. The test-retest reliability of TORR was .71 ($p < .001$). Moreover, TORR was found to be significantly and moderately correlated with Raven's Progressive Matrices (Raven, 1941; $r = .49$, $p < .001$), confirming its convergent validity. The low-moderate correlation of TORR and Shapebuilder (Sprenger et al., 2013), a measure of working memory, at $r = .31$ ($p = .02$), supported its discriminant validity. Further, TORR was found to significantly predict college students' performance on both the verbal and math SAT items ($r = .60$ for verbal and $r = .36$ for math; Alexander et al., 2016), which confirmed its predictive validity.

SAT Math

The final version of the SAT Math (SATM) measuring mathematical performance was used to establish concurrent validity for the MVKR^2. Previous studies found that math performance was closely related with fluid relational reasoning (Zhao et al., 2021) and vocabulary knowledge (Carlson et al., 2011; Dunston & Tyminski, 2013). Thus, it is reasonable to speculate that performance on the MVKR^2, which depend on both fluid relational reasoning and vocabulary knowledge, should be predictive of mathematical performance as assessed by SATM. SATM consisted of 12 selected-response items that were selected from the SAT Math – Initial based on the first pilot data. SATM was compiled by previously administered and publicly released SAT math items. Participants were instructed to work through the given math problems and choose one answer from the five given options. See the sample item and item screening in Supplemental Materials.

Procedures

Participants completed all the measures for the validation phase online via Qualtrics in two sessions. In the first session, they were asked to complete the consent form, demographic measure, the MVKR^2, and the SAT Verbal. In the second session, they were asked to complete the TORR and SAT Math. The procedures followed for this validation phase were the same as those described in the section of Pilot Tests in Supplemental Materials.

Results

Descriptive Statistics

The descriptive statistics, including the means, standard deviations, and data distribution for all the administered measures, are presented in Table 2. The data could be considered normally distributed for the total and scale scores of all the administered measures, given that all their skewness and kurtosis estimates fell within the -2 to 2 range (George & Mallery, 2010). There were no missing values.

We followed the argument-based approach to validation (Kane, 2013) when evaluating the validity of the MVKR^2. Specifically, our analyses focused on scoring, generalization, and extrapolation inferences in Kane's framework. The scoring inference included item analyses (i.e., item difficulty and discrimination) and analyses of the factor structure, which provided backing support for interpreting the scores of MVKR^2 as intended. The generalization inference here referred to the internal consistency reliability that provided warrants for generalizing the interpretations of the MVKR^2 scores over conditions of observation. The extrapolation inference aimed to examine the relations between the scores of the MVKR^2 and scores based on "criterion" assessments, which could also be considered as convergent and concurrent validity assessments.

Scoring Inference

Item Difficulty and Discrimination. The difficulty and discrimination for each item in the MVKR^2 were examined under the classical test theory (Table 3). All items on the MVKR^2 except Antinomy item 4 (.83) fell within the ideal difficulty range of .30 to .80 (Wainer & Thissen, 2001). The most difficult item was Antithesis 4 with a difficulty index of .31. Further, the MVKR^2 items revealed a relatively balanced distribution in terms of difficulty level. Ten items were judged as difficult (.30-.50), 12 as moderately difficult (.50-.70), and 10 as easy items (.70-.83). Thus, the MVKR^2 appeared appropriately challenging for college students.

Next, we examined the discrimination indices for all MVKR^2 items. This index represents the correlation between respondents' performance on a given item and the scale score. A high value indicates that respondents' performance on that item well reflects their overall ability as represented by their performance on the entire scale. The discrimination values of the 32 MVKR^2 items ranged from .26 to .58 with a median of .42. In practice, items with values of .10 or above could be classified as acceptable, .10-.30 as fair, and above .30 as good, in discriminating between high-ability and low-ability test takers (University of Washington, 2021). All items on the MVKR^2 had discrimination values over .10, with 27 items over .30, which means that most MVKR^2 items could well differentiate between those respondents with deeper and shallower vocabulary knowledge.

Factor Structure. The overall factor structure of the MVKR^2 was examined by a series of confirmatory item factor analyses. The overarching goal of the factor analyses was to examine the dimensionality of word relational knowledge assessed by the MVKR^2. In other words, the factor analyses would provide evidence to support whether word relational knowledge, as assessed by the MVKR^2, should be better considered as a unidimensional construct or a multidimensional construct. Consistent with the previous psychometric works on relational reasoning (Alexander, Dumas, et al., 2016; Alexander, Singer, et al., 2016; Dumas & Alexander, 2016; Zhao et al., 2021), we tested a unidimensional model (Model A) and a multidimensional model (Model B). Model A was a one-factor, unidimensional model in which a general latent factor, word relational knowledge, explained the relations among all the 32 items. We further tested a higher-order model (Model B) in which we modeled four latent factors each representing word relational knowledge of a specific higher-order relation as well as an overarching higher-order factor to account for any resulting association among them. If Model B was found to fit the data better than Model A, it would provide one piece of evidence supporting the multidimensional hypothesis.

We further included a bifactor model (Model C) as a source of triangulation to gain validity of the findings. Bifactor models usually include a general factor explaining the shared variances among all the individual items as in unidimensional models (i.e., word relational knowledge). They also include specific grouping factors (i.e., MVKR^2 Analogy, MVKR^2 Anomaly, MVKR^2 Antinomy, MVKR^2 Antithesis), which are usually set to be orthogonal to each other as well as to the general factor, to explain the shared variances among subsets of items over and above the general factor. Bifactor models are commonly used as dimensionality tests in psychometric analyses (Rodriguez et al., 2016; Zhao et al., 2022). If the shared variance explained by the general factor is found to be large enough, the scale can be considered unidimensional (Hoffmann et al., 2021; Rodriguez et al., 2016), even if the bifactor model or other multidimensional models yield acceptable model fit.

The confirmatory item factor analyses were conducted in Mplus 8.0 (Muthén & Muthén, 2017), with each item score declared as binary categorical variable. We fit the three models with the weighted least squares mean and variance adjusted estimator (WLSMV), which is a robust estimator for modeling categorical or ordered data (Brown, 2006; Proitsi et al., 2011). In selecting the best-fitting model, we considered comparative fit index (CFI) ≥ .90, Tucker–Lewis index TLI ≥ .90, root-mean-square error of approximation (RMSEA) ≤ .08, and standardized root-mean-square residual (SRMR) ≤ .10 as acceptable model fit (Hau et al., 2004). The results suggest that Model A fit the data poorly, yet Model B and C both had acceptable fit. Moreover, the explained common variance (ECV) of the general factor in Model C did not exceed the criterion value (0.7), suggesting that the one general factor could not explain enough shared variance among all the MVKR^2 items.

Taken together, these results suggest that word knowledge as measured by MVKR^2 should not be considered simply as a unidimensional construct, but better as a multidimensional construct with four dimensions capturing the knowledge of specific word relations and one overarching dimension of word knowledge accounting for the association among these dimensions, which aligns with the higher-order model (Model B). Details of model specification and comparisons can be found in Supplemental Materials. An examination of dimensionality based on the tested bifactor model was also included in Supplemental Materials (i.e., calculation of the explained common variance).

Generalization Inference

The internal consistency reliability of the four scale factors of the MVKR^2 were represented by the McDonald's ω (McDonald, 1985, 1999). The higher-order model (Model B) was selected as the measurement model for calculating omega for the four scale factors. The omega coefficients for the Analogy, Anomaly, Antinomy, and Antithesis scales were ω=.91, .92, .94, and .93, respectively. We also calculated the internal consistent reliability for the entire test based on the bifactor model (Model C). Omega hierarchical (McDonald, 1985, 1999), which can parse out the variability attributable to subfactors and calculates reliability for a general factor that applies to all items, and is thus recommended for bifactor models (Cho, 2022; McNeish, 2018). Omega hierarchical for the general factor based on Model C was 0.87.

Extrapolation Inference

Convergent and concurrent validity of MVKR^2 were examined through correlations. Correlations indicated that MVKR^2 had moderate to strong correlations with all the other three measures ($.42 \leq r \leq .70$, ps<.001; Table 4). Thus, performance on the MVKR^2 converged with performance on assessments of fluid relational reasoning (i.e., TORR) and vocabulary knowledge (i.e., SATV), and also predicted performance on a distal yet important academic domain (i.e., math performance). Take together, these results provided empirical evidence to the criterion-based validity of the MVKR^2.

Phase III: Unique Contributions of Fluid Relational Reasoning and Vocabulary Knowledge

The second research question guiding this study asked about the unique contributions of fluid relational reasoning ability and vocabulary knowledge to participants' performance on the MVKR^2. To address this question, we examined those contributions of vocabulary knowledge and fluid relational reasoning ability at the test, scale, and item levels.

Test-Level Examination

An SEM analysis using Mplus 8.0 indicated an acceptable fit for the model that examined the contributions of fluid relational reasoning ability and vocabulary knowledge to word relational knowledge on the test level (Figure 6; χ^2 = 2140.44, *df* = 2005, *p* = .018; *RMSEA* = .01, *RMSEA 90% C.I.* = [.007, .019]; *CFI* = .96; *TLI* = .95; *SRMR* = .09). Latent word relational knowledge was modeled by the MVKR^2 items following the higher-order factor structure (Model B). Similarly, fluid relational reasoning ability was modeled by the TORR items following a higher-order factor structure (Alexander, Dumas, et al., 2016). The word relational knowledge factor was further regressed on fluid relational reasoning and SAT Verbal total score. Vocabulary knowledge and fluid relational reasoning uniquely and significantly contributed to word relational knowledge (path coefficients= .61, .36, *p*s<.001). This implies that 37.21% and 12.96% of the variance in word relational knowledge is explained by vocabulary knowledge and fluid relational reasoning ability, respectively. The test-level examination confirmed that performance on the MVKR^2 depends on vocabulary knowledge and fluid relational reasoning with vocabulary knowledge emerging as the stronger predictor for college students.

Scale-Level Examination

The scale-level examination aimed to assess the unique contributions of latent fluid relational reasoning and latent vocabulary knowledge to each scale of MVKR^2. A structural equation modeling (SEM) analysis in Mplus 8.0 with maximum likelihood estimation was performed, loading TORR scale scores on the relational reasoning factor only, SAT verbal total score on the vocabulary knowledge factor only, and MVKR^2 scale scores on both factors (Figure 7). Model fit indices (χ^2 = 51.65, *df* = 24, *p* = .001; *RMSEA* = .06, *RMSEA 90% C.I.* = [.036, .080]; *CFI* = .97, *TLI* = .95, *SRMR* = .06) indicated an acceptable fit. All TORR scales loaded significantly on the relational reasoning factor, as did the SAT Verbal total score on the vocabulary knowledge factor. MVKR^2 scales showed significant loadings on the vocabulary knowledge factor, with Analogy and Antinomy also loading on the relational reasoning factor. Variances explained by relational reasoning for the Analogy, Anomaly, Antinomy, Antithesis scales of the MVKR^2 were 1.74%, 0.48%, 6.40%, and 1.66%, whereas vocabulary knowledge explained 36.70%, 48.70%, 19.98%, and 33.99% of variances. The correlation between relational reasoning and vocabulary knowledge factors was moderate ($r = .44$, $p < .001$). The results suggest that individual differences in vocabulary knowledge were still the driving force in explaining the variability of performance on the MVKR^2 compared to fluid relational reasoning, yet relational reasoning did not contribute to the performance on the Anomaly and Antithesis scales MVKR^2 to the same extent as the Analogy and Antinomy scales.

Item-Level Examination

The item-level examination aimed to discern the contributions of fluid relational reasoning and vocabulary knowledge on MVKR^2 item performance. Using four two-dimensional IRT models, each representing a form of relational reasoning, data were analyzed with the WLSMV estimator in Mplus 8.0. Items from Analogy, Anomaly, and Antithesis scales were loaded on correlated relational reasoning and vocabulary knowledge factors, while for the Antinomy scale, the correlation between these two factors were fixed at a value of 0.5 due to convergence issues. Fit indices confirmed the appropriateness of the models. Most TORR and all SAT Verbal items loaded exclusively on fluid relational reasoning and vocabulary knowledge factors, respectively. The standardized item-factor loadings indicated that vocabulary knowledge predominantly contributed to MVKR^2 performance, with all items in Analogy, Anomaly, and Antinomy scales significantly loaded on the vocabulary knowledge factor. In contrast, fluid relational reasoning only significantly contributed to the Antithesis scale. See details of model specification and factor loading details in Supplemental Materials. Despite earlier findings at test- and scale-levels supporting the importance of relational reasoning, the item-level examination revealed that relational reasoning played a significant role only in the Antithesis scale, potentially influenced by its distinct configuration in that it is the only scale that explicitly asked participants first to identify the continuum represented by two opposite words presented before selecting the correct response in multiple-choice items. It is possible that the added construction component to the antithesis items increased the amount of relational reasoning required for this set of items over those in the other scales. These tentative explanations warrant further research.

Conclusions and Implications

Before revisiting the findings of the current study, there are particular limitations to the present investigation that must be acknowledged.

Limitations

Restricted by available resources in recruiting thousands of participants, this study collected a sample of 338 participants. Although this sample size allowed for valid statistical inferences for most of the analyses reported in this manuscript, the relatively small sample size could have affected analyses and findings in two ways. For one, we were not able to calibrate the items in the MVKR^2 with accurate estimates for item discrimination, difficulty, and guessing parameters under a multidimensional IRT (MIRT) model. Although confirmatory item factor analyses were conducted, they should be considered as assessing the overall factor structure of the MVKR^2 rather than obtaining accurate item parameter estimations for calibration purpose.

Contributions and Implications

In their influential work, Nagy and Scott (2000) emphasized the interrelated nature of vocabulary, highlighting the limitations of existing assessments in capturing the nuances of individuals' knowledge of word relations. This study introduces a novel perspective by assessing vocabulary knowledge through relational reasoning, moving beyond isolated or context-dependent evaluations. Informed by a theoretical model of relational reasoning (Alexander &

DRLRL, 2012b), the Measure of Vocabulary Knowledge through Relational Reasoning (MVKR^2) explores semantic relations among words within four higher-order relations.

A key innovation lies in the application of the relational reasoning framework to assessing nuanced word meanings. Unlike traditional assessments, MVKR^2 evaluates participants' understanding of semantic relations among words within complex networks. This unique approach uncovers in-depth knowledge by requiring individuals to compare meanings and analyze semantic relationships, reflecting the integration of their semantic network. MVKR^2 emerged as a reliable measure with supported validity. Confirmatory item factor analysis aligned with the relational reasoning conceptual framework. Empirical evidence supported concurrent and convergent validity. Difficulty and discrimination analyses confirmed appropriate item difficulty and discrimination, enhancing the measure's robustness.

Examining the relative contributions of fluid relational reasoning and vocabulary knowledge to MVKR^2 performance, the study found that both factors operated in concert, but vocabulary knowledge generally played a more significant role. Scale- and item-level examinations revealed variations in the contribution of relational reasoning across scales. The study's efforts to parse out the contributions of fluid relational reasoning and vocabulary knowledge at the item level offer a modeling approach with potential implications for scoring comprehensive measures like MVKR^2 that assess more than one construct. This approach could provide researchers with two sets of scores reflecting students' fluid relational reasoning ability and vocabulary knowledge on the latent level through one administration of MVKR^2, provided that item parameter estimates are obtained from calibration efforts based on the item-level model as forwarded by the current study.

Looking ahead, future studies should investigate MVKR^2's predictive effect for linguistic tasks (e.g., reading comprehension) or for achievement in language-rich domains (e.g., reading, writing, literary studies, or history). It is also critical to explore its effectiveness across diverse demographic groups and assess its equity through measurement invariance analyses. Lastly, it seems worthwhile to explore whether the relational reasoning framework would also be effective in *teaching* new words to adolescents. If the relational reasoning framework is as effective in vocabulary instruction as in the vocabulary assessment, teachers would observe an increase in students' vocabulary learning outcome by juxtaposing the target word with some familiar words and highlighting the similarity and difference in the word meanings based on the four forms of relational reasoning.

The current investigation represents an initial endeavor to apply the relational reasoning framework to the vocabulary knowledge domain. Consistent with previous research on relational reasoning, the theoretical framework of relational reasoning possesses great potential in guiding and contributing to learning in various domains. Future research should continue harnessing its potential and applying it more broadly to further our knowledge of human learning.

Disclosure Statement

The authors report there are no competing interests to declare.

References

Akbasli, S., Sahin, M., & Yaykiran, Z. (2016). The effect of reading comprehension on the performance in science and mathematics. *Journal of Education and Practice, 7*(16), 108-121.

Alexander, P. A., & the Disciplined Reading and Learning Research Laboratory. (2012a). Reading into the future: Competence for the 21st century. *Educational Psychologist, 47*(4), 259-280. https://doi.org/10.1080/00461520.2012.722511

Alexander, P. A., & the Disciplined Reading and Learning Research Laboratory. (2012b). *Test of Relational Reasoning*. College Park, MD: University of Maryland.

Alexander, P. A., & the Disciplined Reading and Learning Research Laboratory. (2014). *Verbal Test of Relational Reasoning*. College Park, MD: University of Maryland.

Alexander, P. A., & the Disciplined Reading and Learning Research Laboratory. (2018). *Test of Relational Reasoning-Junior* (copyright pending). College Park, MD: University of Maryland.

Alexander, P. A., Dumas, D., Grossnickle, E. M., List, A., & Firetto, C. M. (2016). Measuring relational reasoning. *Journal of Experimental Education, 84*(1), 119-151. https://doi.org/10.1080/00220973.2014.963216

Alexander, P. A., & Kulikowich, J. M. (1991). Domain knowledge and analogic reasoning ability as predictors of expository text comprehension. *Journal of Reading Behavior, 23*(2), 165-190. https://doi.org/10.1080/10862969109547735

Alexander, P. A., Singer, L. M., Jablansky, S., & Hattan, C. (2016). Relational reasoning in word and in figure. *Journal of Educational Psychology, 108*(8), 1140-1152. https://doi.org/10.1037/edu0000110

Anderson, R., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews* (pp. 77-117). Neward, DE: International Reading Association.

Beck, I. L., & McKeown, M. G. (1985). Teaching vocabulary: Making the instruction fit the goal. *Educational Perspectives, 23*(1), 11-15.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guildford.

Butler, S., Urrutia, K., Buenger, A., Gonzalez, N., Hunt, M., & Eisenhart, C. (2010). A review of the current research on vocabulary instruction. *National Reading Technical Assistance Center, RMC Research Corporation, 1*.

Carlson, E., Jenkins, F., Bitterman, A., & Keller, B. (2011). A longitudinal view of the receptive vocabulary and math achievement of young children with disabilities. NCSER 2011-3006. *National Center for Special Education Research*.

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), 1-49. https://doi.org/10.3102%2F00346543063001001

Cho, E. (2022). Reliability and omega hierarchical in multidimensional data: A comparison of various estimators. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000525

Cohen, M. T. (2012). The importance of vocabulary for science learning. *Kappa Delta Pi Record, 48*(2), 72-77.

Deane, P., Lawless, R. R., Li, C., Sabatini, J., Bejar, I. I., & O'Reilly, T. (2014). Creating vocabulary item types that measure students' depth of semantic knowledge. *ETS Research Report Series, 2014*(1), 1-19. https://doi.org/10.1002/ets2.12001

Dumas, D., & Alexander, P. A. (2016). Calibration of the Test of Relational Reasoning. *Psychological Assessment, 28*(10), 1303-1318. https://psycnet.apa.org/doi/10.1037/pas0000267

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, 106*(4), 1021-1035. https://doi.org/10.1037/a0036777.

Dumas, D., & Schmidt, L. (2015). Relational reasoning as predictor for engineering ideation success using TRIZ. *Journal of Engineering Design*, *26*(1-3), 74-88. https://doi.org/10.1080/09544828.2015.1020287

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test*: American Guidance Service.

Dunston, P. J., & Tyminski, A. M. (2013). What's the Big Deal about Vocabulary?. *MatheMatics Teaching in the Middle School, 19*(1), 38-45.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Gentner, D. & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball & V. A. Thompson (eds.) *International Handbook of Thinking & Reasoning* (pp. 186-203). Psychology Press.

George, D., & Mallery, P. (2010). *SPSS for Windows step by step. A simple study guide and reference*. Pearson Education, Inc.

Goswami, U. (1992). *Essays in developmental psychology. Analogical reasoning in children*. Lawrence Erlbaum Associates, Inc.

Grossnickle, E. M., Dumas, D., Alexander, P. A., & Baggetta, P. (2016). Individual differences in the process of relational reasoning. *Learning and Instruction, 42*, 141-159. https://doi.org/10.1016/j.learninstruc.2016.01.013

Hattan, C. (2019). Prompting rural students' use of background knowledge and experience to support comprehension of unfamiliar content. *Reading Research Quarterly, 54*(4), 451-455. https://doi.org/10.1002/rrq.270

Hoffmann, M. S., Moore, T. M., Axelrud, L. K., Tottenham, N., Zuo, X.-N., Rohde, L. A., Milham, M. P., Satterthwaite, T. D., & Abrahão Salum, G. (2021). Reliability and validity of Bifactor models of dimensional psychopathology in youth from three continents. medRxiv. https://doi.org/10.1101/2021.06.27.21259601.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). Iowa tests of basic skills (ITBS) forms A, B, and C. *Rolling Meadows, IL: Riverside Publishing Company*.

Hau, K.-T., Wen, Z.-L., and Cheng, Z.-J. (2004). *Structural equation modeling (SEM) and its application.* Beijing: Educational Science Publishing House.

Kane, M. (2013). The argument-based approach to validation. *School Psychology Review, 42*(4), 448-457.

Klix, F. (1992). Higher order learning mechanisms in knowledge domain. *Zeitschrift für Psychologie mit Zeitschrift für angewandte Psychologie, 200*(2), 91–103.

McDonald, R. P. (1985). *Factor analysis and related methods*. Lawrence Erlbaum

McDonald R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412-433.

Meara, P. (1984). The study of lexis in interlanguage. *Interlanguage*, 225-235.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. *Foreign Language Comprehension and Production*, 103-113.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing, 4*(2), 142-154.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Developmental Psychology, 40*(5), 665–681. https://doi.org/10.1037/0012-1649.40.5.665.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. https://doi.org/10.1207/S15328007SEM0904_8

Muthén, B., & Muthén, L. (2017). Mplus. In van der Linden (Ed), *Handbook of Item Response Theory* (pp. 548-560). Chapman and Hall/CRC.

Nagy, W., & Scott, J., (2000). Vocabulary processes. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research*, Vol. 3 (p. 269–284). Erlbaum Associates.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines, 5*, 12-25.

Nation, I. S. P. (1990). *Teaching and Learning Vocabulary*. Newbury House.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific* *research literature on reading and its implications for reading instruction* (NIH publication No. 00-4769). U.S. Government Printing Office.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*(1), 45-65.

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, 09-29. https://doi.org/10.18806/tesl.v11i1.623

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*(2), 282-296. https://doi:10.1598/RRQ.42.2.4

Proitsi, P., Hamilton, G., Tsolaki, M., Lupton, M., Daniilidou, M., Hollingworth, P., ... & Powell, J. F. (2011). A multiple indicators multiple causes (MIMIC) model of behavioural and psychological symptoms in dementia (BPSD). *Neurobiology of Aging, 32*(3), 434-442.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing, 10*(3), 355-371. https://doi.org/10.1177%2F026553229301000308

Read, J. (1998). *Validating a test to measure depth of vocabulary knowledge.* In A. Kunnan (Ed.), Validation in Language Assessment (pp. 41-60). Erlbaum.

Read, J. A. (2000). *Assessing vocabulary*. Cambridge University Press.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *The Journal of Educational Research, 95*(5), 259-272. https://doi.org/10.1080/00220670209596600

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning, 48*(2), 281-317. https://doi.org/10.1111/1467-9922.00042

Schmitt, N. (1998). Quantifying word association responses: What is native-like? *System, 26*(3), 389-401. https://doi.org/10.1016/S0346-251X(98)00019-0

Snow, C. E., Tabors, P. O., Nicholson, P. A., & Kurland, B. F. (1995). SHELL: Oral language and early literacy skills in kindergarten and first-grade children. *Journal of Research in Childhood education, 10*(1), 37-48.

Sorensen, R. (2003). *A Brief History of the Paradox: Philosophy and the Labyrinths of the Mind*. Oxford University Press.

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*(5), 638-663.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36*(2), 139-152.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. *Journal of Personality and Social Psychology, 41*(1), 37–55. https://doi.org/10.1037/0022-3514.41.1.37

Taboada, A. (2012). Relationships of general vocabulary, science vocabulary, and student questioning with science comprehension in students with varying levels of English proficiency. *Instructional Science, 40*(6), 901-923.

University of Washington. (2021). Understanding Item Analyses. https://www.washington.edu/assessment/scanning-scoring__trashed/scoring/reports/item-analysis/

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (p. 23–72). Lawrence Erlbaum Associates Publishers.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*(1), 13-40. https://doi.org/10.3138/cmlr.53.1.13

Zhang, Y., Hedo, R., Rivera, A., Rull, R., Richardson, S., & Tu, X. M. (2019). Post hoc power analysis: is it an informative and meaningful analysis? *General psychiatry*, *32*(4), 1-4. 10.1136/gpsych-2019-100069

Zhao, H., Alexander, P. A., & Sun, Y. (2021). Relational reasoning's contributions to mathematical thinking and performance in Chinese elementary and middle-school students. *Journal of Educational Psychology, 113*(2), 279-303. https://psycnet.apa.org/doi/10.1037/edu0000595

Zhao, H., Li, X., & Chen, S. (2022). Development of a brief therapist presence inventory in China using multilevel factor analysis and item response theory. *Psychotherapy Research, 33*(4), 508–523. https://doi.org/10.1080/10503307.2022.2143301

Table 1

Scoring Key for the Constructed-Response Items of the Antithesis Scale

Item No.	Acceptable Responses
Samples	
1	temperature
2	closeness
Items	
1	size, magnitude
2	retention, possession, keeping
3	light*, brightness, luminosity
4	decision making, problem solving, practicality, how something is realistic/idealistic
5	flow*, movement, water/liquid speed
6	active*, energy level, liveliness
7	anger, frustration, upset, irritation, annoyance
8	mood, feeling, emotional state, happiness
9	price, cost
10	stability, volatility, change*
11	pitch, tone
12	bravery, courage, fear
13	expertise, competence, ability, skill, performance, experience, knowledge, capability
14	order*, peace*, serenity

Note. * represents the variants of the given word.

Table 2

Descriptive Statistics for the MVKR^2, TORR, SAT Verbal, SAT Math Scores

	Descriptive Statistics						
Variable	***n***	***Min***	***Max***	***M***	***SD***	**Skewness**	**Kurtosis**
MVKR^2 Total	323	4	31.5	20.75	5.85	-.62	-.15
MVKR^2_AG	322	1	8	5.25	1.75	-.41	-.54
MVKR^2_AM	319	0	8	4.47	1.95	-.20	-.72
MVKR^2_AN	319	1	8	5.95	1.84	-.88	.10
MVKR^2_AT	322	0	8	5.24	1.76	-.84	.16
TORR Total	278	3	30	14.09	5.78	.55	-.27
TORR AG	278	0	8	3.09	2.06	.57	-.49
TORR AM	278	0	8	3.39	1.90	.30	-.76
TORR AN	278	0	8	4.08	1.94	.08	-.86
TORR AT	278	0	8	3.53	1.99	.24	-.71
SAT Verbal Total	312	0	12	6.74	2.94	-.18	-.75
SAT Math Total	320	0	12	6.91	3.39	-.22	-1.15

Note. MVKR^2 = Measure of Vocabulary Knowledge through Relational Reasoning; TORR = Test of Relational Reasoning; AG = Analogy; AM = Anomaly; AN = Antinomy; AT = Antithesis.

Table 3

Classical Test Theory Difficulty (% Correct) and Discrimination (Item-Total Correlation) Estimates for the MVKR^2 Items

MVKR^2 Scale	Item Number	Item Difficulty	Item Discrimination
Analogy	1	0.71	.42**
	2	0.64	.42**
	3	0.63	.30**
	4	0.77	.39**
	5	0.63	.47**
	6	0.68	.31**
	7	0.36	.42**
	8	0.58	.30**
Anomaly	9	0.65	.52**
	10	0.72	.46**
	11	0.36	.32**
	12	0.44	.29**
	13	0.61	.39**
	14	0.54	.45**
	15	0.54	.31**
	16	0.37	.47**
Antinomy	17	0.72	.48**
	18	0.71	.36**
	19	0.69	.26**
	20	0.83	.46**
	21	0.74	.54**
	22	0.74	.41**
	23	0.76	.32**
	24	0.43	.43**
Antithesis	25	0.70	.41**
	26	0.36	.55**
	27	0.52	.53**
	28	0.32	.32**
	29	0.46	.58**
	30	0.33	.26**
	31	0.41	.52**
	32	0.63	.52**

Note. MVKR^2 = Measure of Vocabulary Knowledge through Relational Reasoning. **p<.01.

Table 4

Correlations among Total and Scale Scores of MVKR^2, TORR, SAT Verbal, and SAT Math

Variables	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1. MVKR^2 Total	--										
2. MVKR^2 AG	.79***										
3. MVKR^2 AM	.78***	.60***									
4. MVKR^2 AN	.73***	.41***	.39***								
5. MVKR^2 AT	.79***	.49***	.46***	.49***							
6.TORR Total	.42***	.32***	.36***	.38***	.33***						
7.TORR AG	.30***	.22***	.24***	.25***	.27***	.77***					
8.TORR AM	.36***	.27***	.27***	.36***	.27***	.77***	.52***				
9.TORR AN	.26***	.22**	.24***	.20**	.18*	.70***	.35***	.40***			
10.TORR AT	.33***	.23***	.31***	.31***	.24***	.69***	.36***	.36***	.30***		
11.SAT Verbal Total	.70***	.56***	.61***	.49***	.54***	.34***	.27***	.25***	.25***	.24***	
12.SAT Math Total	.46***	.35***	.35***	.38***	.38***	.61***	.51***	.47***	.38***	.40***	.43***

Note. MVKR^2 = Measure of Vocabulary Knowledge through Relational Reasoning; TORR = Test of Relational Reasoning; AG = Analogy; AM = Anomaly; AN = Antinomy; AT = Antithesis. *p<.05, **p<.01, ***p<.001.

Figures

Figure 1

Sample Item from the Peabody Picture Vocabulary Test (PPVT) (citrus).

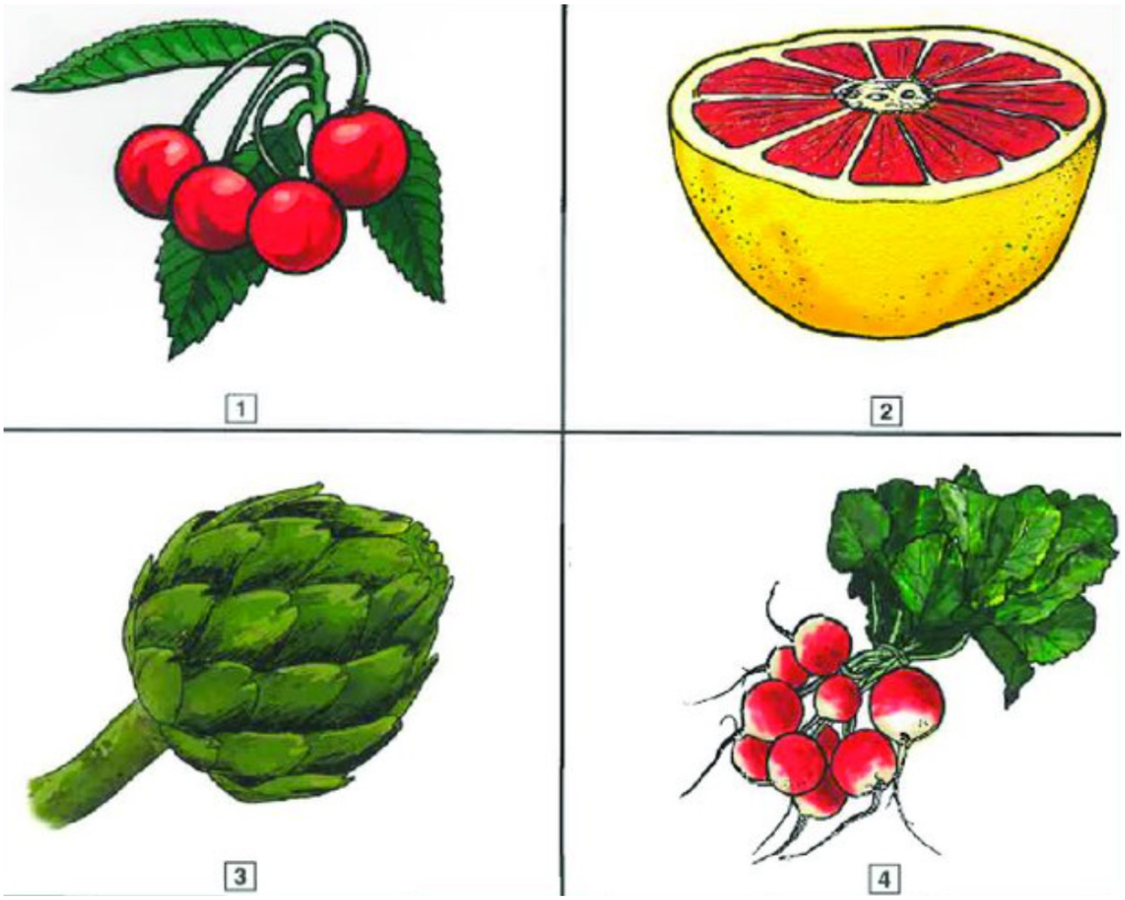


Figure 2

Sample Item from the Current Commonly Used Vocabulary Assessments.

(a)

a. apply	
b. elect	__ choose by voting
c. jump	__ become like water
d. manufacture	__ make
e. melt	

(b)

For this instrument, an examinee indicates whether:

(a) *I have never seen this word*;
(b) *I have seen this word before, but I don't know what it means*;
(c) *I have seen this word before, and I think it means____________ (synonym or translation)*;
(d) *I know this word. It means_____________ (synonym or translation)*;
(e) *I can use this word in a sentence*: _______________.

(c)

Stephen agreed to undertake the _________.
a. purpose
b. task
c. question

launch, conduct, complete:
a. relieve
b. reject
c. undertake

To undertake something is to __________ it.
a. begin
b. continue
c. notice

Note. (a) Vocabulary Level Test; (b) Vocabulary Knowledge Scale; (c) Measure of Students' Depth of Semantic Knowledge by the Educational Testing Service.

Figure 3

Sample Item from the Four Scales of the Measure of Vocabulary Knowledge through Relational Reasoning.

(a) **Directions**: This is a pattern that is not yet complete. Find the missing term from those below that completes the pattern.

WHISPER : SHOUT ::
WHIMPER : ?

A. ORATE
B. WAIL*
C. MUMBLE
D. WEEP

(b) **Directions**: In the item below, three of the four words follow a particular pattern in terms of their meaning. Select the word that ***does not follow the same pattern*** in meaning as the other three.

A. GRIEVED
B. INQUISITIVE*
C. REMORSEFUL
D. SORROWFUL

(c) **Directions**: In the item below, there are ***two distinct sets*** of words (Set 1 and Set 2). Within each set, the words fit together according to their meaning. Select one of the four given words that fits ***only*** in Set 1, but ***not*** Set 2.

Set 1	Set 2
TAIL	BASEMENT
BEAK	HALL
PECK	BUILD
FLY	RENOVATE

A. DOME
B. CLAW*
C. PAVE
D. WING

(d) **Directions**: The words shown below mark the opposite ends of a continuum.

←→
DIMINUTIVE COLOSSAL

In the box below, identify what that ***continuum*** represents.

*size

For this item, select the word with ***the meaning that fits somewhere between*** the two opposite terms.

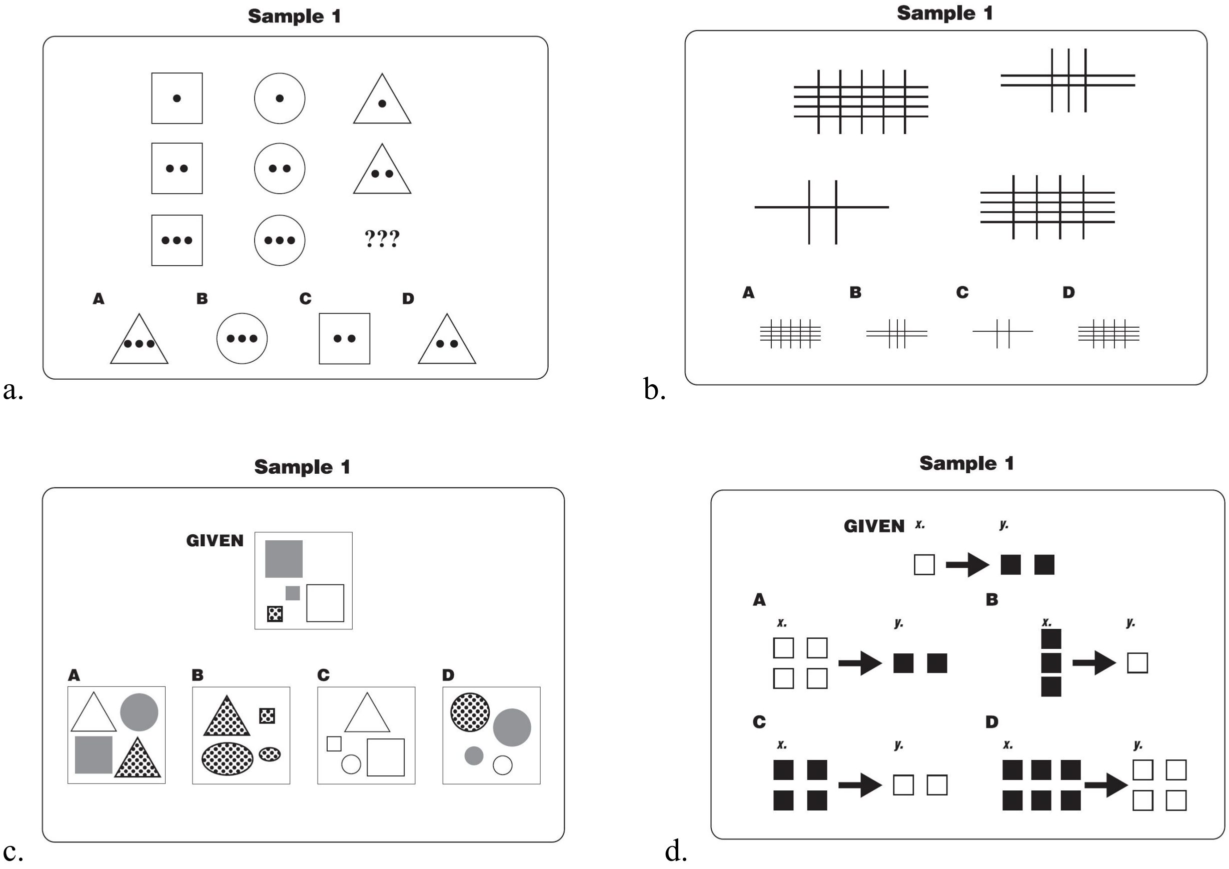
←→
DIMINUTIVE COLOSSAL

A. PETITE*
B. HEAVY
C. BONY
D. MICROSCOPIC

Note. *the correct/acceptable response; (a) Analogy; (b) Anomaly; (c) Antinomy; (d) Antithesis.

Figure 4

Sample Items from the TORR for the (a) Analogy, (b) Anomaly, (c) Antinomy, and (d) Antithesis Scales



a.

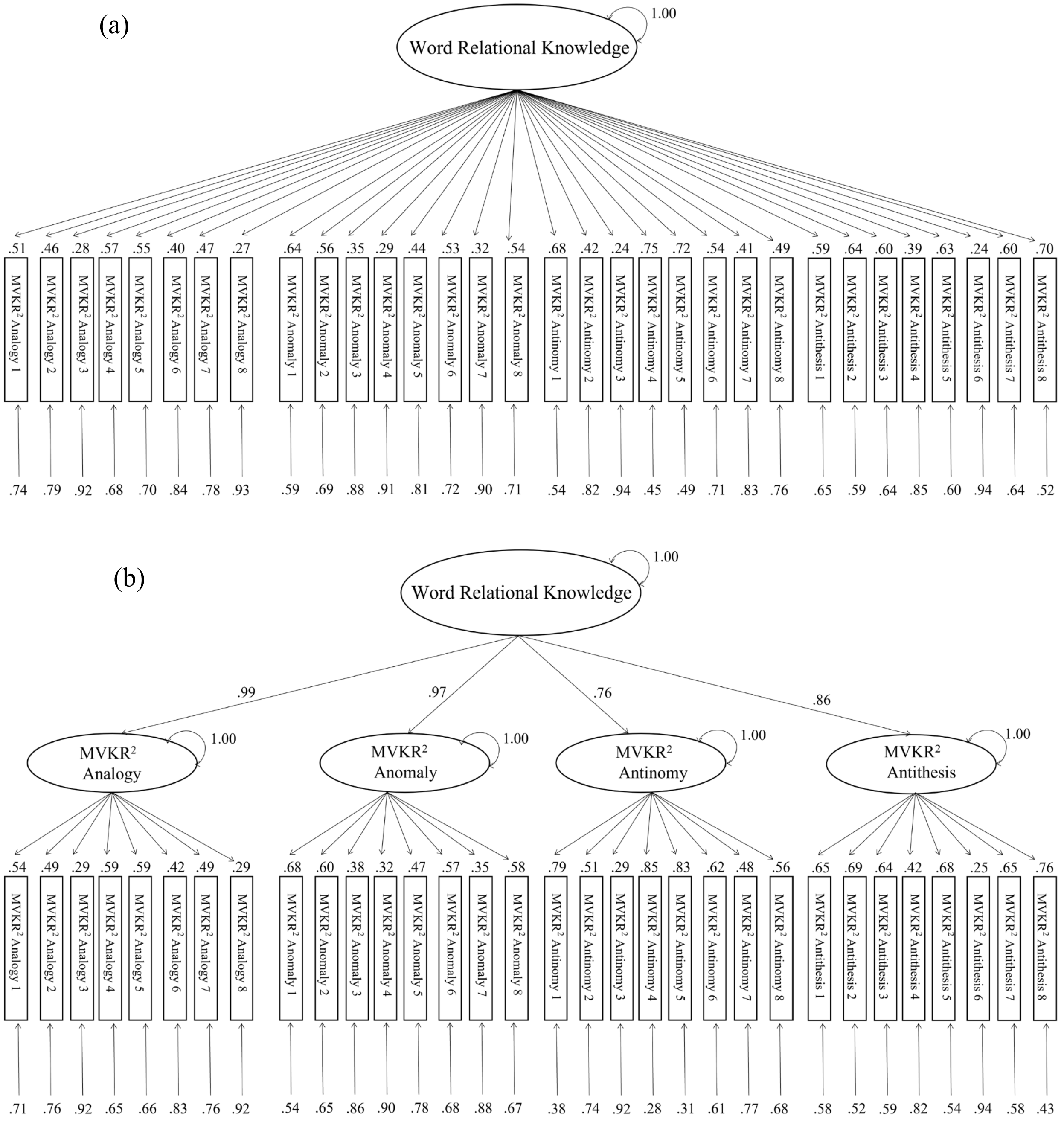
b.

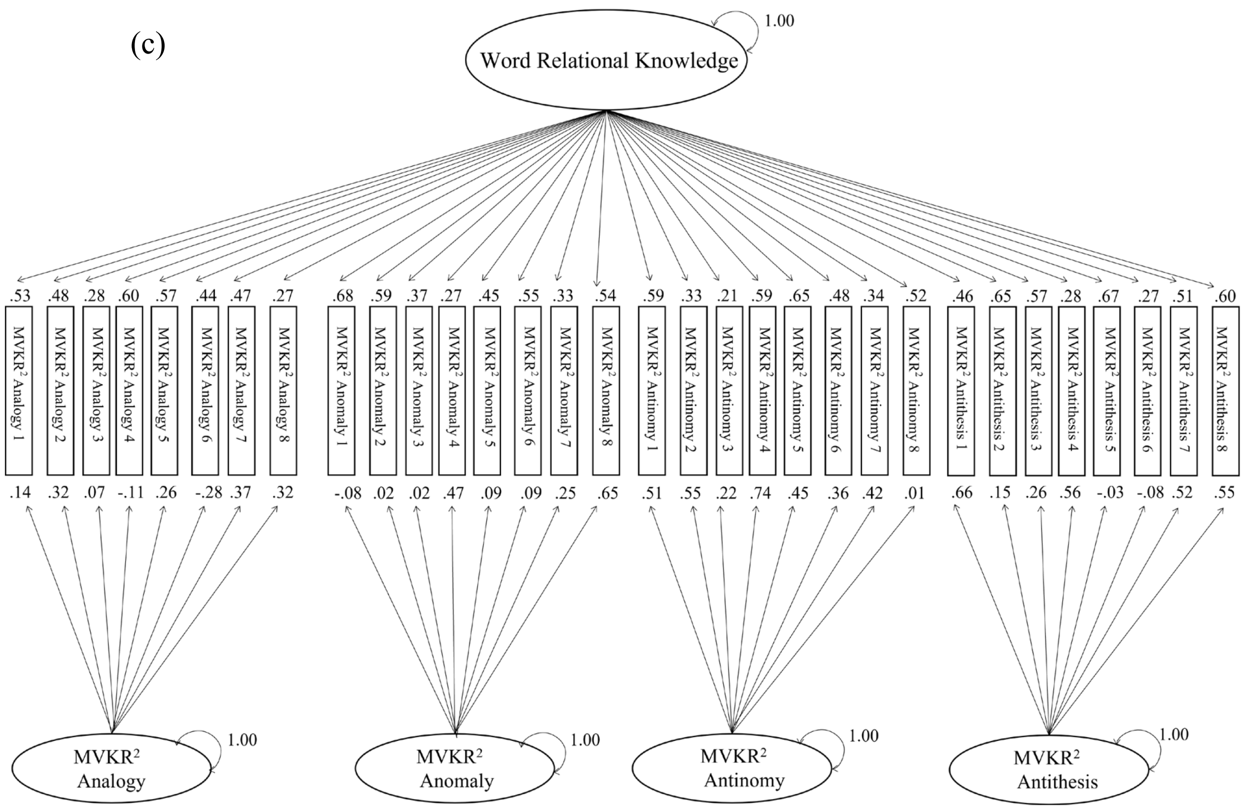
c.

d.

Figure 5

Three Theoretically Viable Models Tested by Means of Confirmatory Item Factor Analyses

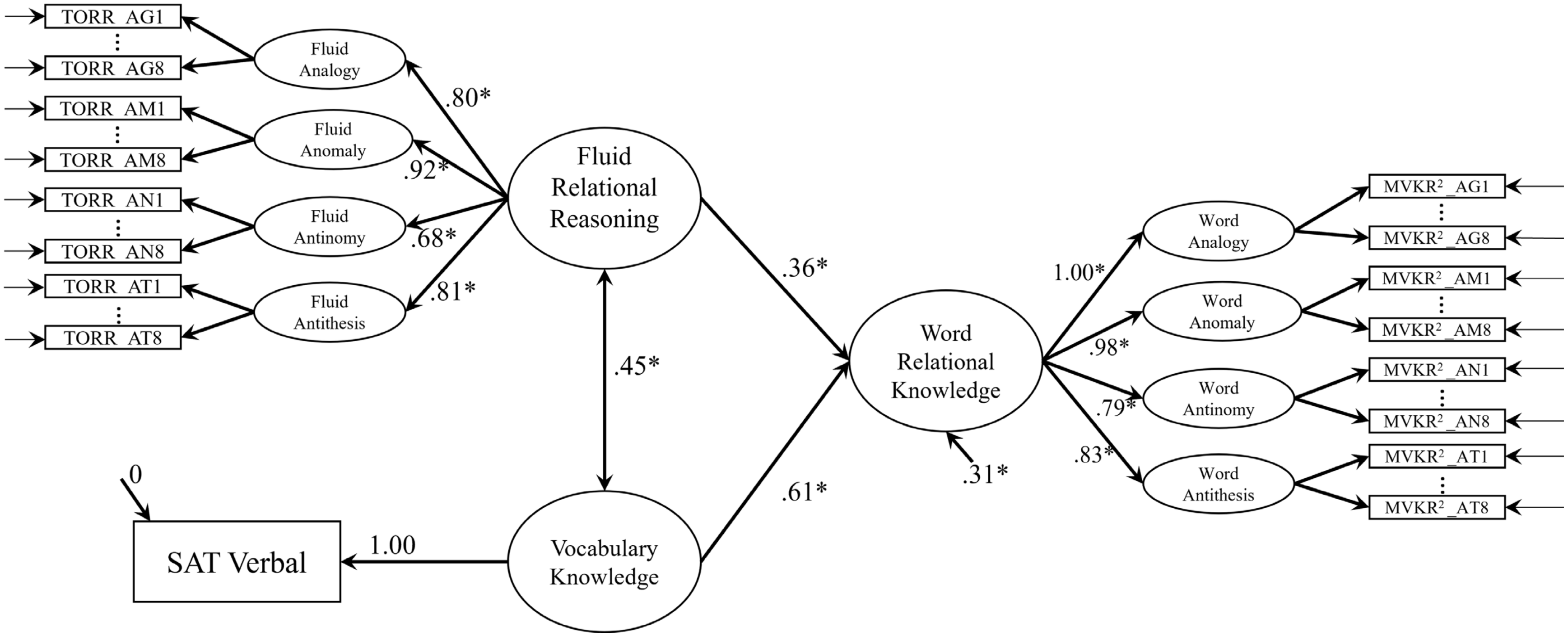




Note. (a) Unidimensional model; (b) Higher-order model; (c) Bifactor model.

Figure 6

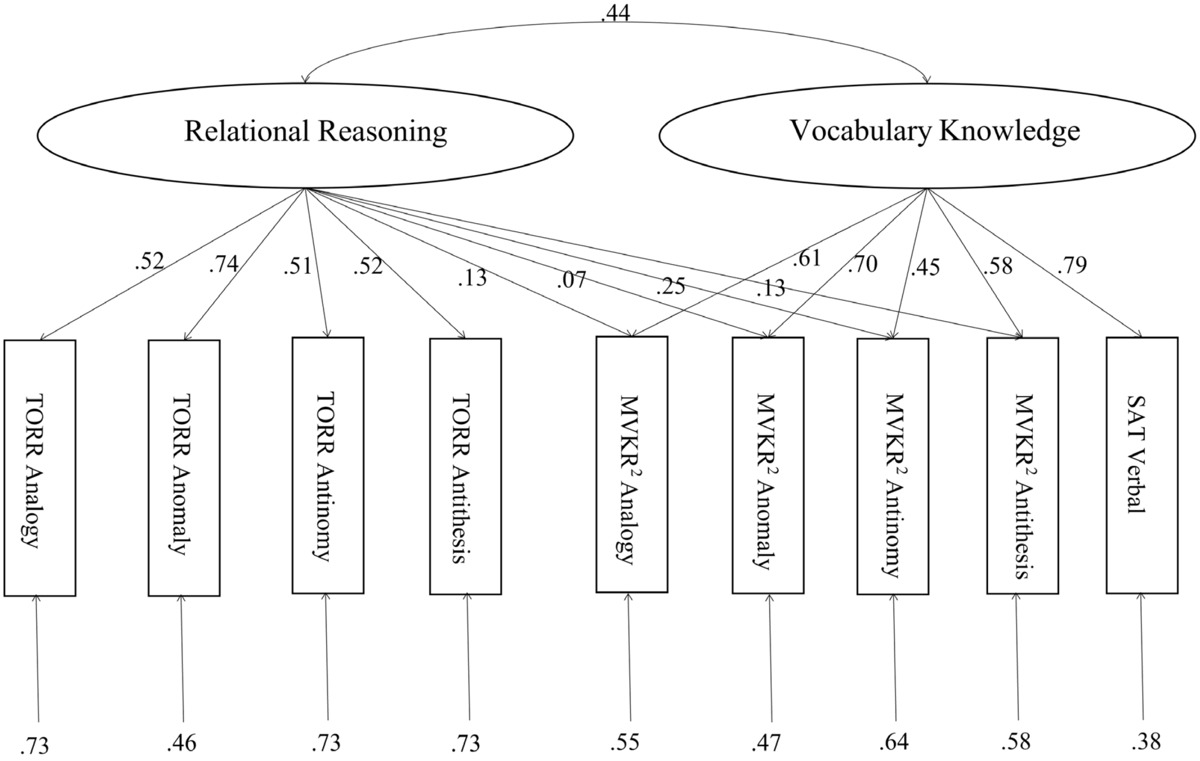
Test-Level Examination of the Contributions of Fluid Relational Reasoning and Vocabulary Knowledge to the Measure of Vocabulary Knowledge through Relational Reasoning



Note. Due to space limit, some path coefficients of the measurement models are omitted. MVKR^2 = Measure of Vocabulary *Knowledge* through Relational Reasoning; TORR = Test of Relational Reasoning; AG = Analogy; AM = Anomaly; AN = Antinomy; AT = Antithesis. *. p < .001.

Figure 7

Scale-Level Examination of the Contributions of Relational Reasoning and Vocabulary Knowledge to the Measure of Vocabulary Knowledge through Relational Reasoning



Note. MVKR^2 = Measure of Vocabulary Knowledge through Relational Reasoning; TORR = Test of Relational Reasoning. All coefficients in this figure are significant, except the loadings of MVKR^2 Anomaly (estimate = .07, *p* = .359) and MVKR^2 Antithesis (estimate = .13, *p* = .084) on Relational Reasoning.