

Check for
updates

Citation: Cheyette, S. J., & Piantadosi, S. T. (2024). Response to Difficulty Drives Variation in IQ Test Performance. *Open Mind: Discoveries in Cognitive Science*, 8, 265–277. https://doi.org/10.1162/opmi_a_00127

DOI:
https://doi.org/10.1162/opmi_a_00127

Supplemental Materials:
https://doi.org/10.1162/opmi_a_00127

Received: 9 May 2023
Accepted: 6 February 2024

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Samuel J. Cheyette
cheyette@mit.edu

Copyright: © 2024
Massachusetts Institute of Technology
Published under a Creative Commons
Attribution 4.0 International
(CC BY 4.0) license



The MIT Press

REPORT

Response to Difficulty Drives Variation in
IQ Test PerformanceSamuel J. Cheyette¹  and Steven T. Piantadosi² ¹Department of Brain & Cognitive Sciences, MIT²Department of Psychology, UC Berkeley**Keywords:** IQ, Bayesian modeling, individual differences**ABSTRACT**

In a large ($N = 300$), pre-registered experiment and data analysis model, we find that individual variation in overall performance on Raven's Progressive Matrices is substantially driven by differential strategizing in the face of difficulty. Some participants choose to spend more time on hard problems while others choose to spend less and these differences explain about 42% of the variance in overall performance. In a data analysis jointly predicting participants' reaction times and accuracy on each item, we find that the Raven's task captures at most half of participants' variation in time-controlled ability (48%) down to almost none (3%), depending on which notion of ability is assumed. Our results highlight the role that confounding factors such as motivation play in explaining individuals' differential performance in IQ testing.

INTRODUCTION

Intelligence tests are central to many areas of applied and theoretical psychology, however the question of what IQ tests measure has been debated for decades (Ceci, 1996; Flynn, 1987; Gould, 1996; Jensen, 1998; Richardson, 2002; Mackintosh, 2011; Mensh & Mensh, 1991; Schönemann, 1983). Large and robust effects of coaching, schooling, practice, and pay (Briggs, 2001; Brinch & Galloway, 2012; Cahan & Cohen, 1989; Cliffordson & Gustafsson, 2008; Duckworth et al., 2011; Kulik, Bangert-Drowns, et al., 1984; Kulik, Kulik, et al., 1984; Powers, 1993; Ritchie & Tucker-Drob, 2018) on IQ test performance demonstrate that individual experiences and incentives affect test outcomes, independent of intellectual ability. Experiments that manipulate the amount of reward provided to participants based on performance find substantial, robust effects on test performance. Figure 1 shows data replotted from Duckworth et al. (2011)'s meta-analysis of prior pay manipulations, showing the overall effect of pay (left) as well as the pay broken down by coding of reward size (color). This illustrates a robust effect (Hedges $g \approx 1$ or roughly 15 IQ points in the best cases) that appears sensitive to the amount of extrinsic reward.

While these results show that individuals will change their performance in response to external incentives, they do not demonstrate that people vary *intrinsically* in the effort and strategies they bring into testing situations. This possibility is important for understanding the construct validity of IQ tasks because individual variation in intrinsic effort or strategy would masquerade as differences in ability. Specifically, the speed-accuracy tradeoff that each individual decides upon should be expected to impact their performance. This possibility was

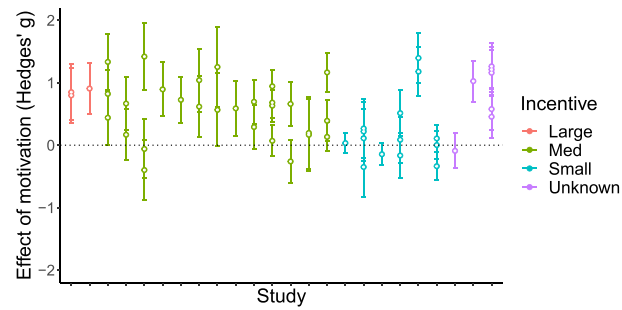


Figure 1. A visualisation of data from Duckworth et al. (2011) showing the effect size of a pay manipulation on IQ tasks (y-axis) across studies (x-axis), broken down by reward size (color). This robustly shows effect of pay manipulations on test outcomes.

highlighted by early experimental psychologists like Thurstone (1937), who articulated the inevitable tradeoff between accuracy and time in testing situations. Figure 2A shows a sketch of the relationship between accuracy (“probability of success”), time, and difficulty highlighted in Thurstone (1937), capturing the idea that difficult items will tend to take more time to achieve a high probability of success. The interrelationship highlights that a finding that time investment differs between individuals is relevant to measuring ability: a person’s ability—perhaps quantified as difficulty at a fixed level of accuracy and RT—cannot be read off of their performance if individuals differ in time investment. Figure 2B and C illustrate this point: assuming a fixed level of ability across a population, natural variation in the maximum time participants’ allot to a question (2b) could lead to substantial variation in Raven’s scores (2c).

For this reason, it is unclear to what extent the positive manifold reported in intelligence research since Spearman (1904) might be explained not through a shared component of intellectual capacity, but through a shared component of effort or time investment in testing tasks. This idea has received surprisingly little attention in psychology’s IQ debates (Goldhammer & Entink, 2011; Goldhammer et al., 2015; Scherer et al., 2015; Tate, 1948; Thissen, 1976). A notable exception is the work of Thissen: following Furneaux (1973), Thissen (1983) showed a correlation of $R = 0.94$ between slowness and outcome on Raven’s tasks (Raven, 2000; Raven et al., 1989). Thissen concluded that “working slowly and carefully is strongly related to the probability of responding correctly, and what is measured is largely slowness.”

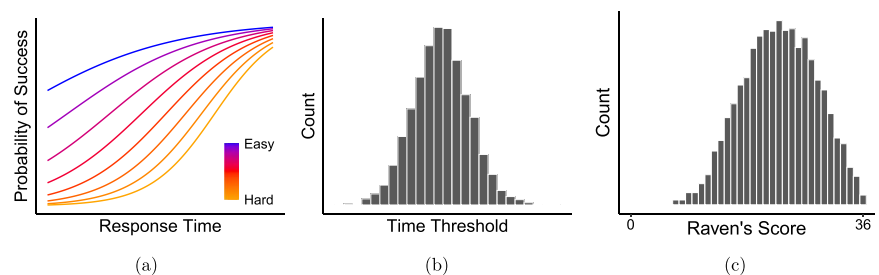


Figure 2. An illustration of the potential issue with uncontrolled variation in response times. (A) The conceptual tradeoff between an item’s difficulty (colors), the time taken on the task (x-axis), and the probability of responding accurately (y-axis). (B) Simulated participants with variation in their maximum time investment on a given question. (C) Simulated Raven’s scores given only the RT-accuracy curves and natural variation in participants’ response time thresholds shown in the other two panels.

It is important to distinguish the idea that performance depends on slow, careful, sustained attention and effort from another popular hypothesis in psychometrics. A considerable body of work has examined how *general* processing speed fits into the picture of psychometric *g* (Bates & Stough, 1997, 1998; Carroll, 1993; Evans & Deary, 1994; Deary & Stough, 1996; Grudnik & Kranzler, 2001; Jensen, 1982, 1985, 2006; Kyllonen & Zu, 2016; Nettelbeck, 1998; Neubauer, 1990; Sheppard & Vernon, 2008; Vernon, 1983). Such work typically quantifies each individual's processing speed on simple perceptual tasks like responding quickly to a light stimulus as in Hick (1952). This hypothesis is distinct from the idea explored by Thissen (1983) because the time spent on each question is dependent on higher-level cognitive processes than those involved in perceptual tasks. A considerable literature in test theory (Gulliksen, 1950; van der Linden, 2009) has examined the relationship between time-limits and performance broadly in testing situations (e.g., Bridgeman et al., 2004; Davidson & Carroll, 1945; Kyllonen & Zu, 2016; Rindler, 1979). This has resulted in proposed measures in psychometrics that combine speed and accuracy (Liesefeld et al., 2015; Liesefeld & Janczyk, 2019; Townsend & Ashby, 1983; Vandierendonck, 2018), or jointly analyze both (Bolsinova et al., 2017; De Boeck & Jeon, 2019; Entink et al., 2009; Kyllonen & Zu, 2016; van der Linden & Fox, 2016; van der Linden et al., 2010). Such tradeoffs are even attested in other species (Bogacz et al., 2010; Chittka et al., 2009; Goldhammer, 2015; Heitz, 2014; Heitz & Schall, 2012; Luce, 1986; Wickelgren, 1977). Yet, in the context of IQ testing, it is standard to compute overall accuracy, and not even look at timing patterns, much less control them.

Here, we build on Thissen (1983) to examine the relationship between individuals' response times across questions (reflecting strategy and effort) and overall test performance in a Raven's task. We aim to update these results with modern methods, including Bayesian data analyses that control for items and participants, large sample sizes, and pre-registered experimental designs and data analysis, and then interpret these findings in the context of the construct validity for these tasks. Several behavioral patterns are possible as items become more difficult throughout a Raven's task: (i) participants could spend more time on *more* difficult items, likely exerting greater effort in order to achieve high accuracy; (ii) participants could spend *less* time on difficult items, perhaps meta-cognitively realizing that a given problem is out of reach; or (iii) participants could be relatively insensitive to item difficulty, perhaps allocating time or effort equally throughout the test. Crucially, participants may show different patterns of behavior across questions, and our analysis aims to determine how variability in these patterns affects their overall score.

EXPERIMENT

Method

We pre-registered an experiment where 300 participants took an online version of Raven's Progressive Matrices (Raven, 2000) in September of 2022.¹ The experiment was run on Prolific, which has been found to yield higher-quality data than other online platforms (Peer et al., 2022). As is standard for this task, participants were told to complete as many of the items as they could in the maximum time of 40 minutes. Participants received compensation of \$7.50 for completion of the task. They were given instructions adapted from the 1988 Advanced Progressive Matrices manual (Raven & Court, 1988) for use in an online study. Unlike standard analyses of this task which focus on overall accuracy, we recorded response time for each

¹ The pre-registration, along with data and analysis, can be found at <https://osf.io/9rz2v/>.

item. These response times reflect either how long it took participants to find a solution, or how long they were willing to spend on a given item before moving on. Following our pre-registration plan, which was determined through a smaller pilot experiment on an independent pool of participants, we removed participants whose median response time was less than 10 seconds. This left 276 total participants in our main analysis. We z-scored RT across all participants in order to use a standardized scale, but maintain intrinsic variation between individuals. We also collected data on participants' demographics and socioeconomic status (e.g., income and education), and asked participants to report how many questions they thought they correctly answered.

Results

Aggregate response times are shown in Figure 3A, which show the RT for each item throughout the task, grouped by accuracy. Participants tended to spend more time on difficult (later) questions, but this effect is primarily driven by those who answer correctly: participants who are incorrect on later questions don't tend to spend more time on them. Differential time investment on hard questions hints that individuals may tend to be inaccurate when they choose to invest less time in a problem. One way to see whether subjects participant—and whether any variation is associated with accuracy—is to run a regression *within* each participant predicting their RT from the item number, using item number as a proxy for difficulty. Figure 3C and D shows these coefficients for each subject (y-axis) plotted against their overall task performance (x-axis). Participants who performed well tended to spend more time working on the later (harder) questions.

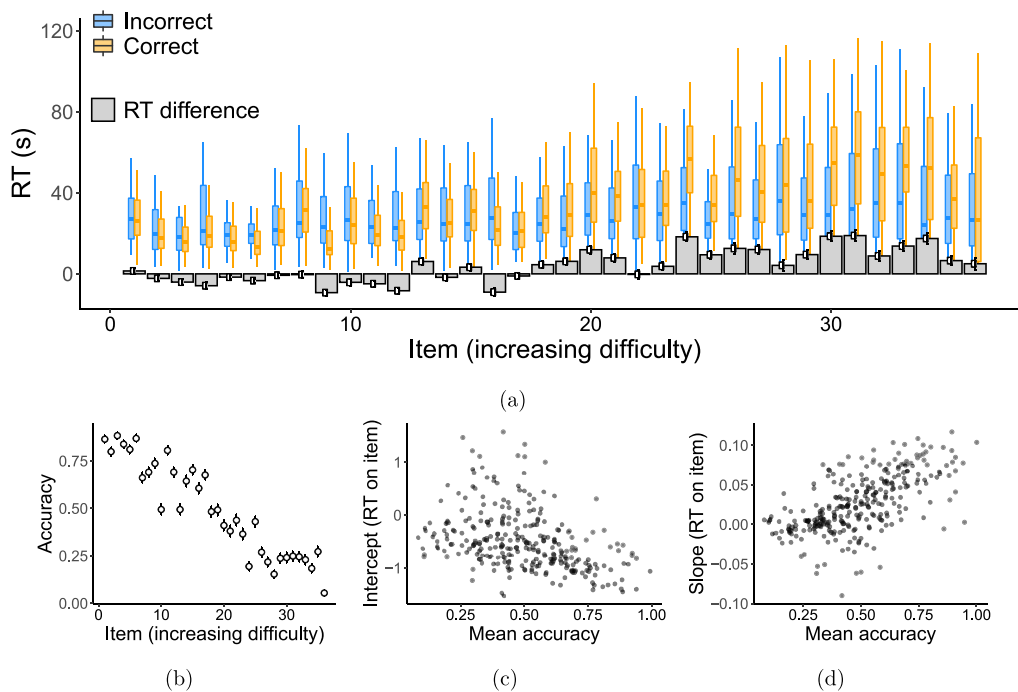


Figure 3. (A) Boxplots showing the distribution of response times to each question, grouped by accuracy (color). The gray bars depict the mean response time difference between correct and incorrect respondents. (B) Validation of the task, showing generally decreasing accuracy for harder items. (C–D) Inferred intercepts (C) and slopes (D) in a regression item (e.g., difficulty) and RT, separately for each subject, plotted against participants' mean accuracy across Raven's problems.

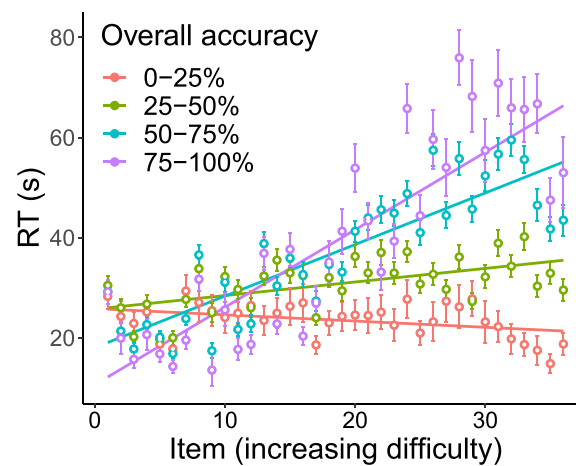


Figure 4. Response time (y -axis) as a function of problem number (x -axis), which is treated as a proxy for difficulty. Participants are grouped (color) according to their overall accuracy on the test. Participants who performed well were the ones who invested the most time on the harder problems.

An aggregate view of this effect is shown in Figure 4, which plots the relationship between RT and accuracy for participants, broken down by their overall accuracy. This figure paints a clear picture that those who performed well on the task tended to spend more time on the harder questions. The effect size between groups is huge: the best-performing quartile of participants spend approximately *three times* as long on the hard questions as the lowest-performing quartile. We emphasize that everyone was given the same instructions on the same task, so these differences represent *intrinsic* variation in how individuals approach the task. Individual subject plots can be found in the SI, and demonstrate a similar pattern.

Both intercepts and slopes are statistically correlated with overall Raven's ($R = -0.34$, $p < 0.001$ and $R = 0.65$, $p < 0.001$, respectively). Partialing the variance in overall performance between slope and intercept coefficients, we find Raven's score much more likely reflects response to difficulty (slope partial $R^2 = 0.36$) as opposed to average amount of overall time spent (intercept partial $R^2 = 0.01$). This indicates that these differences in slope matter to overall performance and thus the difficulty-time slope confounds Raven's measures which use overall performance. Following our preregistration plan, we also quantified variation in subject responses to difficulty by comparing two regression models that predicted RT: one where slopes by item were allowed to vary by subject and one where they were not. Both regressions included coefficients for item and accuracy. This revealed strong evidence in favor of the model that varied slopes by participant (ELPD difference 627.9 with a standard error of 38.5), providing further statistical support to the idea that individuals in a Raven's task respond differently to difficulty.

These differing slopes raise the natural question of how individuals might have performed *if* they had allocated time differently. Such a counterfactual is a step towards quantifying "ability" because it targets a subject's potential—what they could do—rather than what they happened to decide to do in our task. However, it is only a partial step towards ability because it leaves other factors like motivation and coaching uncontrolled. Following our pre-registration plan, we constructed a joint, fully Bayesian data analysis of accuracy, RT, and the latent difficulty of each item. One way this model differs from the previous regression is that the latent difficulty of each item is assumed to affect response time, rather than item number as above

(which is only imperfectly correlated with difficulty). By including RT, this model goes beyond recent item response theory models of the same task (Bürkner, 2020; Myszkowski & Storme, 2018); it differs from Thissen (1983) in that it uses a Bayesian analysis (Fox, 2010) that is hierarchical, allowing us to extract confidence in each individual subject and item parameter, while optimally reducing estimation noise through the use of partial pooling (Gelman & Hill, 2006). The model predicted the z-scored RT of the s 'th subject on the i 'th item, R_{si} as

$$R_{si} \sim \text{Normal}(\beta_0 + \beta_{0s} + (\beta_1 + \beta_{1s}) \cdot d_i, \sigma) \tag{1}$$

where d_i is the latent difficulty of the i 'th item. Here, β_0 and β_1 are the overall subject intercept and slope, which are given $\text{Normal}(0,3)$ priors. β_{0s} and β_{1s} are the s 'th subject's adjustments to those means, which are assumed to follow $\text{Normal}(0, \nu)$ for $\nu \sim \text{Exponential}(1)$. The item difficulties were given a prior $d_i \sim \text{Normal}(0,1)$. The standard deviation of response times was given prior $\sigma \sim \text{Exponential}(0.1)$.

Simultaneously with (1), the probability of responding correctly for subject s on item i (P_{si}) was modeled in a hierarchical logistic setup, such that

$$\text{logit}(P_{si}) = ((\gamma_0 + \gamma_{s0}) + (\gamma_1 + \gamma_{s1}) \cdot R_{si} + (\gamma_2 + \gamma_{s2}) \cdot d_i + \lambda_1 \cdot \beta_{0s} + \lambda_2 \cdot \beta_{1s} + \lambda_3 \cdot \beta_{0s} \cdot \beta_{1s}), \tag{2}$$

where:

$\gamma_0 + \gamma_{s0}$ = subject accuracy intercept

$\gamma_1 + \gamma_{s1}$ = effect of response time (R_{si}) on accuracy

$\gamma_2 + \gamma_{s2}$ = effect of item difficulty (d_i) on accuracy

λ_1 = effect of overall time investment (β_{0s}) on accuracy

λ_2 = effect of RT response to difficulty (β_{1s}) on accuracy

λ_3 = interaction between overall time investment (β_{0s}) and response to difficulty (β_{1s}) on accuracy

Response accuracy was then distributed according to $A_{si} \sim \text{Bernoulli}(P_{si} + \frac{1}{8} \cdot (1 - P_{si}))$, where the $\frac{1}{8} \cdot (1 - P_{si})$ term represents the probability of correctly answering a question by guessing randomly.

Here, $\gamma_0, \gamma_1, \gamma_2, \lambda_1, \lambda_2, \lambda_3$ are group mean parameters and were given $\text{Normal}(0,3)$ priors. The γ_s are individual subject parameters that, as with the β_s parameters, were drawn from $\text{Normal}(0, \nu)$ for $\nu \sim \text{Exponential}(1)$. When combined with (1), this form of model can be thought of as inferring latent participant measurements, β_{0s} and β_{1s} , that characterize how each person responds to difficulty, which are then used as predictors of accuracy, in addition to the RT on each item R_{si} . The net effect is that the other accuracy predictors (e.g., $\gamma_{0s}, \gamma_{1s}, \gamma_{2s}$) are then controlled for the patterns of response to difficulty apparent in RT. This hierarchical setup allows each subject estimate to be informed by the others, but also permits individual variation through the subject-specific parameters. The model was run using Stan (Carpenter et al., 2017), with 4 chains and 5,000 samples of its NUTS sampler (Hoffman & Gelman, 2014). Convergence was assessed using traceplots and \hat{R} values, which were less than 1.01.

Figure 5 shows the inferred individual parameters from this model as a function of each subject's raw accuracy on the Raven's task (x-axis). This provides several intuitive checks that the model is working appropriately—for example, Figure 5C shows that participants tend to be less accurate on more difficult problems since these values are negative. Figure 5E, giving the RT response to difficulty, replicates the analyses above to show that high-performing

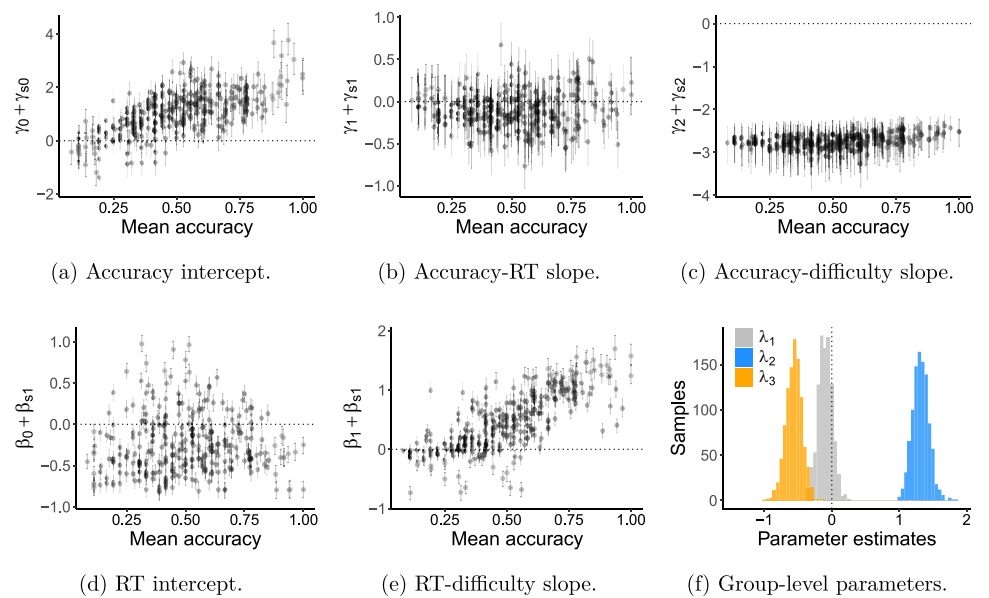


Figure 5. Model parameters predicting RT and accuracy. (A–E) Each point represents a single subject, and their parameter value (y -axis) is plotted against their overall Raven’s accuracy (x -axis). (F) Group-level posterior estimates for λ_1 , λ_2 , and λ_3 .

participants also tended to spend more time on more difficult problems. There are also many participants who had negative or essentially zero difficulty slopes for RT, meaning that they do not spend more time on harder problems. These people tended to perform least well overall. However, the RT intercept (time at average difficulty) in Figure 5D was relatively unrelated to overall correctness, showing that the effects are mostly about response to difficulty rather than starting time investment. Interestingly, the participants with the highest performance overall did *not* show better accuracy slopes (Figure 5B), meaning that their accuracy-per-time-invested did not improve faster than others. However, their accuracy intercepts (Figure 5A) did tend to be higher, which is almost inevitable in this kind of model. Figure 5F shows the values of λ_1 , λ_2 , λ_3 showing that participants who had higher β_{s1} tended to be more accurate.

It is important to note that differences in participants’ accuracy intercepts under this model may, and indeed likely do, reflect many other factors than just ability. That is, the intercepts simply reflect all the remaining variance from the model not explained by reaction time, since we were not measuring or controlling other differences in the model. Familiarity with similar tests, for instance, could explain part of the variance in accuracy that is reflected as differences in participants’ intercepts. In fact, participants who responded that they had taken a similar test scored 3.3 points (20%) higher, on average, than participants who reported that they had not taken a similar test ($F(1, 269) = 16.27$; $p < 0.001$).² Since there was no relationship between RT and having taken a previous test ($F(1, 269) = 0.1$; $p = 0.74$), this portion of variance (about 6%) is simply incorporated into participants’ intercepts. This is of course true as well for the myriad other factors that are not correlated or imperfectly correlated with RT, such as focus.

With that note of caution in mind, the model can still be used to estimate measures of performance controlled both for response time and pattern of response to difficulty, which can provide an upper bound on how well Raven’s can quantify ability. The posterior median accuracy intercept quantifies the accuracy that participants would have at the mean RT, with the

² 7 participants declined to answer this question.

mean response to difficulty, on the easiest items. This is correlated $R = 0.67$ with overall Raven's score. The posterior median average difficulty at which people would be 50% accurate at the average RT is correlated $R = 0.69$ with overall Raven's score. Third, the posterior median time, according to the model, it would take someone to solve the most difficult problem is correlated at $R = 0.17$ with overall Raven's score. This means that, depending on which upper-bound of "ability" we think is the most appropriate formalization, Raven's tasks capture at most approximately *half* ($R^2 = 0.695^2 = 0.48$) of the subject variation in time-controlled ability, and possibly down to virtually none ($R^2 = 0.174^2 = 0.03$).

Re-analysis with Higher-performing Participants

One potential objection to our findings is that, because the experiment was conducted online and without great incentive to perform well, a significant subset of participants may not have been engaged—more than would be expected in traditional test-taking settings—and these participants are driving all the results. It is true that participants in our sample performed somewhat worse on average on our task than in samples reported in the APM manual (Raven & Court, 1988): 52% correct in our sample vs. 53–58% correct in the APM depending on the population tested. To account for the possibility that lower average engagement levels were distorting our results, we ran post-hoc (not pre-registered) analyses using an even stricter exclusion criteria. Specifically, anyone who did not answer all of the first three questions and at least 25% of all questions correctly was excluded. This left 176 participants who answered 58% of the questions correctly on average.

Even in this higher-performing sample, we find that differential time investment on difficult questions is a strong predictor of overall performance. We first re-ran the regressions within individual participants, predicting their response time from item number. Both intercepts and slopes were again correlated with overall Raven's ($R = -0.49$, $p < 0.001$ for intercepts, $R = 0.66$, $p < 0.001$ for slopes); partialing the variance shows that Raven's score is better explained by response to difficulty (slope partial $R^2 = 0.26$) than by overall time spent ($R^2 = 0.01$). We then re-ran the hierarchical Bayesian model and found, similar to the initial results, that the posterior median accuracy intercept explained about half of differences in overall score ($R^2 = 0.53$) and the posterior median time required to solve the most difficult problem explained almost none ($R^2 = 0.01$).

DISCUSSION AND CONCLUSION

These results document substantial variation in how participants respond to difficulty in a standard intelligence task. Moreover, the variation matters: participants' response to difficulty explains 42% of the variance in overall performance. In this case, it is not surprising that a measure like Raven's would correlate with other life outcomes (Mackintosh, 2011; Richardson et al., 2012; Strenze, 2007), just as personality measures do (Duckworth & Seligman, 2005; Duckworth et al., 2019; Heckman & Kautz, 2012; Poropat, 2009). The idea that time investment on Raven's might drive correlations with life outcomes is conceptually close to "grit" (Duckworth et al., 2007; Duckworth & Quinn, 2009), which is an individual measure intended to capture an individual's willingness to work towards a long-term goal (for critiques, see Credé et al. (2017)). Notably, it was not the faster (or slower) workers or thinkers who did well, but rather those who dedicated more time to the hard questions.

An important question is how much the results from our study—which used an online 'crowd labor' marketplace to recruit participants—will generalize to a traditional test setting. In particular, online platforms may incentivize strategic time allocation, and therefore have a

greater time-difficulty tradeoff, relative to an in-person test. However, we believe that our findings have broad applicability, and are likely to generalize, for several reasons. First, recent studies have found that Prolific participants generally exhibit high levels of task engagement, supported by strong scores on tests of attention, comprehension, and reliability (Peer et al., 2022). Second, our re-analysis of high-scoring participants yielded results that were remarkably consistent with the entire sample, suggesting that even within more motivated groups there are still large individual differences in responses to difficulty. Lastly, the growing preference for online platforms in social science research—due to their cost-effectiveness, demographic diversity, and generally high quality of data—underscores that, if nothing else, our findings are relevant to contemporary social scientists interested in individual differences.

Our results align with a recent study by Schirner et al. (2023), which found that participants in the Human Connectome Project who had higher Penn Matrix Reasoning scores were those who took longer on hard questions. They linked the differential time allocation to easy and hard problems to measures of functional connectivity, finding that slower solvers had higher resting state connectivity. Simulations from a network model, which represents relationships between brain regions and mutual patterns of excitation and inhibition, identified ratios of excitation and inhibition between regions as a plausible neural candidate underlying differences in functional connectivity and, they argue, the difference between high-*g* and low-*g* individuals. However, that work leaves explanations at a cognitive level largely unaddressed.

There are several possible drivers of the relationship between response to difficulty and success on reasoning tasks. First, people's decisions about how much time to invest in each problem may be driven by meta-cognitive awareness or belief about their likelihood of finding the correct solution in a reasonable amount of time. Participants may give-up on questions they judge to be too difficult, and this may even reflect a sensible test-taking strategy, since the test has an overall time limit. However, very few participants (4%) ran out of time at the end, making it less likely that participants who invested less time on hard questions were using a rational strategy to maximize performance. Furthermore, while confidence is a well-known factor affecting test-taking (Ergene, 2003; Stankov & Lee, 2008; Stankov et al., 2012), differences in test strategy due to confidence is only weakly supported by our data: subjects' overall score was correlated with a confidence rating they provided at the end ($R = 0.52$, $p < 0.001$), but their confidence was only weakly correlated ($R = 0.17$, $p = 0.003$) with the average *time* they spent on the task (i.e., explaining less than 3% of the variance in total test time). We note though that feedback was not provided in the task, so any person's judgements about their own ability must come from intrinsic beliefs or suppositions about what the correct answers were or how easy they were to find.

A second, non-exclusive, possibility is that participants vary intrinsically in how much effort they are willing to invest in the task. When the reward size is not directly or obviously coupled to outcomes, participants may defaultly choose to invest variable amounts of time and energy. This idea is supported by the moderate to large effect sizes reviewed above for how task incentives affect performance (Duckworth et al., 2011). Such a finding has the potential to explain other demographic influences on Raven's performance—for example, people with less schooling may be less familiar with or comfortable with testing situations and the sustained energy and attention they require; and people from lower-socioeconomic levels may intrinsically make a different tradeoff with their time.

Either possibility—rational meta-cognitive strategies or intrinsic variation in effort—is markedly different from the standard interpretation of IQ tests as providing a measure of “ability.”

The notion that “intelligence” is what such tests quantify *by definition* has found some popularity (Boring, 1961; Van der Maas et al., 2014), but the view becomes difficult to sustain once alternative predictors of performance are clearly articulated. The amount of time someone allocates to a task is, we believe, not what anyone actually means by “intelligence.” Indeed, given variation in time investment, attempts to develop factor-analytic theories of intelligence seem doomed to uninterpretability: once the underlying measures are highly confounded by individual variation in effort or interest, the resulting factor structure means little. A way out of this is to focus on uncovering mechanisms and testing them empirically.

We emphasize that the amount of time spent on each item is likely only a proxy for real *cognitive* approaches to solving Raven’s tasks, and should not be confused for the real cognitive mechanism generating success in the task. For example, some authors have developed computational models which formalize mechanistic hypotheses about how intelligent agents may solve Raven’s or Raven’s-like problems (Depeweg et al., 2018; Gonthier & Roulin, 2020; Hernández-Orallo et al., 2016; Kunda et al., 2013; Little et al., 2012; Lovett et al., 2010; Carpenter et al., 1990), often searching over logical, rule-like representations, a recently popular approach to explaining high-level cognition (Rule et al., 2020). Other work has documented the effects of speed in generating possible rules (Verguts et al., 1999). Verguts and De Boeck (2002) showed that people’s search for rules preferentially re-uses rules they previously encountered—a finding which might provide a cognitive basis for practice and coaching effects. Carpenter et al. (1990) used eye-tracking and verbal reports from subjects engaged in a standard Raven’s task and showed that participants incrementally find components of rules, emphasizing the search process required to solve these problems. Work has shown that eye movements reflect different strategies for finding solutions (Hayes et al., 2011; Vigneau et al., 2006), and in fact that eye-movement differences may confound claimed processing time correlations (Bors et al., 1993).

A focus on understanding the real mechanisms of performance—developing models which can themselves solve problems like Raven’s—is a promising way to resolve the field’s century-long debate about the construct validity of intelligence measures. Timing decisions are one of the most basic components of mechanisms, but success is only possible when strategic decisions are combined with the right representations and inference procedures, which remain unclear. It is notable that neglect of mechanism has prevented the field from centering perhaps *the most basic* fact about such a widely used psychometric test: that the people who score highly are those who invest the most time on hard questions.

FUNDING INFORMATION

This work was supported by a seed grant from the Institute for Brain & Cognitive Sciences at UC Berkeley to Samuel Cheyette & Steven Piantadosi.

AUTHOR CONTRIBUTIONS

SJC & STP contributed equally to conceptualization, design, and writing. SJC led the data analysis with support from STP.

DATA AVAILABILITY STATEMENT

Data and model code are freely available at <https://osf.io/9rz2v/>.

REFERENCES

- Bates, T., & Stough, C. (1997). Processing speed, attention, and intelligence: Effects of spatial attention on decision time in high and low IQ subjects. *Personality and Individual Differences*, 23(5), 861–868. [https://doi.org/10.1016/S0191-8869\(97\)00089-5](https://doi.org/10.1016/S0191-8869(97)00089-5)
- Bates, T., & Stough, C. (1998). Improved reaction time method, information processing speed, and intelligence. *Intelligence*, 26(1), 53–62. [https://doi.org/10.1016/S0160-2896\(99\)80052-X](https://doi.org/10.1016/S0160-2896(99)80052-X)
- Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed-accuracy tradeoff. *Trends in Neurosciences*, 33(1), 10–16. <https://doi.org/10.1016/j.tins.2009.09.002>, PubMed: 19819033
- Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, 82(4), 1126–1148. <https://doi.org/10.1007/s11336-016-9537-6>, PubMed: 27738955
- Boring, E. G. (1961). Intelligence as the tests test it. In J. J. Jenkins & D. G. Paterson (Eds.), *Studies in individual differences: The search for intelligence* (pp. 210–214). Appleton-Century-Crofts. <https://doi.org/10.1037/11491-017>
- Bors, D. A., MacLeod, C. M., & Forrin, B. (1993). Eliminating the IQ-RT correlation by eliminating an experimental confound. *Intelligence*, 17(4), 475–500. [https://doi.org/10.1016/0160-2896\(93\)90014-V](https://doi.org/10.1016/0160-2896(93)90014-V)
- Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, 41(4), 291–310. <https://doi.org/10.1111/j.1745-3984.2004.tb01167.x>
- Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. *Chance*, 14(1), 10–18. <https://doi.org/10.1080/09332480.2001.10542245>
- Brinch, C. N., & Galloway, T. A. (2012). Schooling in adolescence raises IQ scores. *Proceedings of the National Academy of Sciences*, 109(2), 425–430. <https://doi.org/10.1073/pnas.1106077109>, PubMed: 22203952
- Bürkner, P.-C. (2020). Analysing standard progressive matrices (SPM-LS) with Bayesian item response models. *Journal of Intelligence*, 8(1), Article 5. <https://doi.org/10.3390/jintelligence8010005>, PubMed: 32033073
- Cahan, S., & Cohen, N. (1989). Age versus schooling effects on intelligence development. *Child Development*, 60(5), 1239–1249. <https://doi.org/10.2307/1130797>, PubMed: 2805901
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>, PubMed: 36568334
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97(3), 404–431. <https://doi.org/10.1037/0033-295X.97.3.404>, PubMed: 2381998
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511571312>
- Ceci, S. J. (1996). *On intelligence*. Harvard University Press. <https://doi.org/10.4159/9780674029316>
- Chittka, L., Skorupski, P., & Raine, N. E. (2009). Speed-accuracy tradeoffs in animal decision making. *Trends in Ecology & Evolution*, 24(7), 400–407. <https://doi.org/10.1016/j.tree.2009.02.010>, PubMed: 19409649
- Cliffordson, C., & Gustafsson, J.-E. (2008). Effects of age and schooling on intellectual performance: Estimates obtained from analysis of continuous variation in age and length of schooling. *Intelligence*, 36(2), 143–152. <https://doi.org/10.1016/j.intell.2007.03.006>
- Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, 113(3), 492–511. <https://doi.org/10.1037/pspp0000102>, PubMed: 27845531
- Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement*, 5(4), 411–427. <https://doi.org/10.1177/001316444500500408>
- Deary, I. J., & Stough, C. (1996). Intelligence and inspection time: Achievements, prospects, and problems. *American Psychologist*, 51(6), 599. <https://doi.org/10.1037/0003-066X.51.6.599>
- De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, Article 10, Article 102. <https://doi.org/10.3389/fpsyg.2019.00102>, PubMed: 30787891
- Depeweg, S., Rothkopf, C. A., & Jäkel, F. (2018). Solving Bongard problems with a visual language and pragmatic reasoning. *arXiv*. <https://doi.org/10.48550/arXiv.1804.04452>
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6), 1087–1101. <https://doi.org/10.1037/0022-3514.92.6.1087>, PubMed: 17547490
- Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, 91(2), 166–174. <https://doi.org/10.1080/00223890802634290>, PubMed: 19205937
- Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108(19), 7716–7720. <https://doi.org/10.1073/pnas.1018601108>, PubMed: 21518867
- Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, 16(12), 939–944. <https://doi.org/10.1111/j.1467-9280.2005.01641.x>, PubMed: 16313657
- Duckworth, A. L., Taxer, J. L., Eskreis-Winkler, L., Galla, B. M., & Gross, J. J. (2019). Self-control and academic achievement. *Annual Review of Psychology*, 70, 373–399. <https://doi.org/10.1146/annurev-psych-010418-103230>, PubMed: 30609915
- Entink, R. H. K., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74(1), 21–48. <https://doi.org/10.1007/s11336-008-9075-y>, PubMed: 20037635
- Ergene, T. (2003). Effective interventions on test anxiety reduction: A meta-analysis. *School Psychology International*, 24(3), 313–328. <https://doi.org/10.1177/01430343030243004>
- Evans, R. B., & Deary, I. J. (1994). Sensory discrimination and intelligence: Postmortem or resurrection? *American Journal of Psychology*, 107(1), 95–115. <https://doi.org/10.2307/1423292>, PubMed: 8185025
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 101(2), 171–191. <https://doi.org/10.1037/0033-2909.101.2.171>
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media. <https://doi.org/10.1007/978-1-4419-0742-4>
- Furneaux, W. (1973). Intellectual abilities and problem-solving behaviour. In *The measurement of intelligence* (pp. 212–237). Springer. https://doi.org/10.1007/978-94-011-6129-9_14

- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511790942>
- Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, 13(3–4), 133–164. <https://doi.org/10.1080/15366367.2015.1100020>, PubMed: 26807063
- Goldhammer, F., & Entink, R. H. K. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, 39(2–3), 108–119. <https://doi.org/10.1016/j.intell.2011.02.001>
- Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, 3(1), 21–40. <https://doi.org/10.3390/jintelligence3010021>
- Gonthier, C., & Roulin, J.-L. (2020). Intraindividual strategy shifts in Raven's matrices, and their dependence on working memory capacity and need for cognition. *Journal of Experimental Psychology: General*, 149(3), 564–579. <https://doi.org/10.1037/xge0000660>, PubMed: 31318256
- Gould, S. J. (1996). *The mismeasure of man*. WW Norton & Company.
- Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. *Intelligence*, 29(6), 523–535. [https://doi.org/10.1016/S0160-2896\(01\)00078-2](https://doi.org/10.1016/S0160-2896(01)00078-2)
- Gulliksen, H. (1950). *Theory of mental tests*. Routledge. <https://doi.org/10.4324/9780203052150>
- Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision*, 11(10), 10. <https://doi.org/10.1167/11.10.10>, PubMed: 21926182
- Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4), 451–464. <https://doi.org/10.1016/j.labeco.2012.05.014>, PubMed: 23559694
- Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8, Article 150. <https://doi.org/10.3389/fnins.2014.00150>, PubMed: 24966810
- Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, 76(3), 616–628. <https://doi.org/10.1016/j.neuron.2012.08.030>, PubMed: 23141072
- Hernández-Orallo, J., Martínez-Plumed, F., Schmid, U., Siebers, M., & Dowe, D. L. (2016). Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, 230, 74–107. <https://doi.org/10.1016/j.artint.2015.09.011>
- Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1), 11–26. <https://doi.org/10.1080/17470215208416600>
- Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.
- Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93–132). Springer. https://doi.org/10.1007/978-3-642-68664-1_4
- Jensen, A. R. (1985). The nature of the Black-White difference on various psychometric tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, 8(2), 193–219. <https://doi.org/10.1017/S0140525X00020392>
- Jensen, A. R. (1998). The g factor and the design of education. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, instruction, and assessment* (pp. 111–132). Routledge.
- Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Elsevier. <https://doi.org/10.1016/B978-0-08-044939-5.X5000-9>
- Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, 95(2), 179–188. <https://doi.org/10.1037/0033-2909.95.2.179>
- Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, 21(2), 435–447. <https://doi.org/10.3102/00028312021002435>
- Kunda, M., McGreggor, K., & Goel, A. K. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test using iconic visual representations. *Cognitive Systems Research*, 22–23, 47–66. <https://doi.org/10.1016/j.cogsys.2012.08.001>
- Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, 4(4), Article 14. <https://doi.org/10.3390/jintelligence4040014>
- Liesefeld, H. R., Fu, X., & Zimmer, H. D. (2015). Fast and careless or careful and slow? Apparent holistic processing in mental rotation is explained by speed-accuracy trade-offs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(4), 1140–1151. <https://doi.org/10.1037/xlm0000081>, PubMed: 25528084
- Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, 51(1), 40–60. <https://doi.org/10.3758/s13428-018-1076-x>, PubMed: 30022459
- Little, D. R., Lewandowsky, S., & Griffiths, T. L. (2012). A Bayesian model of rule induction in Raven's progressive matrices. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 34, pp. 1918–1923).
- Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping model of Raven's Progressive Matrices. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, pp. 2761–2766).
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195070019.001.0001>
- Mackintosh, N. J. (2011). *IQ and human intelligence*. Oxford University Press.
- Mensh, E., & Mensh, H. (1991). *The IQ mythology: Class, race, gender, and inequality*. SIU Press.
- Myszkowski, N., & Storme, M. (2018). A snapshot of g? Binary and polytomous item-response theory investigations of the last series of the Standard Progressive Matrices (SPM-LS). *Intelligence*, 68, 109–116. <https://doi.org/10.1016/j.intell.2018.03.010>
- Nettelbeck, T. (1998). Jensen's chronometric research: Neither simple nor sufficient but a good place to start. *Intelligence*, 26(3), 233–241. [https://doi.org/10.1016/S0160-2896\(99\)80006-3](https://doi.org/10.1016/S0160-2896(99)80006-3)
- Neubauer, A. C. (1990). Speed of information processing in the Hick paradigm and response latencies in a psychometric intelligence test. *Personality and Individual Differences*, 11(2), 147–152. [https://doi.org/10.1016/0191-8869\(90\)90007-E](https://doi.org/10.1016/0191-8869(90)90007-E)
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643–1662. <https://doi.org/10.3758/s13428-021-01694-3>, PubMed: 34590289
- Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, 135(2), 322–338. <https://doi.org/10.1037/a0014996>, PubMed: 19254083
- Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice*, 12(2), 24–30. <https://doi.org/10.1111/j.1745-3992.1993.tb00530.x>
- Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, 41(1), 1–48. <https://doi.org/10.1006/cogp.1999.0735>, PubMed: 10945921

- Raven, J. C., & Court, J. H. (1988). *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press.
- Raven, J. C., Court, J. H., & Raven, J. E. (1989). *Standard progressive matrices*. Australian Council for Educational Research Limited.
- Richardson, K. (2002). What IQ tests test. *Theory & Psychology, 12*(3), 283–314. <https://doi.org/10.1177/0959354302012003012>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353–387. <https://doi.org/10.1037/a0026838>, PubMed: 22352812
- Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement, 16*(4), 261–270. <https://doi.org/10.1111/j.1745-3984.1979.tb00107.x>
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science, 29*(8), 1358–1369. <https://doi.org/10.1177/0956797618774253>, PubMed: 29911926
- Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences, 24*(11), 900–915. <https://doi.org/10.1016/j.tics.2020.07.005>, PubMed: 33012688
- Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence, 48*, 37–50. <https://doi.org/10.1016/j.intell.2014.10.003>
- Schirner, M., Deco, G., & Ritter, P. (2023). Learning how network structure shapes decision-making for bio-inspired computing. *Nature Communications, 14*(1), Article 2963. <https://doi.org/10.1038/s41467-023-38626-y>, PubMed: 37221168
- Schönemann, P. H. (1983). Do IQ tests really measure intelligence? *Behavioral and Brain Sciences, 6*(2), 311–313. <https://doi.org/10.1017/S0140525X00016125>
- Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences, 44*(3), 535–551. <https://doi.org/10.1016/j.paid.2007.09.015>
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*(2), 201–292. <https://doi.org/10.2307/1412107>
- Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*(4), 961–976. <https://doi.org/10.1037/a0012546>
- Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*(6), 747–758. <https://doi.org/10.1016/j.lindif.2012.05.013>
- Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401–426. <https://doi.org/10.1016/j.intell.2006.09.004>
- Tate, M. W. (1948). Individual differences in speed of response in mental test materials of varying degrees of difficulty. *Educational and Psychological Measurement, 8*(3-1), 353–374. <https://doi.org/10.1177/001316444800800307>, PubMed: 18885688
- Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13*(3), 201–214. <https://doi.org/10.1111/j.1745-3984.1976.tb00011.x>
- Thissen, D. (1983). Timed testing: An approach using item response theory. In *New horizons in testing* (pp. 179–203). Elsevier. <https://doi.org/10.1016/B978-0-12-742780-5.50019-6>
- Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika, 2*, 249–254. <https://doi.org/10.1007/BF02287896>
- Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge University Press.
- van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*(3), 247–272. <https://doi.org/10.1111/j.1745-3984.2009.00080.x>
- van der Linden, W. J., & Fox, J.-P. (2016). Joint hierarchical modeling of responses and response times. In *Handbook of item response theory* (pp. 509–528). Chapman and Hall/CRC.
- van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5), 327–347. <https://doi.org/10.1177/0146621609349800>
- Van der Maas, H. L. J., Kan, K.-J., & Borsboom, D. (2014). Intelligence is what the intelligence test measures. Seriously. *Journal of Intelligence, 2*(1), 12–15. <https://doi.org/10.3390/jintelligence2010012>
- Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *Journal of Cognition, 1*(1), Article 8. <https://doi.org/10.5334/joc.6>, PubMed: 31517182
- Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices Test. *European Journal of Cognitive Psychology, 14*(4), 521–547. <https://doi.org/10.1080/09541440143000230>
- Verguts, T., De Boeck, P., & Maris, E. (1999). Generation speed in Raven's progressive matrices test. *Intelligence, 27*(4), 329–345. [https://doi.org/10.1016/S0160-2896\(99\)00023-9](https://doi.org/10.1016/S0160-2896(99)00023-9)
- Vernon, P. A. (1983). Speed of information processing and general intelligence. *Intelligence, 7*(1), 53–70. [https://doi.org/10.1016/0160-2896\(83\)90006-5](https://doi.org/10.1016/0160-2896(83)90006-5)
- Vigneau, F., Caissie, A. F., & Bors, D. A. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence, 34*(3), 261–272. <https://doi.org/10.1016/j.intell.2005.11.003>
- Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*(1), 67–85. [https://doi.org/10.1016/0001-6918\(77\)90012-9](https://doi.org/10.1016/0001-6918(77)90012-9)