Journal of Contextual Behavioral Science 30 (2023) 155–168

Contents lists available at ScienceDirect

Journal of Contextual Behavioral Science

journal homepage: www.elsevier.com/locate/jcbs

On the measurement of relational responding

Jamie Cummins [a,b]

[a] Department of Experimental-Clinical and Health Psychology, Ghent University, Belgium
[b] Institute of Psychology, University of Bern, Switzerland

ARTICLE INFO

Keywords:
Relational responding
Relational reasoning
Measurement
Item response theory
Measurement precision

ABSTRACT

Psychologists increasingly recognize the importance of relational responding in understanding human behavior. As a result, there is a growing need for valid, reliable, and precise measures of relational responding. One promising measure is the Relational Abilities Index (RAI). However, its measurement properties have not been explored in-depth. There is little understanding, for example, of how precise the RAI is as a measure of individual-level abilities. In this paper I examine the precision of the RAI at the individual-level, quantify the item characteristics of the task using Item Response Theory, and more generally assess its quality as a measure of relational responding. Although broadly promising, the measure exhibits several substantial shortcomings which limit its utility and interpretability. I make recommendations for how to improve the task and highlight the importance of in-depth examinations of measurement for future research.

1. Introduction

A quiet "relational revolution" has begun to sweep across psychological science. Across fields, researchers are converging on a critical idea: that the cornerstone of human language and cognition is our ability to relate stimuli in arbitrary and complex ways (McLoughlin et al., 2020). This convergence has happened for good reason: relational responding has been implicated in a varied array of phenomena, including (but certainly not limited to) academic performance (Alexander et al., 2016), perspective-taking (Montoya-Rodríguez et al., 2017), and psychopathology (Hayes, Merwin, et al., 2021). In behavior analysis, the primary approach that advocates for this relational view is Relational Frame Theory (RFT; Hayes et al., 2001). RFT states that the ability to relate stimuli in arbitrary and complex ways can be considered an operant response class, typically referred to as arbitrarily applicable relational responding (AARR; Hayes et al., 2021a). This perspective is unique to RFT: most theoretical accounts from other fields tend to conceptualize relational responding as a mental mechanism, rather than as an operant behavior (McLoughlin et al., 2020). Importantly, conceiving of relational responding as an operant comes with the necessary implications that relational responding should be subject to environmental control, and that fluency in relational responding should be trainable through multiple exemplar training (Cassidy et al., 2010).

Both implications are well-supported. On the one hand, research has shown that relational responding can be brought under operant control, both in terms of the specific types of relational responses that people emit (i.e., C*rel* control; Hughes et al., 2019; Perez et al., 2017) as well as the specific properties of stimuli which people relate (i.e., C*func* control; Delabie et al., 2022; Finn & De Houwer, 2021). On the other hand, there are now many studies which demonstrate that relational responding fluency can be trained directly (Cassidy et al., 2016; Colbert et al., 2018; Dixon, Belisle, McKeel, et al., 2017a,b; McLoughlin et al., 2021). Two "relational training" protocols have been developed within behavior analysis which are explicitly derived from this operant view of relational responding: SMART (Strengthening Mental Abilities with Relational Training; Cassidy et al., 2011) and the PEAK relational training system (Dixon, Belisle, McKeel, et al., 2017a,b). Both SMART and PEAK have shown efficacy at improving relational responding abilities (Beck et al., 2023; Dixon et al., 2021; May et al., 2022). As well as this, there is growing evidence that these training programmes may also lead to improvements on other outcomes such as IQ and educational attainment, which serves to further support the idea of relational responding as foundational in human cognition (e.g., Hayes & Stewart, 2016; but see May et al., 2022; Thirus et al., 2016).

In tandem with the "relational revolution", another revolution has begun to stir within psychological science: a revolution of measurement. It has become increasingly clear that psychological science has at times been lax in evaluating the validity of its measurement procedures. Indeed, there is now substantial evidence that the validity of psychological measures has been selectively reported (Hussey & Hughes, 2020), and that many measurement procedures used in psychological science are either (i) not measuring what they claim to measure, or (ii)

E-mail address: Jamie.Cummins@UGent.be.

https://doi.org/10.1016/j.jcbs.2023.10.003
Received 7 July 2023; Received in revised form 2 October 2023; Accepted 8 October 2023
Available online 10 October 2023

are measuring what they claim to measure in a rather imprecise manner (Flake & Fried, 2020; Lilienfeld & Strother, 2020). This is problematic; we cannot produce good or useful theories in the absence of reliable and accurate measurement (Eronen & Bringmann, 2021). Given the recent Association for Contextual Behavioral Science Task Force report which emphasized the need for more precise measurement (particularly at the individual level; Hayes, Merwin, et al., 2021), this issue is now more relevant than ever.

With all the above said, it is worthwhile to inquire into the integrity of measures of relational responding. One popular measure for this purpose is the Relational Abilities Index (RAI; Colbert et al., 2017). The RAI is a repeatable assessment battery which tests participants' abilities in different types of relational responding. The number of relations which are assessed in the measure has increased over time. Originally, only 2 relations were assessed (opposition and quantity; Colbert et al., 2017); subsequently, a 5-relation version was developed (assessing opposition, quantity, difference, temporality, and analogy; Colbert et al., 2019). Most recently, an 8-relation version has been used (assessing the 5 relations from the 5-relation version as well as hierarchy, deictic, and mathematical relations; Cummins, Nevejans, Colbert, & De Houwer, 2023). Each relation is assessed across a series of trials where yes/no answers are required. Participants are presented with a series of premises (e.g., A is the same as B, B is the same as C), and a question relating to these premises (e.g., is A the same as C?). As participants progress in each subscale, the subscale's trials increase in terms of the number of premises involved, the number of required derivations to answer correctly, and several other criteria which RFT predicts should require more "complex" relational responding (Hayes et al., 2001).

In a recent examination of the task, Cummins and colleagues (2022) investigated both the split-half and 1-week test-retest reliability of the measure. Although the full RAI score exhibited excellent psychometrics, the subscales of the measure fared differently: split-half reliability varied from adequate to poor, whereas test-retest reliability was generally poor. Colbert et al. (2017) did not report subscale-level psychometrics but found similarly that the overall test-retest of the RAI was adequate. In a different vein, Ruiz Méndez and colleagues (2022) recently used a 4-relation RAI to examine whether the overall number of errors in trials in each subscale increased as the task progressed, as would be predicted based on an RFT perspective on the task (Cassidy et al., 2016; Colbert et al., 2017). The results were mixed; two subscales exhibited this expected pattern, whereas two did not.

The low reliability of the RAI subscales should be particularly concerning for Contextual Behavioral Science (CBS) researchers, given that the precision of individual-level scores (i.e., the scores of specific single individuals) in a measurement procedure is a direct function of its reliability (Dudek, 1979). In other words: if the reliability of a subscale is poor, then a given participant's score will not tell us much about their fluency in the probed relational response class. Although metrics for individual-level precision of scores in such tasks are not currently used often within CBS (or in psychology more generally), such metrics do exist. For example, statistical approaches such as the standard error of measurement (Dudek, 1979), the rescaled estimates of true scores (RETS; Schmukle, 2023), and bootstrapping (Mooney et al., 1993) can all be used to quantify the (im)precision of a measurement procedure at the individual-level; see for example Hussey's (2020) application of bootstrapping in the Implicit Relational Assessment Procedure, and Cummins and Hussey (2023).

This paper has several research questions (RQs) which will be addressed by different analyses (these RQs are detailed specifically in the Analytic Strategy section below). First, I aimed to further investigate the psychometric properties of the RAI. I estimated the individual-level precision of RAI scores, as well as the implications that the current levels of precision have for the relative discriminability of RAI subscale scores (RQ1). Thereafter, I also examined the potential impact that increasing the number of trials in RAI may have on this precision (RQ2). I then conduct similar analyses relating to the split-half reliability of the measure (RQ3 & RQ4). Finally, I used Item Response Theory (IRT) to investigate a core assumption of the measure: namely, that the complexity (and by extension, difficulty) of items within each subscale increases as items progress (RQ5; Cassidy et al., 2016). Together, these investigations aim to identify whether RAI in its current form is suitable for the goals of relational responding researchers, and if not, to identify the ways in which it is suboptimal.

2. Method

2.1. Sample

All data, processing code, and analysis code are openly available via the Open Science Framework (https://osf.io/rqav2/?view_only=18200a40711040fb8ad34d99ab9fb27a). The data used in this study consist of both human and simulated observations, analyzed separately. The human observations are derived from published and unpublished data. Specifically, RAI data are taken from the time 1 assessments in Experiment 2 of [REDACTED FOR ANONYMITY] and data from unpublished studies using the 8-relation RAI. These data were only used if the RAI was the first measure participants completed and if participants were typically developed adults between 18 and 40 years old. In total, the human data consist of 264 participants (116 men, 147 women, 1 non-binary) with a mean age of 28.86 years (SD = 5.87 years). These participants were recruited either via Prolific Academic or word-of-mouth. The simulated observations were produced based on the density distribution of the scores for the human observations. Firstly, I extracted the kernel density function for each separate subscale of the human RAI data using the bkde() function from the KernSmooth R package (Wand, 2021). Kernel density estimation in essence is a method which allows for the relatively precise approximation of complex distributions, which in turn can then be used to simulate new observations from the approximated distribution (Chen, 2017). I simulated 1000 new observations of scores for each of these subscales based on this density distribution, bounding observations between 0 and 1 (1000 was chosen to provide a sufficient range of observations while also limiting the computational complexity of the to-be-run models). The distributions of the human and simulated data can be inspected in Fig. 1.

2.2. Materials

Relational Abilities Index (RAI). Detailed information regarding the RAI procedure can be found in Colbert et al. (2019). In short, the RAI in all data sources was identical to that used within Experiment 2 of [REDACTED FOR ANONYMITY], consisting of 8 relational subscales divided into 16 trials each (opposition, difference, quantity, temporal, containment, analogy, deictic, and mathematical relations). Each trial required a binary yes/no response to a relational question based on a set of relational premises. Trials in each subscale are stated to scale up increasingly in difficulty as trials increase, by varying the number of relational premises, the number of derivations required to be made by the participant in answering the question, and other similar dimensions (Cassidy et al., 2016; Colbert et al., 2017, 2019). Fig. 2 illustrates examples of trials from each of the 8 subscales.

2.3. Analytic Strategy

2.3.1. Analytic methods

Given that some of the analytic methods employed in the manuscript may be novel to readers, in this section I provide an overview of these methods, broadly explain their logic and background, and clarify the way in which they are useful for the purposes of this paper.

Individual-level estimation of scores. When we compute scores in a measure, there is always uncertainty associated with these scores, and this uncertainty decreases as our number of measurements increase. If a coin is flipped 100 times and 70% of the results are heads, we can be

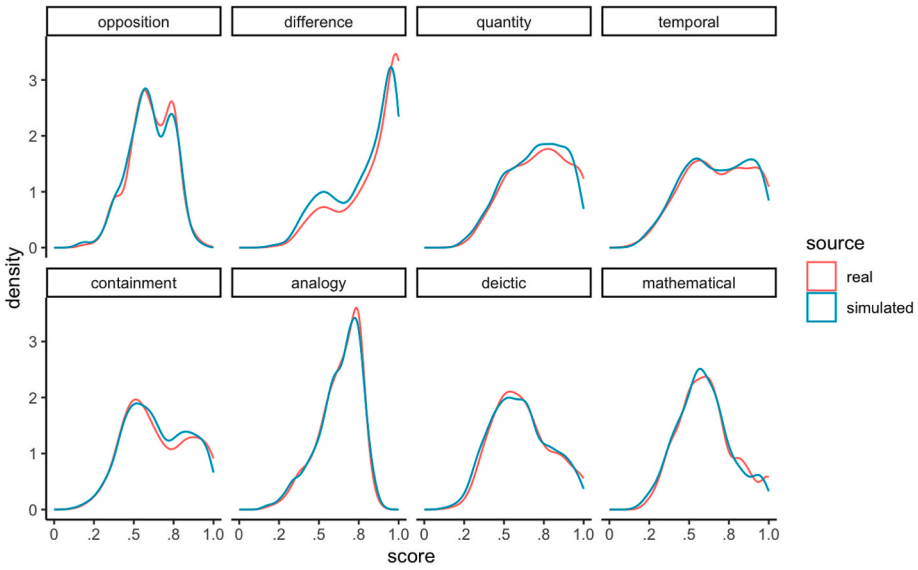

Fig. 1. Density distributions of the (i) real and (ii) simulated RAI subscale data.

much more confident that the coin is unfair than if the coin showed 70% heads after only 10 flips. We are used to employing this logic when it comes to the measurement of a group of participants: we increase our sample size to improve the precision of our measurement, and we estimate the precision of that measurement by examining the width of (e.g.) 95% confidence intervals (CIs). We also then use these confidence intervals to make inferences; in general, in a two-sided statistical test with an alpha level of 0.05, if the 95% CI excludes zero then the associated p-value will be less than 0.05 (Knol et al., 2011).

In the context of individual-level measurement, we surprisingly rarely compute the confidence intervals associated with individual persons' scores. However, just like with group-level analyses, there is necessarily uncertainty associated with these scores, and estimating this uncertainty can provide us with insight into how meaningful individual persons' scores are. To address RQ1 and RQ2, it was therefore necessary to first identify a method for extracting individual-level confidence intervals on the RAI data. Critically, scores in the RAI follow a very distinct distribution: namely, responses in each trial can either be correct (i.e., a success) or wrong (i.e., a failure). As such, the RAI can be conceptualized as a series of Bernoulli trials which form a binomial distribution; the sum score which is computed for the RAI or its individual subscales can be considered the "success" parameter of the binomial distribution for a given participant (this renders this score quite similar to the mean number of heads/tails in the context of coin-flipping). Confidence intervals on this binomial distribution can then be trivially computed using Clopper & Pearson's (1934) method; computing confidence intervals from the binomial distribution is a highly conventional method, akin to the familiarity of computing standard deviation or z-scores based on the Gaussian distribution (Edwards, 1960).

It is worth noting that it is likely not accurate to conceive of the RAI in terms of the binomial distribution. The binomial distribution assumes both independence between trials and an identical probability of success on all trials. These assumptions are violated in how the RAI is conceived; minimally, there is an assumption that later trials have a lower probability of success than earlier trials. However, there is a discrepancy between this conceptualization of the RAI and its use of sum scores/averages as its score. The use of sum scores or averages (as has been done exclusively with the RAI to date) also does not account for dependencies between trials, nor differences in the relative probabilities of success across different trials (cf. McNeish & Wolf, 2020). As such, the use of the binomial distribution here to quantify confidence intervals does not mean to imply that the assumptions of the binomial hold; rather, it is simply a reflection of the assumptions that are already baked into the use of the sum score and its derivatives.

Item Response Theory. The modal set of methods used to investigate the psychometric properties of psychological tasks can be broadly referred to as Classical Test Theory (CTT). CTT in general focuses on the overall scores which are produced in each test, estimating (for example) how stable these scores are across time (test-retest reliability) or when sampling from different test items (e.g., split-half reliability). CTT (in its most fundamental form) assumes that all items in the examined measure contribute equally to the total test score (Wang & Osterlind, 2013), and approaches to item development are typically done with respect to how adding/dropping items affects the overall test score (e.g., does dropping a certain item increase reliability; Raykov, 2008). Although CTT is clearly useful, it does not concern itself with the properties of items within the test themselves. In contrast, Item Response Theory (IRT; Embretson & Reise, 2013) is fundamentally concerned with the nature and characteristics of specific items, particularly as they pertain to the latent ability Θ that the scale seeks to assess. In the case of the RAI, this latent ability can be conceptualized as relational responding to a particular type of relation (i.e., depending on the subscale). IRT, then, can allow us to determine how specific trials in each subscale relate to this latent ability.

In the RAI, there is an assumption that later trials in a given subscale are in general more difficult than earlier trials due to their greater relational complexity (Cassidy et al., 2016; Colbert et al., 2017). Models in IRT can allow us to address this question by determining the difficulty of items in each subscale of the RAI *inductively*. IRT consists of a range of models which vary in terms of their complexity. The simplest of these is the dichotomous 1-parameter logistic model (1-PL model; Rasch, 1960). In short, this model assumes that all features of all items in a scale are identical, except for a single element: the difficulty of the item. Given these assumptions, the 1-PL model estimates Item Characteristic Curves (ICCs), which illustrate the functional relationship between the probability of a correct response on the given item and the latent trait being assessed. The point on the "ability" axis at which this curve is most steep represents the "difficulty" of the item, which is also referred to as the *location* parameter (Lord, 2012). Items which are more difficult have a higher location parameter value, whereas items which are easier have lower location parameter values.

The 1-PL model can also be used to estimate Test Characteristic Curves (TCCs) and Test Information Curves (TIC) for the entire scale(s) of interest (Embretson & Reise, 2013). The TCC simply represents the

Fig. 2. Example trials from each of the 8 subscales of the RAI. Reading from left to right and top to bottom, the represented subscales are: opposition, difference, quantity, temporal, containment, analogy, deictic, and mathematical relations.

sum of the ICCs for each subscale item weighted by the score for each item. This provides an overview of the relationship between the overall subscale score and the ability level of participants. The TIC is, in essence, a plot which visualizes the measurement error in the scale associated with different levels of the latent ability being assessed (this contrasts with CTT, which assumes that measurement error is uniformly distributed; Wang & Osterlind, 2013). Lower levels of information are associated with higher degrees of measurement error. For this paper, both the ICC and TIC can provide valuable information about the subscales of the RAI.

Some readers will already note that the assumptions of the 1-PL model align well with those of the RAI. Indeed, other IRT models which could be selected (e.g., the 2-PL model) make additional assumptions which have not been explicitly stated in the context of the RAI. For example, the *discriminability* of items can be allowed in addition to the difficulty of items, which would mean that some items could be better at distinguishing between participants of different abilities than others. In such a case, the associated curves for each item could vary not only in terms of their location in the plot, but also in terms of the shapes of each item curve. Although IRT analyses would typically conduct a comparison of these various models and choose the best-fitting candidate for further analysis, the analytic approach here seeks to test the extant theoretical claims relating to the RAI, *not* to identify the best fitting IRT model for the data. As such, I opted to use the 1-PL and not compare different IRT configurations (although future work aimed at further developing the RAI could greatly benefit from doing this). It should be noted, however, that I did not fix the discriminability *between* subscales; as a consequence, the shapes of the ICCs were identical within subscales, but differed between subscales.

2.3.2. Research questions

RQ1. How precise is the existing RAI at the individual-level? The analytic approach used across all sub-questions for RQ1 is directly inspired by (and near-identical to) the method employed by Hussey (2020).

RQ1.1. What is the average CI width for participants across subscales? This first research question aimed to get at the simple idea of how wide, on average, the confidence intervals of human participants were. This can provide a sense of the precision of the subscale in general. I addressed this question using a multilevel linear model (see Bates et al., 2015), with CI width modelled as the dependent variable, subscale modelled as a predictor, and study id and participant id modelled as random intercepts. The Wilkinson notation of this model was as follows:

$$\text{CI Width} \sim 1 + \text{subscale} + (1 \mid \text{study id} \,/\, \text{participant id})$$

I then used a likelihood ratio test to compare this model with a model identical to this but without subscale included as a predictor (i.e., to assess whether the RAI subscale had a meaningful impact on CI width). The mean CI width for each subscale was then estimated based on the marginal means for each subscale.

RQ1.2. What proportion of participants' scores exclude the possibility of random responding across subscales? This question investigated the proportion of participants in each subscale whose scores were significantly greater than chance level. A similar strategy was used here as in RQ1.1, but using a multilevel logistic model (rather than linear model), with significance (TRUE or FALSE) used as the dependent variable. Significance was determined based on whether the estimated confidence intervals for each participant excluded 0.5 (given that the nonoverlap of confidence intervals with a point estimate value is equivalent to a slightly more conservative one-sample *t*-test; Knol et al., 2011). The predictor and random effects were identical to those of RQ1.1. Specifically:

$$\text{Discriminable from chance} \sim 1 + \text{subscale} + (1 \mid \text{study id} \,/\, \text{participant id})$$

As before, a likelihood ratio test was used to determine whether discriminability from chance was predicted significantly better by factoring in the subscale, and the mean proportion of discriminability from chance for each subscale were estimated based on the marginal mean predictions from the model.

RQ1.3. What proportion of participants' scores are discriminable from one another across subscales? Not being differentiable from chance-level responding is not in-and-of-itself an indicator of measure quality. For example, it might be the case that difference responding is generally less complex than analogical responding; greater chance-level responding would therefore necessarily be expected for analogical responding compared to difference (indeed, different relational responses being differentially complex is perfectly in line with RFT; Hayes et al., 2001). What can be more informative is how well the measure serves to discriminate *between* individuals. This is precisely the motivation behind RQ1.3. To address this, within each subscale each participant's score was compared with every other participant's score. Participants were registered as being significantly different from another participant in instances where

$$|x_1 - x_2| > 1.96 \times \sqrt{SE_1^2 + SE_2^2} \tag{1}$$

where X_1 is the first participant, X_2 is the comparison participant, SE_1 is equal to the standard error of the first participant's score, and SE_2 is equal to the standard error the second participant's score. This method of comparison is more accurate to detect true differences between observations than merely comparing the (non-)overlap of confidence intervals (see Austin & Hux, 2002; Cornell Statistical Consulting; Tan & Tan, 2010; Cornell Statistical Consulting Unit, 2008). The individual-level SEs were derived from the basic formula

$$SE = \frac{(CI_{upper} - CI_{lower})}{(1.96 * 2)} \tag{2}$$

After comparing each participant, the observations were entered into a multilevel logistic regression, with significant difference (TRUE or FALSE) entered as the dependent variable. The predictors and random effects were otherwise identical to RQ1.1 and RQ1.2. Specifically:

$$\text{Discriminable from other participant} \sim 1 + \text{subscale} + (1 \mid \text{study id} \,/\, \text{participant id})$$

Once again, a likelihood ratio test was run to compare this model to one which excluded subscale as a predictor, and estimated marginal means were used to identify the mean proportion of discriminability from other participants.

RQ2. How would increasing trial lengths affect the RAI's individual-level precision? RQ1 focuses on evaluating the individual-level precision of the RAI along three dimensions. Individual-level precision is essentially a form of estimation; as we know from estimation at the group-level, increasing sample size can improve estimates of models. By extension, if we increase our sample size within each participant (i.e., the number of trials completed in the RAI), then the individual-level precision of the measure should improve. RQ2's analyses therefore

echoed those of RQ1 but investigated the levels of individual-level precision that would be achieved at varying trial lengths using a simulation-based approach, simulating data for 1000 observations as described in the Sample section above. Confidence intervals for the simulated scores were subsequently estimated assuming varying numbers of trials: ranging from 16 to 256 trials in steps of 16.

RQ2.1. How do average CI widths change as a function of trial length across subscales? This research question focused on estimating how average CI widths varied as a function of the number of trials used in the task. This was achieved arithmetically, since given a proportion correct and number of trials the confidence intervals of the binomial distribution can be computed trivially without the need for simulation (Clopper & Pearson, 1934). However, it is important to note that this tells us little about the average CI widths across subscales, given that subscales' scores were not uniformly distributed, and they may not be identically distributed between subscales. Therefore, I investigated this question while also modelling this between-subscale variability in the simulated data. The model used for this question was broadly the same as in RQ1.1, with the addition of number of trials as an added predictor as well as its interaction with subscale, and the removal of study id as a random effect. The Wilkinson notation for this model is:

$$\text{CI Width} \sim 1 + \text{n\_trials} + \text{subscale} + \text{n\_trials} : \text{subscale} + (1 \mid \text{participant id})$$

A likelihood ratio test was used to compare this model to one which did not factor in the number of trials into its predictions, and the interaction between number of trials and subscale was both verified statistically and presented visually.

RQ2.2. How does discriminability from random responding change with increasing trial lengths? This question investigated the rate of discriminability of participants from chance-level responding as a function of increasing trial lengths. RQ2.2's model therefore paralleled that of RQ1.2's, but with the addition of number of trials as an added predictor variable, as well as its interaction with subscale. In other words:

$$\text{Discriminable from chance} \sim 1 + \text{n\_trials} + \text{subscale} + \text{n\_trials} : \text{subscale} + (1 \mid \text{participant id})$$

A similar analytic approach to RQ2.1 (likelihood ratio test, test of main and interaction effects, and visualization of relationship) was used.

RQ2.3. How does discriminability from other participants change with increasing trial lengths? This question investigated the rate of discriminability of participants from other participants as a function of increasing trial lengths. Like the previous two analyses in RQ2, I had initially planned to model the number of trials, and its interaction with subscale, as additional predictors to its counterpart model from RQ1.3. Specifically:

$$\text{Discriminable from other participant} \sim 1 + \text{n\_trials} + \text{subscale} + \text{n\_trials} : \text{subscale} + (1 \mid \text{study id} \,/\, \text{participant id})$$

Critically, however, the number of possible comparisons between each participant on each trial-type at each number of trials was extremely large (with 1000 participants, 8 trial-types, and 16 different numbers of trials yielding approximately 128,000,000 comparisons).

Table 1
Comparisons of mean individual-level CI width across RAI subscales. P-values for each subscale are relative to the reference category; the p-value for the reference category is compared to a null effect.

Subscale	Estimated marginal mean CI Width (95% CIs)	p
Opposition (reference)	.48 (.44, .52)	<.001
Difference	.33 (.29, .37)	<.001
Quantity	.41 (.37, .46)	<.001
Temporal	.42 (.38, .46)	<.001
Containment	.43 (.39, .47)	<.001
Analogy	.47 (.43, .51)	= .520
Deictic	.45 (.41, .49)	<.001
Mathematical	.46 (.42, .50)	= .002

Running this model would have taken quite some time. As a more computationally manageable alternative, I instead switched to the use of a fixed-effects linear regression model. Specifically, I computed a "discriminability score" for each participant at each number of trials and for each subscale (i.e., a value between 0 and 1 indicating the proportion of participants which could be discriminated from the present participant). A score of 1 indicated that this participant was discriminable from all other participants, whereas a score of 0.5 indicated that this participant was discriminable from exactly half of all other observations. Although nonlinear regression methods such as beta regression would generally be more appropriate for modelling such proportion data, I opted for a simpler linear model due to the relative ease of interpretation of linear model coefficients, which will aid in future decision-making for RAI researchers. This regression can be specified as:

$$\text{Discriminability score} \sim 1 + \text{n\_trials} + \text{subscale} + \text{n\_trials} : \text{subscale}$$

I then examined whether each of the terms in the model were significant predictors of discriminability, and also plotted the estimated marginal predictions for each subscale across the number of trials.

RQ3. What is the split-half reliability of the existing RAI across subscales? This research question is relatively straightforward and simply consisted of examining the split-half reliability of the RAI, which has not been reported extensively to date.

RQ4. How does the split-half reliability of the RAI change across subscales with increasing trial lengths? This research question was similar to RQ3, but used the Spearman-Brown prophecy formula to project the split-half reliability of each RAI subscale for various trial lengths (Brown, 1910; Spearman, 1910).

RQ5. Do items in each RAI subscale align with the theoretical expectations of RFT? The ordering of scale items in the subscales of the RAI is not random. Scale items within subscales are assumed to follow a general increase in difficulty as blocks progress (Cassidy et al., 2016; Colbert et al., 2017). The most extreme constraint of this assumption would be one where, in every instance, Trial N would be more difficult than Trial N-1, which would be more difficult than Trial N-2, and so on. The psychometric term for such an arrangement of trials is a Guttman scale. Methods from both Classic Test Theory (CTT) and Item Response

Theory (IRT) can be used to assess this. From CTT, we can examine the proportion of *Guttman errors* for each subscale (i.e., instances where a correct response on Trial N was immediately preceded by an incorrect response on Trial N-1; Meijer, 1994). Although users and developers of the RAI have not explicitly stated that the RAI ought to strictly conform to a Guttman structure, it is still useful to examine these results in the context of what has been explicitly stated (namely, that trials *in general* should be more difficult as the task progresses). Though we might expect a greater number of violations/errors under this less strict assumption, we should still broadly expect a relatively low number.

After this, I also examined the difficulty of items in each subscale *inductively* using IRT. I firstly fit 1-PL models to each of the 8 RAI subscales separately. Next, for each item in each subscale I extracted the difficulty (aka location) coefficients estimated by the 1-PL models and graphically examined the relative correspondence between the theoretically proposed difficulties and observed difficulties. I also plotted the associated ICCs for each subscale. I also plotted TICs for each subscale to examine the relative measurement error associated with the subscales at varying levels of participant ability.

3. Results

RQ1. The individual-level precision of the RAI in existing data

For RQ1.1 (the mean precision at the individual-level), in a likelihood ratio test the model including subscale as a predictor of mean CI width fit the data significantly better than a null model excluding this predictor, $\chi^2(7) = 761.45$, $p < .001$. For the model itself, the subscale factor was dummy coded with the first subscale (i.e., opposition) of the RAI used as the reference category. The results of this model can be seen in Table 1. Although CI widths varied significantly between subscales, it is critical to note that these widths were generally range wide, with mean estimates across subscales ranging from 0.33 to 0.48.

For RQ1.2 (the mean discriminability from zero), comparing the model which included subscale as a predictor to the model without subscale as a predictor, the subscale-included model once again fit the data better, $\chi^2(7) = 604.35$, $p < .001$. Overall, discriminability from chance differed significantly between subscales, $p < .001$. The specific estimated marginal proportions for each subscale can be found in Table 2. The performances of the subscales varied widely, but generally were not desirable: except for the difference subscale, less than 50% of participants on a given subscale could be successfully discriminated from chance responding.

For RQ1.3 (the mean discriminability from other participants), once more the model with subscale as a predictor was a better fit than the intercept-only model, $\chi^2(7) = 37669$, $p < .001$. The specific predictions and estimated marginal means can be found in Table 3. Once again, however, discriminability was not optimal, ranging from ~6% (in the analogy subscale) to 33% (in the temporal subscale). In other words: even in the best performing subscale, a participant on average could only be discriminated from 33% of other participants.

Table 2
Comparisons of discriminability from chance of individual scores across RAI subscales. P-values for each subscale are relative to the reference category; the p-value for the reference category is compared to a null effect.

Subscale	Estimated proportion of participants discriminable from chance responding (95% CIs)	p
Opposition (reference)	.02 (.01, .05)	<.001
Difference	.85 (.72, .93)	<.001
Quantity	.38 (.22, .57)	<.001
Temporal	.32 (.18, .51)	<.001
Containment	.26 (.14, .43)	<.001
Analogy	.01 (.01, .03)	= .372
Deictic	.12 (.06, .24)	<.001
Mathematical	.09 (.04, .17)	<.001

Table 3
Comparisons of discriminability from other participants across RAI subscales. P values for each subscale are relative to the reference category; the p-value for the reference category is compared to a null effect.

Subscale	Estimated proportion of participants discriminable from each other (95% CIs)	p
Opposition (reference)	.07 (.07, .08)	<.001
Difference	.32 (.30, .33)	<.001
Quantity	.29 (.28, .30)	<.001
Temporal	.33 (.32, .35)	<.001
Containment	.32 (.31, .33)	<.001
Analogy	.07 (.06, .07)	<.001
Deictic	.24 (.23, .25)	<.001
Mathematical	.20 (.19, .21)	<.001

RQ2. Simulating differing trial lengths on the RAI's individual-level precision

RQ2.1 focused on estimating how average CI widths varied as a function of the number of trials used in the task. The simplest approach to answering this question would be to simply examine how CI widths vary as a function of increasing trial lengths and proportions of correct responses (i.e., the RAI subscale score). This general relationship can be seen in Fig. 3.

However, modelling the simulated data directly can also reveal critical insights and account for heterogeneity between subscales. The model fitted to the simulated data fit the data significantly better than an equivalent model without number of trials included, $\chi^2(8) = 109352$, $p < .001$. The omnibus main effects and interaction effect were all significant, $p < .001$. The broad pattern of results can be seen in Fig. 4.

For RQ2.2 (comparing discriminability from chance), the model including number of trials fit the data better than a model excluding this predictor, $\chi^2(14) = 6285.4$, $p < .001$. The omnibus main effects and interaction effect were all significant, $p < .001$. On average across subscales, every 16 trials increased the proportion of participants discriminable from 0.5 by about 0.034. Fig. 5 illustrates this.

For RQ2.3, both the main terms and the interaction term in this model were significant predictors of the discriminability score ($p < .001$). Fig. 6 illustrates the estimated marginal predictions of discriminability for each subscale across increasing numbers of trials.

RQ3. Split-half reliability of the RAI subscales in existing data

Split-half reliability of the RAI has generally not been reported frequently. Although split-half reliability for the RAI was previously computed by Cummins, Hussey, and Spruyt (2022), a larger sample size naturally provides a better estimate. The Spearman-Brown corrected odd-even split-half scores for each of the 8 subscales can be found in Table 4.

RQ4. Impact of subsequent trials on split-half reliability

The Spearman-Brown prophecy formula can conveniently be used to estimate the impact of adding further trials on split-half reliability (Remmers & Ewart, 1941). By extension, it can also be manipulated to solve for the number of trials required to achieve a particular level of reliability. Here, I firstly examined the impact doubling the number of trials would have on each subscale's reliability. As well as this, I also examined how many trials would be required for each subscale separately to achieve specific criterion levels of split-half reliability (i.e., 0.80, 0.90, and 0.99). The results from these analyses are presented in Table 4.

RQ5. Do item difficulties in each RAI subscale align with the theoretical expectations of RFT?

I used a combination of approaches from CTT and IRT to examine

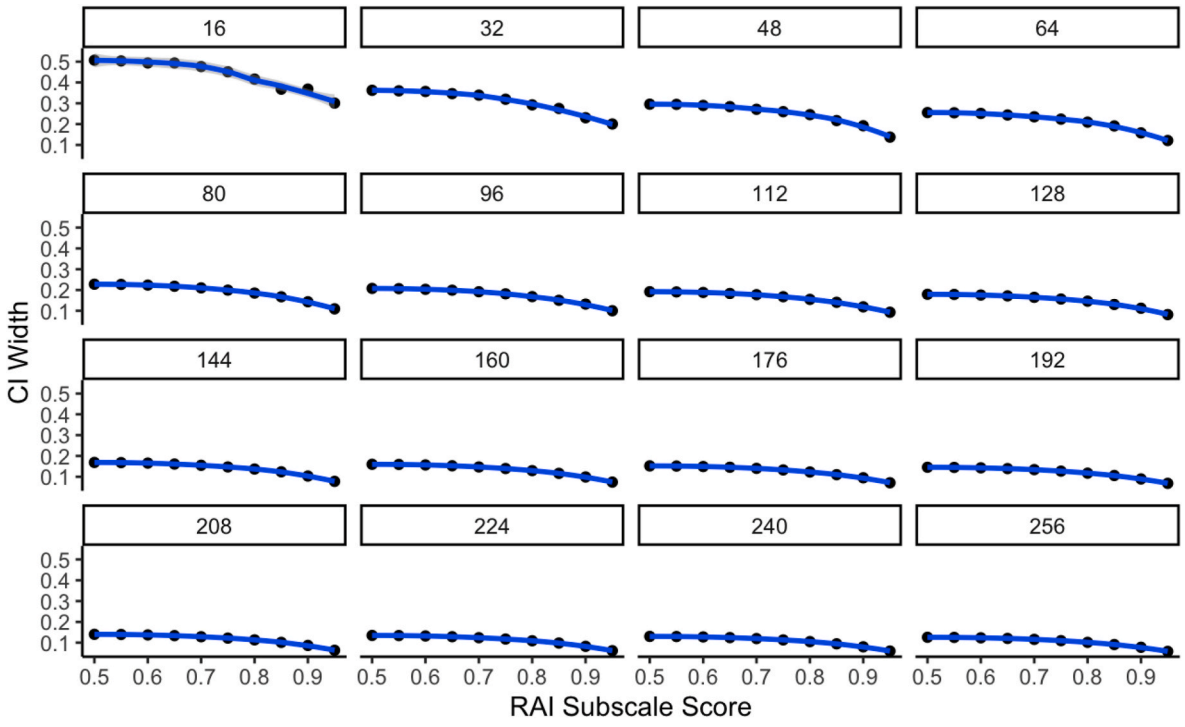

Fig. 3. The impact of increasing trial numbers on RAI subscale CI widths.

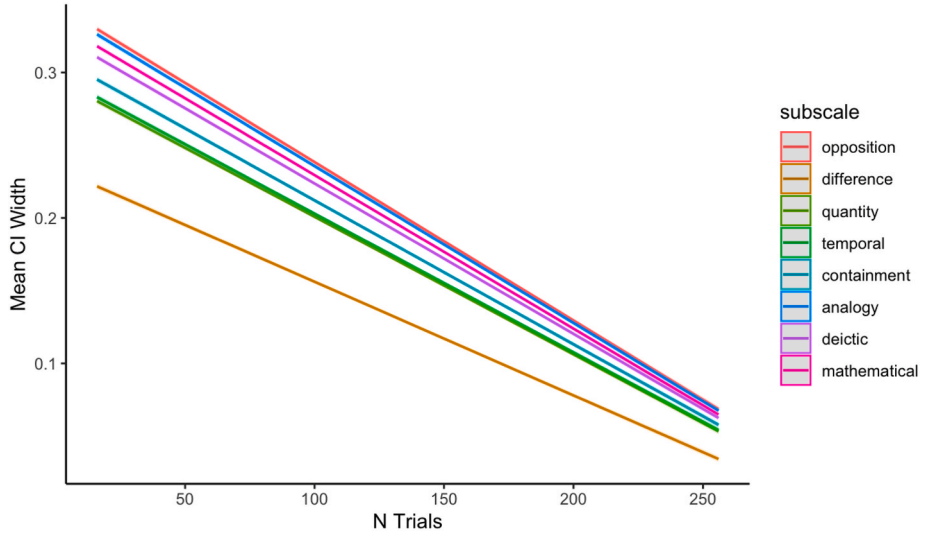

Fig. 4. The change in CI width in the simulated data as a function of increasing trial lengths.

whether the characteristics of the subscales and their items corresponded with the conceptual accounts of the RAI and RFT more generally (i.e., that more complex relational responding ought to entail more difficulty; Cassidy et al., 2016; S. C. Hayes et al., 2001). First, I examined the proportion of Guttman errors for each subscale (i.e., instances where a correct response on Trial N was immediately preceded by an incorrect response on Trial N-1). Table 5 provides results for the mean proportion of Guttman errors for each subscale; in general, these scales did not abide by a strict Guttman scale, and there were substantial proportions of Guttman errors that suggested that item difficulties did not scale in the manner expected.

These results strongly suggest that the difficulties of items in RAI subscales do not conform to a Guttman structure, and that there are structural violations even of the assumption that the trials *in general* increase in difficulty. This was also apparent when plotting the extracted item difficulties from the IRT models against the theoretically predicted difficulties of the items (i.e., their ordering within the task). Fig. 7 illustrates these relationships. It is clear from these plots that item difficulty did not systematically increase across trials in the subscales; in fact, in many cases it appears that earlier trials proved more difficult than later trials (in line with the above findings relating to violations of Guttman structure).

Examining the Item Characteristic Curves associated with each subscale revealed a further interesting pattern: in general, the difficulties associated with each item within each subscales were very similar. The two exceptions to this fact, the opposition and analogy subscales, demonstrate notably poorer discriminability (i.e., the slopes of their curves are substantially smaller) than the other 6 subscales.

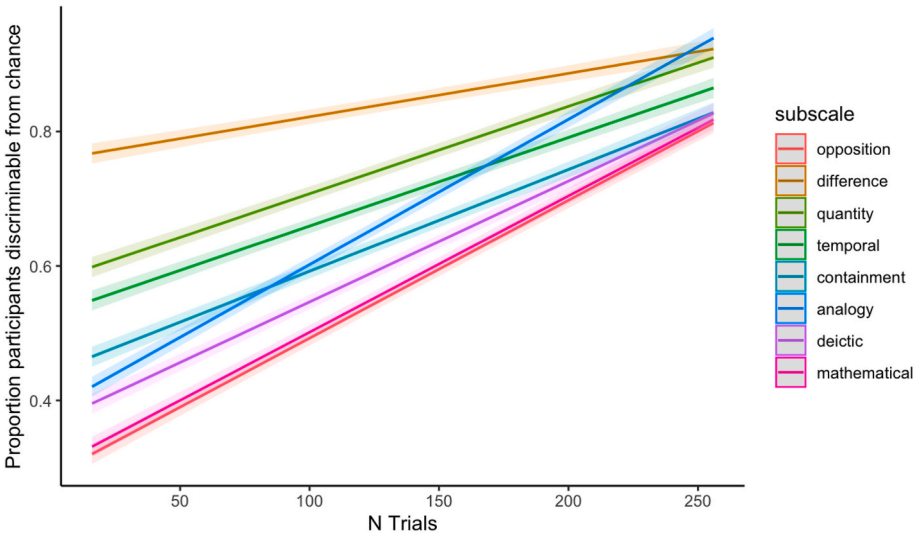

Fig. 5. The estimated marginal predictions for discriminability of RAI subscale scores from chance-level responding (i.e., a score of 0.5) as a function of increasing number of trials in the subscales.

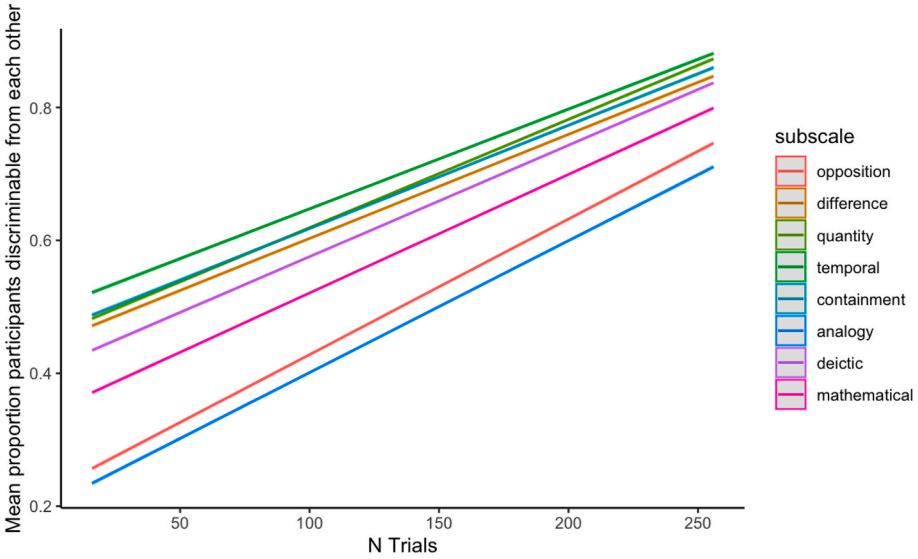

Fig. 6. Estimated marginal predictions for the proportion of discriminability for each of the subscales across increasing numbers of trials.

Table 4
Estimated split-half reliabilities for each subscale, as well as prophesized requirements for various split-half related criteria.

Scale	Split-half	Split-half if doubled	Trials needed for …		
			.8	.9	.99
Opposition	.45	.62	78	176	1963
Difference	.81	.90	15	34	372
Quantity	.67	.80	32	71	780
Temporal	.77	.87	19	43	473
Containment	.74	.85	22	51	557
Analogy	.35	.52	119	267	2942
Deictic	.54	.70	55	123	1349
Mathematical	.64	.78	36	81	891

Table 5
Scale properties of each of the RAI subscales in terms of their adherence to a Guttman structure.

Subscale	Mean proportion of Guttman errors
Opposition	.81
Difference	.90
Quantity	.83
Temporal	.82
Containment	.81
Analogy	.79
Deictic	.80
Mathematical	.80

Fig. 8 displays the ICCs for each subscale.

Finally, I plotted Test Characteristic Curves (TCCs) and Test Information Curves (TICs) for each subscale. The ideal TIC for a scale would be a straight line which intercepts the y-axis at a large value, indicating the test is equally and highly informative at every level of participant ability. Typically, however, TICs are shaped as bell curves; in such cases, fatter tails and high peaks are desirable. Fig. 9 plots the TCCs for each subscale against one another; Fig. 10 does the same for TICs. The TICs demonstrated that for all subscales, test information was highest (and therefore measurement error was lowest) for participants of below average ability. For participants of beyond average ability, the

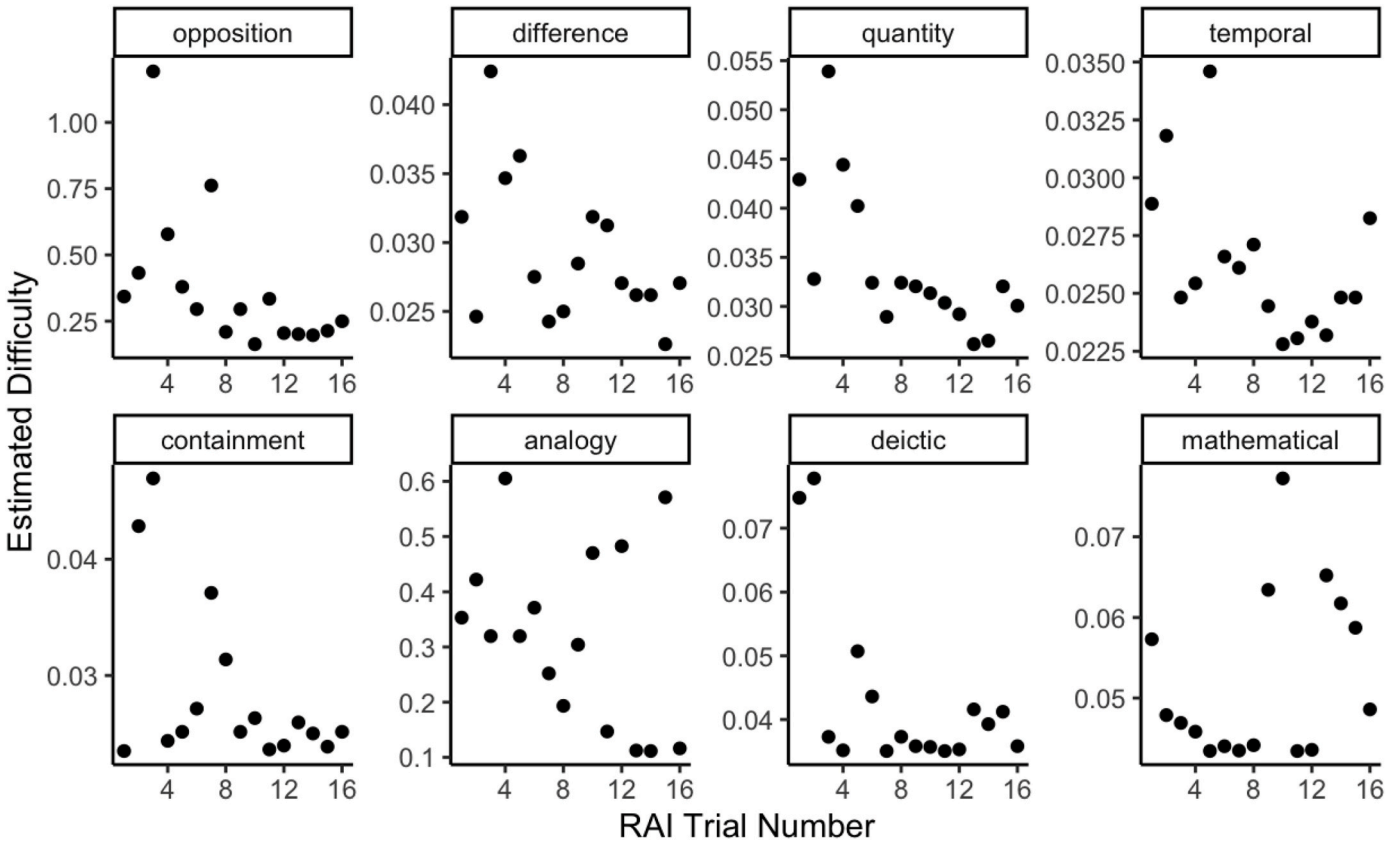

Fig. 7. Comparisons of theoretically predicted trial difficulties (in the form of RAI trial number) against estimated difficulty from IRT models for each of the 8 subscales.

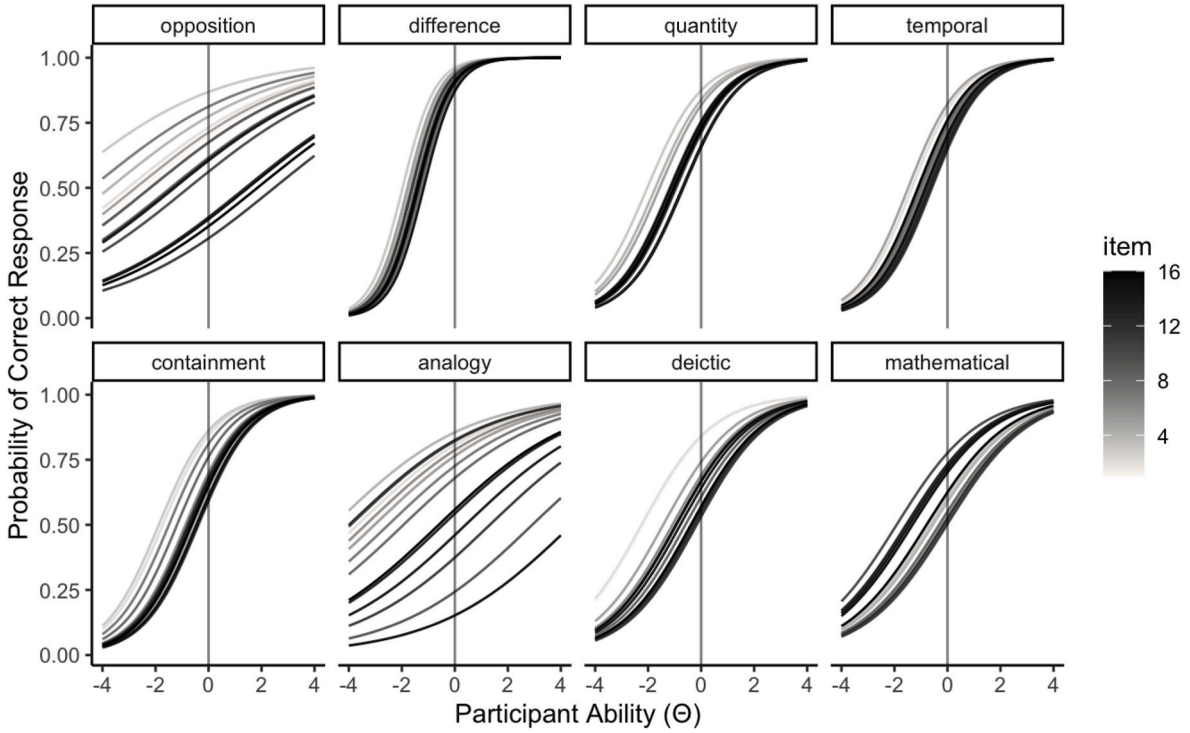

Fig. 8. Item Information Curves for each of the RAI subscales. The vertical line corresponds to zero on the x-axis, indicating an average ability participant. Darker colored lines indicate trials which are presented later in the RAI; in principle we should expect to see a clear gradient of light-to-dark as we move from left-to-right in the plot.

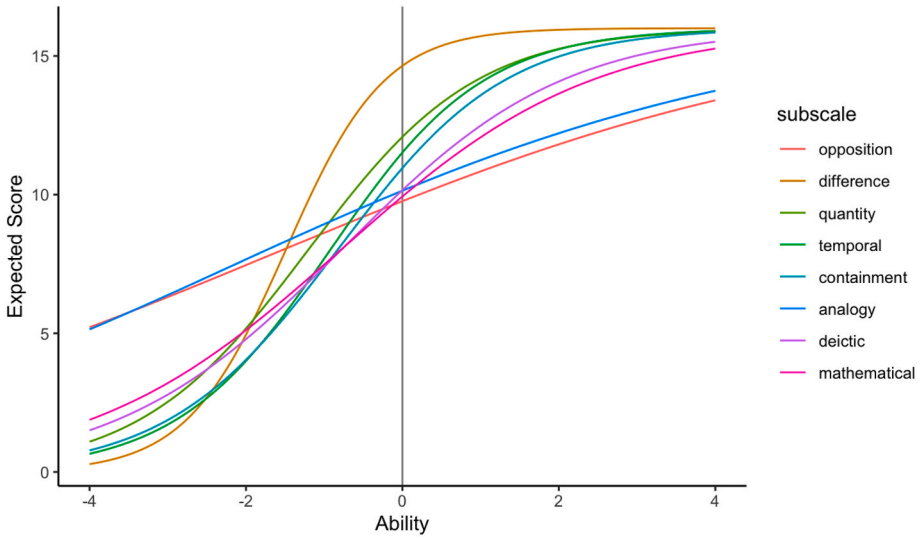

Fig. 9. Test Characteristic Curves for each of the 8 RAI subscales.

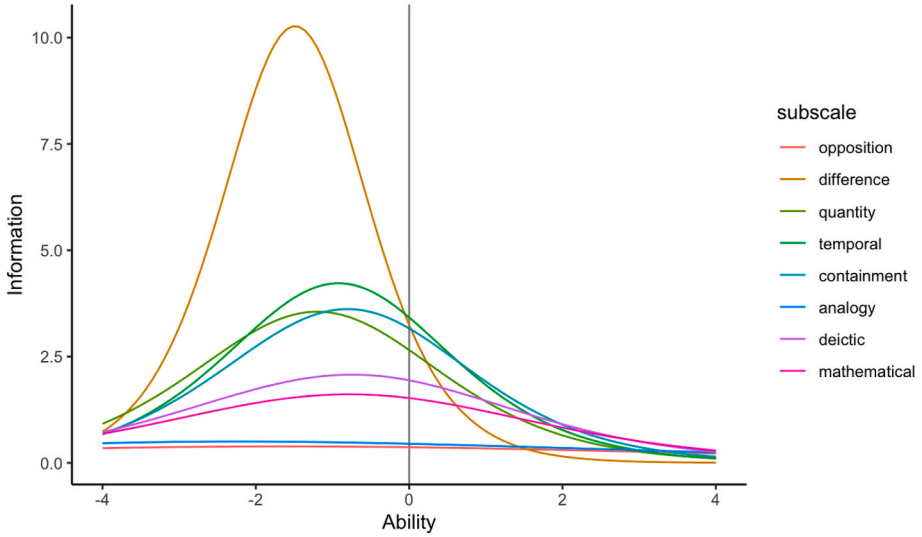

Fig. 10. Test Information Curves for each of the 8 RAI subscales.

subscales' informativeness quickly diminished.

4. Discussion

4.1. Overview of results

Good measurement is fundamental to good theory. For theory on relational responding, measures like the RAI are critical to advance our knowledge. The analyses above, however, suggest that the RAI has plenty of room for improvement. Perhaps the most critical finding is that the RAI subscales' precision at the individual-level is deeply lacking. The median CI width for scores across subscales ranged between 0.33 and 0.48 (excessively wide for a scale with a range of 0–1). In terms of inter-participant discriminability, a given participant was discriminable from, on average, 7–32% of other participants (depending on the subscale). This is a particularly critical issue for CBS, given the recent push for a more idiographic level of analysis (Hayes, Merwin, et al., 2021). Fortunately, individual-level precision can be improved by adding further trials to the subscale being completed. Of course, participants cannot simply complete trials in a relational responding task ad infinitum. What would benefit further research, then, would be to clarify in advance which relations are of interest, and to assess those specific relations in greater depth. Indeed, Figs. 4–6 provide a helpful illustration for future researchers on precisely how many trials would be needed for a given subscale to achieve a particular degree of individual-level performance.

Beyond individual-level precision, the RAI subscales exhibited variation in their split-half reliability. Whereas the difference, temporal, and containment subscales performed rather well even with the current 16 trial format, others such as analogy and opposition were very poor. Perhaps the biggest surprise is the poor performance of opposition, given that it is considered to be the basis for more complex types of relational responding (Mulhern, 2022). Notably, because opposition is the first subscale that is administered in the RAI, one explanation for this could be that participants are simply unfamiliar with the task environment during this initial subscale, leading to more inconsistency in their responding. I conducted a brief exploratory study to test this explanation, with 50 participants completing the opposition subscale and then

the difference subscale (i.e., as in the standard RAI), and 50 participants completing the difference subscale and then the opposition subscale.[1] However, split-half reliability in the opposition subscale did not differ significantly between these conditions (and in fact was slightly worse when opposition was completed second). As with individual-level precision, a solution to improve split-half reliability once again resides in increasing the number of trials in the task; Table 4 provides a convenient guide for future research in this regard.

Notably, violations of Guttman structure were observed very frequently across all subscales of the RAI. The subscales in general simply did not conform to their theoretically proposed structures. Looking further in terms of the fitted 1-PL models, all subscales except analogy and mathematical responding appeared to show a pattern where earlier trials were initially quite difficult, and then this difficulty sharply *decreased* in the middle stages of the subscales, with some increases towards the final few trials in some cases. At a more general level, the analogy and opposition subscales performed poorly in terms of their overall TCCs and TICs. The other six subscales were not ideal, providing limited information about individuals above average ability and generally being constrained to low levels of difficulty. The analogy and opposition subscales, however, simply provided little information at *any* level of participant ability.

4.2. Implications and future research

Beyond the need for increasing number of trials and a closer focus on Guttman structure, there are some further recommendations for use of the RAI that follow from this work. In terms of individual-level precision, researchers should carefully consider exactly how precise they require the specific subscales of interest to be. One method to do this may be to adopt the logic of the "smallest effect size of interest" (SESOI) approach to power analysis and sample size planning (Anvari & Lakens, 2021). In other words, a researcher planning to use the RAI to make inferences about individual-level effects may specify *a priori* the number of participants they would ideally like a given participant to be detectably different from, the desired median CI width, the desired discriminability from zero, or some combination of these features. With these in mind, researchers can design an RAI which can provide them with the specificity they desire. What is clear, as has been iterated across recent psychological literature, is that we must reject a "one-size-fits-all" approach to measurement and cater measurement decisions to specific research questions (Cummins, Hussey, & Spruyt, 2022; Flake et al., 2017).

Although increasing the number of trials may improve the measurement properties of the RAI, it is somewhat of a "brute-force" solution to the problem. One potentially more elegant alternative would be to improve the trial items themselves. The IRT analyses above provide clues as to how this can be done. For several of the subscales, the ICCs demonstrate that the trials in these subscales are generally the same (low) level of difficulty. Consequently, these subscales provide limited information about higher-ability individuals. Development of these subscales could therefore focus on adding further trials of greater difficulty to improve the coverage of these subscales. One other simple way to do this would be to adjust the trials to include more than two response options (e.g., by requiring the discrimination of the correct stimulus to a question such as "which stimulus is the biggest?"). Currently, chance-level responding will produce 50% correct responses in the task, which constrains both the overall range of possible scores to observe and the individual-level precision. By extending the number of possible responses, this would increase the variance of observed scores and improve the diagnosticity of the measure at the individual-level. Future work could also attempt to model guessing directly within the employed IRT models (San Martín et al., 2013).

In this paper, I used IRT primarily to *test* specific proposed properties of the RAI (e.g., adherence to a Guttman structure). However, IRT can also be used in a more iterative manner to *improve* measures (Embretson & Reise, 2013). Future research should certainly do this, and any insights or improvements in the measure may also have implications for the advancement of RFT. It should be noted that the strategies employed to date to implement incremental difficulty across trials in the RAI were derived from assumptions within RFT: namely, that additional derivations, a greater number of reversals of relations, etc. should be more difficult for participants (Cassidy et al., 2016; Colbert et al., 2017; Hayes et al., 2001). Given that these factors appeared to have little impact on the difficulty of RAI trials, these results signal the necessity to empirically assess the theoretical assumptions that are baked into such tasks.

Of course, there may be many reasons why these results were observed. One potential explanation could relate to the ordering of trials in the RAI. In most of the subscales, there was a clear pattern where the initial trials in the measure were most difficult, with difficulty then decreasing and plateauing until the final number of trials. RAI trials are presented in a fixed order, and each subscale is presented as a full block of trials. In progressing from one subscale to the next, participants may perform more poorly on the initial trials of the newly started subscale simply because there is a sudden change in the required relational response. As participants progress beyond these initial trials and become more acquainted with the type of relational problems with which they are presented, responding may then become more fluent, before the final most difficult trials then lead to an increase in difficulty once more. This idea could be tested quite easily by comparing the standard RAI to one where all trials are presented in a random sequence.

4.3. Considerations and limitations

Many of the analyses used to test the research questions in this paper are relatively novel to Contextual Behavioral Science and research on Relational Frame Theory generally. On the one hand, this is a strength of the paper; previously unused methods can naturally provide novel insights into the measures that we use. On the other hand, readers may question whether these methods are appropriate for CBS research. I believe the answer is yes. Although novel to CBS, Item Response Theory in general is a well-established, theoretically coherent alternative to Classical Test Theory (Embretson & Reise, 2013; Johansson et al., 2023). Indeed, its utility in improving scale items (as discussed in the previous section) represents a valuable addition not only for the development of the RAI, but also measures used within CBS (and beyond) more generally. In terms of the individual-level estimation approach used here, the same response applies; although not commonly employed within CBS, the estimation of (im)precision around individual participants' scores has been done extensively in literature on psychological assessment (e.g., Dudek, 1979). Indeed, recently Hussey (2020) applied a very similar approach to the Implicit Relational Assessment Procedure (commonly used within CBS). As an ACBS Task Force report (Hayes, Merwin, et al., 2021) recently recommended, CBS requires novel psychometric methods which can be used as quality standards for the assessment of the idiographic utility of measures; this is precisely what the method employed here (and by Hussey, 2020) achieves, and future research could endeavor to conduct similar analyses in other measures used within CBS for idiographic inferences.

In terms of the analysis of individual-level discriminability (from both chance responding and from other participants), several points of consideration arise based on the results reported here. Most critically is the question of what the "ideal" benchmark for these different forms of discriminability *ought to be*. As one reviewer noted, it would probably be unreasonable to suspect that every participant should be discriminable from every other participant, and we certainly should not expect that every participant should be discriminable from chance. Additionally, we

[1] Data and processing/analysis scripts for this exploratory study can be found on the Open Science Framework; https://osf.io/vu9aw/?view_only=eb4ec2dd710e4ce686e6a8733868f07d.

have no sense of how the current quality of the RAI compares to other measures of relational responding. It may well be the case that relative to other extant measures the RAI performs relatively well in this regard. This is, of course, an empirical question. With all of this said, however, it is worth reflecting on what we as a community might consider as an acceptable degree of discriminability for our instruments. This will of course vary depending on the context in which the measure is to be used and the specific goals of the researcher(s) using the measure.

One potential avenue for identifying more meaningful benchmarks for acceptable degrees of discriminability could be to examine the quality of other measurement instruments from other fields. In the context of psychological assessment generally, it is important to note that the most commonly-used measures (e.g., IQ tests, neuropsychological batteries) exhibit rather high reliabilities (e.g., $r = 0.97$ for the WISC IQ test, $r = 0.80$ for the RBANS neuropsychological battery; Cheng et al., 2011; Gygi et al., 2017). These high reliabilities will necessarily be associated with high Standard Errors of Measurement (and by extension, narrow confidence intervals around scores individual-level scores; Dudek, 1979). Extant data from these assessments (and similar) could be reused to calculate the mean discriminability of participants from one another to serve as an initial starting point to compare the RAI against. Of course, this will be an imperfect comparison; the distribution of scores in the subscales of the RAI are demonstrably not Gaussian, which will necessarily have an impact on discriminability of scores at different points in the distribution. Regardless of these specifics, it would be remiss to assume that the current state of the measure is its best possible iteration, particularly given its poor performance on more traditional psychometrics.

5. Conclusion

The results from this study paint a clear picture: as far as the measurement properties of the RAI is concerned, there is room for improvement. RAI scores are not particularly diagnostic at the individual level, and the group-level psychometric properties of the task are suboptimal. Fortunately, these problems ultimately are solvable in a clear manner, and this paper has provided a roadmap for researchers on how to achieve certain levels of precision and/or psychometric integrity when using the measure in future. More problematically, however, a central assumption of RAI subscales (i.e., that they conform to a Guttman structure) does not appear to be strongly evident in any of the RAI subscales. The methods of IRT provide an excellent set of tools for refining the scale in the future in this regard. However, the observed violations of Guttman structure in the current form of the RAI present a challenge to the conception of complexity in relational responding which drove the development of this scale ordering in the first place. A greater consideration of measurement issues may provide greater insight not only into the utility of our measures, but the validity of the theoretical assumptions upon which they are based.

Author note

JC, Department of Experimental Clinical- and Health Psychology, Ghent University and Institute of Psychology, University of Bern. Funding for this research was provided by FWO grant 1202624N to JC and Ghent University Grant BOF22/MET_V/002 to Jan De Houwer. Correspondence concerning this article should be sent to jamie.cummins@ugent.be. I thank Arianna Zanatta, Maura Nevejans, and Jan De Houwer for comments on earlier drafts of this manuscript.

Ethical approval was obtained through the Ghent University Ethics Committee, approval number 2020/74. Informed consent was obtained from all participants prior to participation.

Declaration of competing interest

The author has no conflict of interest to declare.

References

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

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology, 96*, Article 104159. https://doi.org/10.1016/j.jesp.2021.104159

Austin, P. C., & Hux, J. E. (2002). A brief note on overlapping confidence intervals. *Journal of Vascular Surgery, 36*(1), 194–195. https://doi.org/10.1067/mva.2002.125015

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, C., Garcia, Y., Brothers, L., Mahoney, A., Rancourt, R., & Andrews, M. (2023). A systematic review of the impact of derived relational responding technology in raising intelligence scores. *Psychological Record*. https://doi.org/10.1007/s40732-023-00546-0

Brown, W. (1910). Some experimental results in the correlation of mental Abilities1. *British Journal of Psychology, 1904-1920, 3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Cassidy, S., Roche, B., Colbert, D., Stewart, I., & Grey, I. M. (2016). A relational frame skills training intervention to increase general intelligence and scholastic aptitude. *Learning and Individual Differences, 47*, 222–235. https://doi.org/10.1016/j.lindif.2016.03.001

Cassidy, S., Roche, B., & Hayes, S. C. (2011). A relational frame training intervention to raise intelligence quotients: A pilot study. *Psychological Record, 61*(2), 173–198. https://doi.org/10.1007/BF03395755

Cassidy, S., Roche, B., & O'Hora, D. (2010). Relational frame theory and human intelligence. *European Journal of Behavior Analysis, 11*(1), 37–51. https://doi.org/10.1080/15021149.2010.11434333

Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology, 1*(1), 161–187. https://doi.org/10.1080/24709360.2017.1396742

Cheng, Y., Wu, W., Wang, J., Feng, W., Wu, X., & Li, C. (2011). Reliability and validity of the repeatable battery for the assessment of neuropsychological status in community-dwelling elderly. *Archives of Medical Science : AMS, 7*(5), 850–857. https://doi.org/10.5114/aoms.2011.25561

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*(4), 404–413. https://doi.org/10.2307/2331986

Colbert, D., Dobutowitsch, M., Roche, B., & Brophy, C. (2017). The proxy-measurement of intelligence quotients using a relational skills abilities index. *Learning and Individual Differences, 57*, 114–122. https://doi.org/10.1016/j.lindif.2017.03.010

Colbert, D., Malone, A., Barrett, S., & Roche, B. (2019). The relational abilities Index+: Initial validation of a functionally understood proxy measure for intelligence. *Perspectives on Behavior Science, 43*(1), 189–213. https://doi.org/10.1007/s40614-019-00197-z

Colbert, D., Tyndall, I., Roche, B., & Cassidy, S. (2018). Can SMART training really increase intelligence? A replication study. *Journal of Behavioral Education, 27*(4), 509–531. https://doi.org/10.1007/s10864-018-9302-2

Cornell Statistical Consulting Unit. (2008). *Overlapping confidence Interval and statistical significance. Studylib.Net*. https://studylib.net/doc/18221119/overlapping-confidence-interval-and-statistical-significance.

Cummins, J., & Hussey, I. (2023). *The precision of 6 implicit measures at the individual level.* Unpublished manuscript.

Cummins, J., Hussey, I., & Spruyt, A. (2022). The role of attitude features in the reliability of IAT scores. *Journal of Experimental Social Psychology, 101*, Article 104330. https://doi.org/10.1016/j.jesp.2022.104330

Delabie, M., Cummins, J., Finn, M., & De Houwer, J. (2022). Differential Crel and Cfunc acquisition through stimulus pairing. *Journal of Contextual Behavioral Science, 24*, 112–119. https://doi.org/10.1016/j.jcbs.2022.03.012

Cummins, J., Nevejans, M., Colbert, D., & De Houwer, J. (2023). On the structure of relational responding. *Journal of Contextual Behavioral Science, 27*, 16–25.

Dixon, M. R., Belisle, J., Hayes, S. C., Stanley, C. R., Blevins, A., Gutknecht, K. F., Partlo, A., Ryan, L., & Lucas, C. (2021). Evidence from children with autism that derived relational responding is a generalized operant. *Behavior Analysis in Practice, 14*(2), 295–323. https://doi.org/10.1007/s40617-020-00425-y

Dixon, M. R., Belisle, J., McKeel, A., Whiting, S., Speelman, R., Daar, J. H., & Rowsey, K. (2017a). An internal and critical review of the PEAK relational training system for children with autism and related intellectual disabilities: 2014–2017. *The Behavior Analyst, 40*(2), 493–521. https://doi.org/10.1007/s40614-017-0119-4

Dixon, M. R., Belisle, J., Stanley, C. R., Speelman, R. C., Rowsey, K. E., Kime, D., & Daar, J. H. (2017b). Establishing derived categorical responding in children with disabilities using the PEAK-E curriculum. *Journal of Applied Behavior Analysis, 50*(1), 134–145. https://doi.org/10.1002/jaba.355

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86*, 335–337. https://doi.org/10.1037/0033-2909.86.2.335

Edwards, A. W. F. (1960). The meaning of binomial distribution. *Nature, 186*(4730). https://doi.org/10.1038/1861074a0. Article 4730.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science, 16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Finn, M., & De Houwer, J. (2021). The selective action of Cfunc control. *Journal of the Experimental Analysis of Behavior, 116*(3), 314–331. https://doi.org/10.1002/jeab.717

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Gygi, J. T., Hagmann-von Arx, P., Schweizer, F., & Grob, A. (2017). The predictive validity of four intelligence tests for school grades: A small sample longitudinal study. *Frontiers in Psychology, 8*, 375. https://doi.org/10.3389/fpsyg.2017.00375

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational Frame theory: A post-skinnerian account of human language and cognition*. Springer.

Hayes, S. C., Law, S., Assemi, K., Falletta-Cowden, N., Shamblin, M., Burleigh, K., Olla, R., Forman, M., & Smith, P. (2021a). Relating is an operant: A fly over of 35 Years of RFT research. *Perspectivas em Análise do Comportamento, 12*(1). https://doi.org/10.18761/PAC.2021.v12.RFT.02. Article 1.

Hayes, S. C., Merwin, R. M., McHugh, L., Sandoz, E. K., A-Tjak, J. G. L., Ruiz, F. J., Barnes-Holmes, D., Bricker, J. B., Ciarrochi, J., Dixon, M. R., Fung, K. P.-L., Gloster, A. T., Gobin, R. L., Gould, E. R., Hofmann, S. G., Kasujja, R., Karekla, M., Luciano, C., & McCracken, L. M. (2021b). Report of the ACBS Task Force on the strategies and tactics of contextual behavioral science research. *Journal of Contextual Behavioral Science, 20*, 172–183. https://doi.org/10.1016/j.jcbs.2021.03.007

Hayes, J., & Stewart, I. (2016). Comparing the effects of derived relational training and computer coding on intellectual potential in school-age children. *British Journal of Educational Psychology, 86*(3), 397–411. https://doi.org/10.1111/bjep.12114

Hughes, S., Ye, Y., & De Houwer, J. (2019). Evaluative conditioning effects are modulated by the nature of contextual pairings. *Cognition & Emotion, 33*(5), 871–884. https://doi.org/10.1080/02699931.2018.1500882

Hussey, I. (2020). *The Implicit Relational Assessment Procedure is not suitable for individual use*. PsyArXiv. https://doi.org/10.31234/osf.io/w2ygr

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

Johansson, M., Preuter, M., Karlsson, S., Möllerberg, M.-L., Svensson, H., & Melin, J. (2023). *Valid and reliable? Basic and expanded Recommendations for psychometric Reporting and quality assessment. OSF preprints*. https://doi.org/10.31219/osf.io/3htzc

Knol, M. J., Pestman, W. R., & Grobbee, D. E. (2011). The (mis)use of overlap of confidence intervals to assess effect modification. *European Journal of Epidemiology, 26*(4), 253–254. https://doi.org/10.1007/s10654-011-9563-8

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology, 61*(4), 281–288. https://doi.org/10.1037/cap0000236

Lord, F. M. (2012). *Applications of item response theory to practical testing problems*.

May, R. J., Tyndall, I., McTiernan, A., Roderique-Davies, G., & McLoughlin, S. (2022). The impact of the SMART program on cognitive and academic skills: A systematic review and meta-analysis. *British Journal of Educational Technology, 53*(5), 1244–1261. https://doi.org/10.1111/bjet.13192

McLoughlin, S., Tyndall, I., & Pereira, A. (2020). Convergence of multiple fields on a relational reasoning approach to cognition. *Intelligence, 83*, Article 101491. https://doi.org/10.1016/j.intell.2020.101491

McLoughlin, S., Tyndall, I., Pereira, A., & Mulhern, T. (2021). Non-verbal IQ gains from relational operant training explain variance in educational attainment: An active-controlled feasibility study. *Journal of Cognitive Enhancement, 5*(1), 35–50. https://doi.org/10.1007/s41465-020-00187-z

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods, 52*(6), 2287–2305. https://doi.org/10.3758/s13428-020-01398-0

Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311–314. https://doi.org/10.1177/014662169401800402

Montoya-Rodríguez, M. M., Molina, F. J., & McHugh, L. (2017). A review of relational frame theory research into deictic relational responding. *Psychological Record, 67*(4), 569–579. https://doi.org/10.1007/s40732-016-0216-x

Mooney, C. Z., Duval, R. D., & Duvall, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. SAGE.

Mulhern, T. (2022). Relational frames of opposition and distinction. In T. Mulhern (Ed.), *Relational Frame theory: Made simple* (pp. 71–107). Springer International Publishing. https://doi.org/10.1007/978-3-031-19421-4_4.

Perez, W. F., Kovac, R., Nico, Y. C., Caro, D. M., Fidalgo, A. P., Linares, I., de Almeida, J. H., & de Rose, J. C. (2017). The transfer of Crel contextual control (same, opposite, less than, more than) through equivalence relations. *Journal of the Experimental Analysis of Behavior, 108*(3), 318–334. https://doi.org/10.1002/jeab.284

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Paedagogiske Institut.

Raykov, T. (2008). Alpha if item deleted: A note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology, 61*(2), 275–285. https://doi.org/10.1348/000711007X188520

Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula, III. *Journal of Educational Psychology, 32*, 61–66. https://doi.org/10.1037/h0061781

San Martín, E., Rolin, J.-M., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika, 78*(2), 341–379. https://doi.org/10.1007/s11336-013-9322-8

Schmukle, S. C. (2023). *Unbiased confidence intervals for individual scores in psychological testing: The rescaled estimated true score (RETS) approach*. unpublished manuscript.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920, 3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Tan, S. H., & Tan, S. B. (2010). The correct interpretation of confidence intervals. *Proceedings of Singapore Healthcare, 19*(3), 276–278. https://doi.org/10.1177/201010581001900316

Thirus, J., Starbrink, M., & Jansson, B. (2016). Relational frame theory, mathematical and logical skills: A multiple exemplar training intervention to enhance intellectual performance. *International Journal of Psychology and Psychological Therapy/Revista Internacional de Psicologia y Terapia Psicologica, 16*(2), 141–155.

Wand, M. (2021). *KernSmooth: Functions for kernel smoothing supporting* (Vol. 1995). Wand & Jones (2.23-20) [R].

Wang, Z., & Osterlind, S. J. (2013). Classical test theory. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 31–44). SensePublishers. https://doi.org/10.1007/978-94-6209-404-8_2.