

Speed–Accuracy Trade-Off? Not So Fast: Marginal Changes in Speed Have Inconsistent Relationships With Accuracy in Real-World Settings

Benjamin W. Domingue , **Klint Kanopka**, **Ben Stenhaus**,
Michael J. Sulik

Stanford Graduate School of Education

Tanesia Beverly^a

University of Connecticut

Law School Admissions Council

Matthieu Brinkhuis^a 

Utrecht University

Ruhan Circi^a

American Institutes for Research

Jessica Faul^a

University of Michigan

Dandan Liao^a

Cambium Assessment, Inc.

Bruce McCandliss^a and **Jelena Obradović**^a

Stanford Graduate School of Education

Chris Piech^a

Stanford University

Tenelle Porter^a

University of California, Davis

Project iLEAD Consortium^a

University of California San Francisco

James Soland^a 

University of Virginia

Jon Weeks^a

Educational Testing Service

Steven L. Wise^a

NWEA

Jason Yeatman^a

Stanford Graduate School of Education

Stanford University School of Medicine

^aThese authors are listed alphabetically.

The speed–accuracy trade-off (SAT) suggests that time constraints reduce response accuracy. Its relevance in observational settings—where response time (RT) may not be constrained but respondent speed may still vary—is unclear. Using 29 data sets containing data from cognitive tasks, we use a flexible method for identification of the SAT (which we test in extensive simulation studies) to probe whether the SAT holds. We find inconsistent relationships between time and accuracy; marginal increases in time use for an individual do not necessarily predict increases in accuracy. Additionally, the speed–accuracy relationship may depend on the underlying difficulty of the interaction. We also consider the analysis of items and individuals; of particular interest is the observation that respondents who exhibit more within-person variation in response speed are typically of lower ability. We further find that RT is typically a weak predictor of response accuracy. Our findings document a range of empirical phenomena that should inform future modeling of RTs collected in observational settings.

Keywords: response time; IRT; speed–accuracy trade-off; conditional accuracy function

1. Introduction

The basic notion of the speed–accuracy trade-off (SAT) is an intuitively appealing one: An individual’s slow, deliberative decisions should, all else equal, be more accurate than rushed decisions (Wickelgren, 1977; see also Figure 1). Beyond its intuitive appeal, it has been verified using extensive work in



FIGURE 1. Prototypical speed–accuracy trade-off. For an individual, increases in response time are, all else equal, expected to translate into increase in accuracy relative to expectation (gray line); this is indicated by the upward slope of the blue line. Note that there is no time limit considered in this hypothetical scenario.

experimental settings, wherein a variety of manipulations are used to induce changes in speed (Heitz, 2014). While there is great power in using experimental manipulation of time pressure for the identification of this phenomena, experimental results do not necessarily generalize to nonexperimental settings where additional factors may impact the choice of speed and the resulting level of accuracy. The ubiquity of digital interfaces for all manner of widely varying psychometric instruments has rapidly increased the availability of response time (RT) data in observational settings. This increase in RT data increases the need for models—both conceptual and statistical—for understanding such data and also increases the importance of questions about the generalizability of insights regarding RT derived from experimental settings.

In settings wherein time pressures are not explicitly being manipulated, the SAT may still be a relevant model of behavior. Earlier work has described this kind of SAT, based on idiosyncratic within-person changes in speed during the measurement process, as a “micro” SAT (Dennis & Evans, 1996) in contrast with the “macro” SAT, which is typically targeted via direct experimental manipulation. Initial empirical work supported the concept (Lappin & Disch, 1972; Schouten & Bekker, 1967). Such work posits that individuals are continuously making choices about trade-offs between speed and accuracy in the course of responding, thus making it a relevant phenomenon even when speed is not being explicitly manipulated. In nonexperimental work, respondents are potentially making decisions about time usage due to other pressures (i.e., boredom, fatigue, or testing anxiety may play a role in some settings) that may have different implications for accuracy. This study probes the general utility of the SAT in anticipating behavior across a broad variety of cognitive tasks wherein we study the association of speed changes with accuracy.

The increase in RT data is also leading to the development of a suite of statistical approaches for the study of RT, especially in conjunction with response accuracy (Molenaar et al., 2018; Ranger et al., 2015; Ratcliff et al., 2016; van der Linden, 2007). These approaches account, or do not, for the SAT in several ways. For example, the hierarchical model (van der Linden, 2007)—which has been widely used in educational measurement settings—posits no within-person interplay between speed and accuracy by assuming that speed is constant (as it might be in, for example, a high-stakes measurement scenario with no time constraints). An alternative viewpoint posits that RT is a mixture of guessing and solution behavior wherein these two modes have different implications for accuracy (Wang & Xu, 2015). Other approaches (e.g., the drift diffusion model, Ratcliff et al., 2016) explicitly link RT and accuracy based on the models of decision making (such an approach has experimental support, Palmer et al., 2005) and still others (van Rijn & Ali, 2018) upweight rapid responses in terms of how they inform inferences about respondent ability. These approaches all make presumptions about interplay between RT and accuracy that may not be empirically supported in specific contexts.

While it is clear that the SAT is a useful hypothesis for describing behavior in some settings, we argue that it deserves further scrutiny when applied to non-experimental data across a range of tasks. We aim to study, in a variety of data, whether the general intuition behind the SAT holds. Conceptually, this study builds on work suggesting that additional time spent on a response does not always increase its accuracy (Bolsinova & Molenaar, 2018; Goldhammer et al., 2014; Ranger et al., 2021). For example, Chen et al. (2018) discuss a curvilinear relationship between RT and accuracy: Increases in time spent on an item were associated with increases in accuracy, but only up to a certain point.

In the spirit of earlier work on the SAT (Pew, 1969), we explore this issue using a large number of data sets containing both response accuracy and time from various cognitive tasks. We combine this data with a flexible exploratory approach describing the relationship between within-person variation in time usage and accuracy. We first use an item response model to generate an estimate of the probability of accuracy for a person-item interaction. We then use within-person variation in RT to ask whether extra time spent on an item tends to yield marginal increases in accuracy net of this probability. In such cases, the basic logic of the SAT holds, but, of course, it need not.

Alongside this main question, we ask several additional questions pertaining to interplay between speed and accuracy. We focus on the issues of interest that have seen relatively limited empirical work (especially across diverse data). We ask whether there is heterogeneity in the association between time usage and accuracy as a function of the task or item's level of difficulty (i.e., the probability of accuracy as specified by an item response model). We ask about the existence of item-level and person-level variation in the degree to which marginal changes in time predict change in accuracy. Finally, given the interest in formal models linking time and accuracy, we use RT to predict accuracy in out-of-sample analyses. Collectively, answers to these questions offer insight as to the degree to which the SAT should be embedded in our conceptual and statistical models for responses not collected in experimental settings.

2. Methods

2.1. Data

We consider item response data sets containing a variety of tasks and with respondents of various ages; they are documented in the Supplemental Information (SI). The primary criteria for inclusion were as follows: (1) Time pressures were not experimentally manipulated across the tasks,¹ (2) the data came from cognitive tasks, and (3) accuracy can be appropriately modeled as a monotonically increasing function of some latent trait. Data that are appropriately modeled using item response theory (IRT) models with monotonic item response functions would thus be permissible. In contrast, data from measures of affective traits (e.g., personality) or otherwise characterized by nonmonotonic models—e.g., “D”/“unfolding” models

TABLE 1.

Descriptive Statistics for the Data Sets (Including Time Limits for Those Data Sets That Impose Them at the Item Level)

Data	# People	# Items	# Interactions	Mean Time (s)	Time Limit (s)
Lexical	93	15	66,059	0.6	
RR98 accuracy	30	33	12,194	0.7	
Hearts flowers	255	8	5,071	0.8	1.5
Lexical Decision Task	104	495	51,480	0.9	
ECLS flanker	12,008	20	239,963	0.9	10.0
ECLS DCCS	12,023	30	360,430	1.1	10.0
Motion	106	30	31,778	1.2	10.0
Multi-Source Interference Test	740	24	16,739	1.3	2.5
Reading fluency	3,943	315	212,507	1.4	
Reading comp	3,947	448	165,630	1.5	
Arithmetic	895	173	133,796	1.7	
Groupitizing	481	88	40,450	2.2	
Rotation	95	10	950	2.6	7.5
Set	355	10	3,550	5.0	20.0
Letter chaos	233	10	2,330	5.7	20.0
Add subtract	16,190	60	200,297	6.0	20.0
Working memory	194	4	1,365	6.9	
Mult. div.	14,184	60	174,517	7.0	20.0
Health and Retirement Study	2,215	20	36,785	8.2	
Chess	258	80	19,135	9.7	30.0
PISA reading	42,398	223	1,850,217	11.0	
PERC	1,680	15	25,132	16.9	
MITRE-ETS	801	95	75,912	18.3	90.0
Assistments	2,306	3,518	131,864	21.8	
National Social Life, Health, and Aging Project	2,210	13	28,717	31.2	
Programme for the International Assessment of Adult Competencies	2,278	104	55,563	38.4	
PISA math	21,995	60	323,887	62.6	
NWEA Grade 3	49,998	5,181	1,952,749	64.1	
NWEA Grade 8	49,984	6,049	1,888,845	79.0	

Note. PISA = Programme for International Student Assessment; ECLS = Early Childhood Longitudinal Studies; DCCS = dimensional change card sort; PERC = Persistence, effort, resilience and challenge-seeking task.

(Molenaar et al., 2015)—would not be included. We focus on data that had responses scored in two categories (e.g., correct or incorrect).² Collectively, these data draw from measures that span a range of constructs measured across the lifecourse.

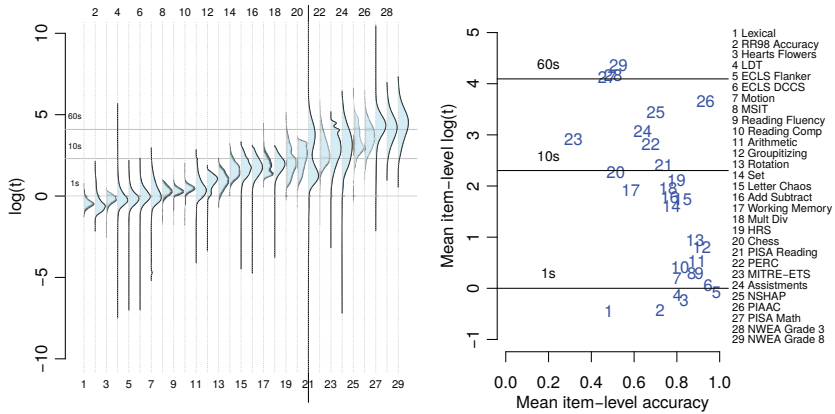


FIGURE 2. Response time (RT). Note. Left: Boxplots of RT (logged) for each of the data sets. Right: Comparison of mean item-level accuracy (x-axis) and RT (y-axis) across the items. Horizontal lines show 1 second, 10 second, and 1 minute increments.

Descriptive statistics, including the size of each data set, are in Table 1. Data range widely in scale; e.g., 30 people or < 10 items to 50,000 people and thousands of items. Note that the number of responses diverges widely in many cases from the product of the number of people times the number of items. For reasons largely of design (i.e., the assessment was delivered via blocks or adaptively), not all individuals attempt all items in some data (e.g., Programme for International Student Assessment [PISA], NWEA). In other cases, items are attempted multiple times.

Figure 2 describes RT in these data. Given the skew associated with time, we use logged time throughout. Tests vary substantially in terms of the amount of time required per interaction. Some tests have items that require less than 1 second on average, while others have items that require more than 1 minute. We order the data by mean RT in our presentation of results. There is also variation in the difficulty of the items, as proxied by average percentage correct, across the assessments. Some of the tests have items for which only half of the responses are correct while others have items for which responses are nearly always correct. We control for this variation via item response models.

2.2. Analysis

There are several conceptual rationales for combining information on RT and accuracy (De Boeck & Jeon, 2019). RT can be used as collateral information to improve the prediction of accuracy, RT can be incorporated for the purpose of studying the underlying cognitive processes, or RT can be explicitly incorporated into scoring rules. The approach used here—in particular, combining

probabilities from item response models with fixed effects—has features of the first two approaches.³

We view our approach as one designed to flexibly model the impacts on the accuracy of within-person changes in time use. In particular, we argue that our approach is reasonable for elucidating key facts about the SAT that we study here but we would not consider it a plausible generative model; indeed, the breadth of data here may require qualitatively different types of generative models. In being an approach focused not on testing models that may have generated the data but instead on key empirical relationships, we view it as being in the tradition of Tukey and others advocating for such exploratory approaches (Fife & Rodgers, 2021; Tukey et al., 1977). Our confidence in our approach’s ability to capture key features of the SAT is based on extensive simulation studies in the context of several different models for the joint distribution of time and accuracy, see SI.

All analyses are conducted in R. We use *mirt* software (Chalmers et al., 2012) to estimate IRT models and the *fixest* package to estimate fixed effect models (Bergé et al., 2018). Code is available (https://github.com/ben-domingue/rt_meta).

2.2.1. Mapping speed–accuracy curves. Our first aim is to estimate within-person speed–accuracy curves (i.e., conditional accuracy functions, Maris & Van der Maas, 2012). To do this, we rely on a flexible approach to recovering these curves that allows us to identify a variety of different configurations of speed and accuracy. Alongside information about RT, we utilize the estimates of the probability of a correct response via the application of a specific item response model. We combine these estimates of accuracy with RT in a linear probability model-based approach to identify speed-accuracy curves. The flexibility comes at the cost of some slightly unconventional choices (i.e., the linear probability model), but we illustrate its robust performance under a variety of assumptions in simulation studies (see SI).

Let x_{ij} be a dichotomously scored responses from person j to item i (so $x_{ij} \in \{0, 1\}$). The first element of our approach involves estimating p_0 , the probability of a correct response generated from the application of an item response model; specifically, the Rasch (1993) model. We estimate

$$p_0 = \Pr(x_{ij} = 1) = \sigma(\theta_j - \delta_i), \quad (1)$$

where θ_j and δ_i are the person-level and item-level parameters, respectively, and $\sigma(z) = (1 + \exp(-z))^{-1}$. Estimation is performed using two approaches. When a conventional item response matrix can be constructed, we use conventional IRT approaches (Chalmers et al., 2012); when this is not possible—in particular, when respondents take multiple attempts at an item—we use a random effects model (De Boeck et al., 2011), where both person and item effects are treated as random.

We then use p_0 in our attempt to model associations between marginal within-person changes in time usage and accuracy. Denoting the time required for the production of person j 's response to item i as t_{ij} , we allow for nonlinear effects in time. This flexibility is important as additional RT may be unassociated with gains in accuracy past a certain threshold (Pachella et al., 1968) or otherwise nonlinear (Chen et al., 2018). We do this by mapping $\log t_{ij}$ onto a b-spline basis; we denote this as $b(\log t_{ij})_k$.⁴ We consider as a baseline model

$$x_{ij} \sim \text{Normal}\left(L(b(\log t_{ij})_k, p_{0,ij}) + \lambda_j + \gamma_i, \sigma_x^2\right), \tag{2}$$

where $L()$ indicates a linear function of its arguments (e.g., $L(x, y) = \alpha x + \beta y$). Note that we rely upon a linear probability model. While unconventional for modeling binary item responses, it is more common in other settings (Cheung, 2007; Jaccard & Brinberg, 2021). We use this approach as it allows for the computational flexibility of including person- and item-level fixed effects (λ_j and γ_i , respectively). The use of fixed effects is key to identification of the SAT as a within-person phenomenon as it allows us to control for all time-invariant properties of persons and items, including a person's typical time usage and an item's typical time demand. Given that we are ignorant as to the true relationship between speed and accuracy across the data considered here, we use splines as a flexible means of mapping speed–accuracy relationships; as we illustrate in the SI, this approach reliably allows us to uncover a variety of relationships between these quantities.

We again emphasize that we are *not* asserting that Equation 2 is the true data generating process. It doesn't, for example, allow for possible variation in item discrimination and is also in the form of a linear probability model. We view such an exploratory approach as appropriate given that it as a flexible yet robust method for uncovering the SAT that arises due to a variety of different mechanisms for jointly generating time and accuracy; its flexibility is key given the range of data we utilize here. This robustness is demonstrated in Section 2 of the SI, wherein we conduct a wide variety of simulation studies demonstrating the efficacy of our approach. It is, for example, able to detect the key features of the SAT even if the item's p values are far from 0.5 on average, can detect a variety of shapes of the SAT, and functions appropriately when a variety of models are used to generate data. As with many approaches to studying accuracy or time usage, our approach assumes no change in a respondent's *overall* ability or speed through the assessment (i.e., θ_j and λ_j are static) and relies on a relatively constrained model to generate p_0 ; we discuss potential limitations stemming from these constraints below.

2.2.2. Heterogeneity in SAT curves. Note that Equation 2 assumes that changes in accuracy are independent of the difficulty of the interaction; a marginal increase in time on an item that is relatively hard for a person is assumed to

be as useful as a marginal increase in time on an item that is easy for a person. We now relax this assumption. To explore heterogeneity as a function of p_0 , we consider

$$x_{ij} \sim \text{Normal}\left(\text{SL}(b(\log t_{ij})_k, p_{0,ij}) + \lambda_j + \gamma_i, \sigma_x^2\right), \quad (3)$$

where $\text{SL}()$ is a saturated linear function of its arguments (e.g., $\text{SL}(x, y) = \alpha x + \beta y + \eta xy$, with the one caveat that we do not include interaction terms between splines). We then consider $\frac{\partial f}{\partial \log t}$, where f is the center of the normal density in Equation 3. The goal is to explicitly identify regions of (p_0, t) space, where additional time predicts an increase ($\frac{\partial f}{\partial \log t} > 0$) or decrease ($\frac{\partial f}{\partial \log t} < 0$) in accuracy.

2.2.3. Item- and person-level analyses. To study the associations of marginal increases in time with accuracy for individual items, we consider the following model separately for each item

$$x_j \sim \text{Normal}\left(\beta_1(\log t_j) + \beta_2 p_{0,p}, \sigma_x^2\right), \quad (4)$$

where j indexes all individuals. The estimate of β_1 is an indicator of the marginal association between time and accuracy for each item. To determine whether there is a patterning of this indicator of association with the item’s difficulty, we also consider $r(\beta_1, \delta_i)$ (with δ_i from Equation 1).

To study person-level associations between speed and ability (i.e., θ in Equation 1), we estimate

$$\widetilde{(\log t_{ij})} \sim \text{Normal}(-1 \cdot \tau_j, \sigma_\tau^2), \quad (5)$$

where $\widetilde{(\log t_{ij})}$ represents demeaned (at item-level, so as to omit the between-item variation in time intensity) RTs and we additionally assume $\tau_j \sim \text{Normal}(0, \sigma_\tau^2)$. We multiply τ by -1 , so that τ represents speed (i.e., a higher τ will be associated with lower time). We first examine $r(\tau_j, \theta_j)$ so as to determine whether higher ability respondents tend to be faster or slower responders. Motivated by previous observations of within-person variation in speed (Wise, 2015), we then consider such variation. Focusing on items with at least 100 responses, we find the quantile in the RT distribution of each response (i.e., the rank) for a person and take the standard deviation of that quantity (which we denote σ_{rank}).⁵ We then consider $r(\theta_j, \sigma_{\text{rank}})$ as an indication of whether within-person variation in speed is associated with ability.

2.2.4. Predictive accuracy. Finally, we ask about the relative gain in the prediction of accuracy that we get from RT. The goal here is to benchmark the potential value of RT as a predictor of accuracy; such findings will supplement those of the SAT-focused analyses in helping to offer insight about interplay between speed and accuracy. We focus on the predictive power of RT for an item response by

comparing the accuracy of predictions in a 10% hold-out-sample of item responses using models trained in the remaining 90%.⁶ For this exercise, we first standardize RT within each item. Predictive performance is based on a transformation of the likelihood meant to provide intuition about item-level responses; if ℓ_{ij} is the log-likelihood for a response with predicted accuracy of P_{ij} ,

$$\ell_{ij} = x_{ij}(\log P_{ij}) + (1 - x_{ij})(\log 1 - P_{ij}), \tag{6}$$

we consider $\exp(\bar{\ell}_{ij})$ (where the average is taken over j and i).

We consider six models for P_{ij} in Equation 6 (denoted A–F) that utilize information on person- and item-level accuracy, information about the individual’s overall time usage, and combinations thereof. As context for evaluating gains in each data set, we first predict (A) using the invariant proportion of correct responses in each data set, $P_{ij} = \bar{x}$. We then consider item-level variation in accuracy and predict based on (B) the proportion correct by item, $P_{ij} = \sum_j x_{ij}/n_i$, where there are n_i responses to item i . We now incorporate person-level information using three quantities: the individual’s proportion of correct responses, the individual’s mean standardized RT, and, due to conceptual (Davidson et al., 2006) and empirical (Su & Davison, 2019) interest in RTs for correct responses, the individual’s mean standardized RT for correct responses.⁷ For each of these three predictors, z , we predict (C–E) based on fitted logistic regression models containing the item proportion correct and one of the three predictors; that is, $P_{ij} = \sigma(b_0 + b_1 \sum_j x_{ij}/n_i + b_2 z_{ij})$, where b_0, b_1 , and b_2 are estimated via logistic regression. Finally, we use both time and accuracy information and predict (F) based on both the individual’s proportion correct responses and mean standardized RT. Note that out-of-sample responses are predicted purely on the basis of in-sample information (i.e., out-of-sample RT is not used). We consider analyses focusing on item-level RT in the SI.⁸

3. Results

3.1. Mapping the SAT

We first construct baseline speed–accuracy curves using the approach described in Equation 2 (see Figure 3). Each panel in that figure has a similar form; they are also similar to the format of Figure 1. The x -axis captures time spent on the item.⁹ The y -axis shows changes to the estimated accuracy net of p_0 . The curves show the estimated changes in accuracy as a function of time; recall that the SAT would suggest that such lines be monotonically increasing as longer responses are associated with increases in accuracy.

We readily observe a large variety of behavior in terms of the within-person relationship between RT and accuracy. In some cases (e.g., Lexical, Arithmetic), longer RTs do generally translate into increased accuracy. However, this is not universally true. For example, longer time can be uniformly associated with a

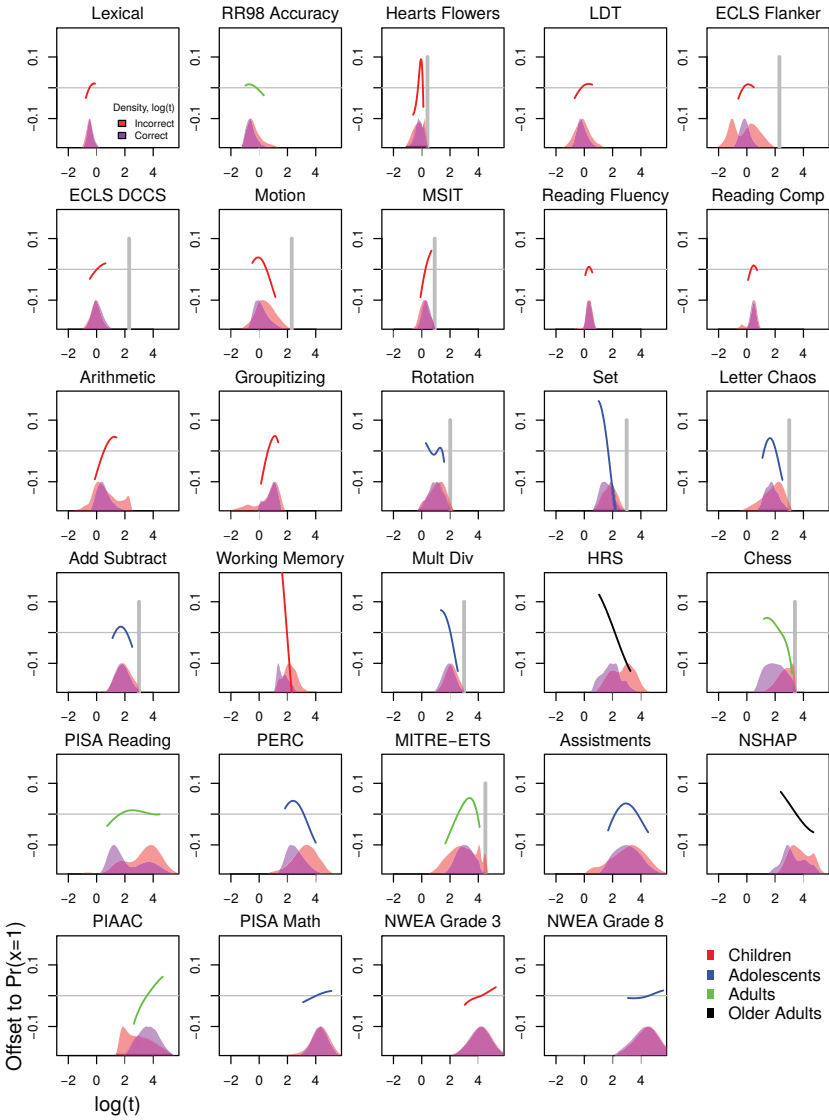


FIGURE 3. Estimated association between response time (RT) and changes to accuracy (net of p_0). For each test, the x-axis spans from the .1 to .9 quantiles of observed $\log(t)$. The y-axis focuses on offsets to the test mean of (item response theory–based) $\Pr(x = 1)$. Curves represent the estimated accuracy as a function of time. Densities at bottom of panel show the distribution of RTs for each test separately by response type. Vertical lines represent the time limits (where applicable). Line color represents the respondent age.

decline in accuracy (e.g., working memory, National Social Life, Health, and Aging Project [NSHAP]). In other cases (e.g., rotation, reading fluency), associations with accuracy for additional RT can be positive or negative. While these results suggest that a wide variety of relationships are possible, we emphasize two points of consistency here and further elaborate on some other potential explanatory mechanisms in the discussion.

Note the role of time limits. Consider the hearts flowers and rotation tasks. For those, we observe steep declines in accuracy as a function of time increases when RTs are near their maximum. In these cases, we hypothesize that respondents began to choose answers with less certainty when they neared the time limit for each task. Note that we also detect a relative increase in the density of incorrect responses prior to the time limit for these two data sets. We further illustrate the role of time limits along the lines described here using variation in time pressure in additional data from the hearts flowers task, see SI.

Within age, we generally observe variation in curve shape. However, if we focus on older respondents (the Health and Retirement Study [HRS] and NSHAP data), we observe strong negative slopes. In the context of these data, we hypothesize that the nature of the curve is due in part to both the age of the respondent and the type of task in these data. We further investigated this possibility using the Programme for the International Assessment of Adult Competencies (PIAAC) data, see SI; this analysis supports the supposition that the nature of the HRS and NSHAP tasks play some role (it does not seem to be simply the age of the respondent).

We considered several sensitivity analyses to complement the results in Figure 3. We consider p_0 values generated from an alternative item response model (i.e., the 2PL). We modeled responses using logistic regression rather than the linear probability model. We also considered results based on the first residualized RTs for person and item fixed effects. Results from analyses are described in Section 3 of the SI; our conclusion that a wide variety of relationships between speed and accuracy are possible in observational data is robust to these alternative specifications.

Figure 3 focuses on associations between RT and accuracy net of the underlying difficulty (i.e., p_0) of the interaction. We now ask whether there may be heterogeneous effects by constructing curves similar to the ones shown in Figure 3 but that vary by the difficulty of the interaction. Rather than focusing on the curve, we focus on the curve's instantaneous slope (i.e., $\frac{\partial f}{\partial \log t}$).

3.2. Heterogeneity as a Function of p_0

Conceptually, the analysis of heterogeneity as a function of p_0 is equivalent to asking whether the shape of the curve shown in Figure 1 is sensitive to the value of p_0 (i.e., the location of the horizontal gray line). Results based on the approach in Equation 3 are shown in Figure 4. In this figure (as in Figure 3), the x -axis

Speed–Accuracy Trade-Off?

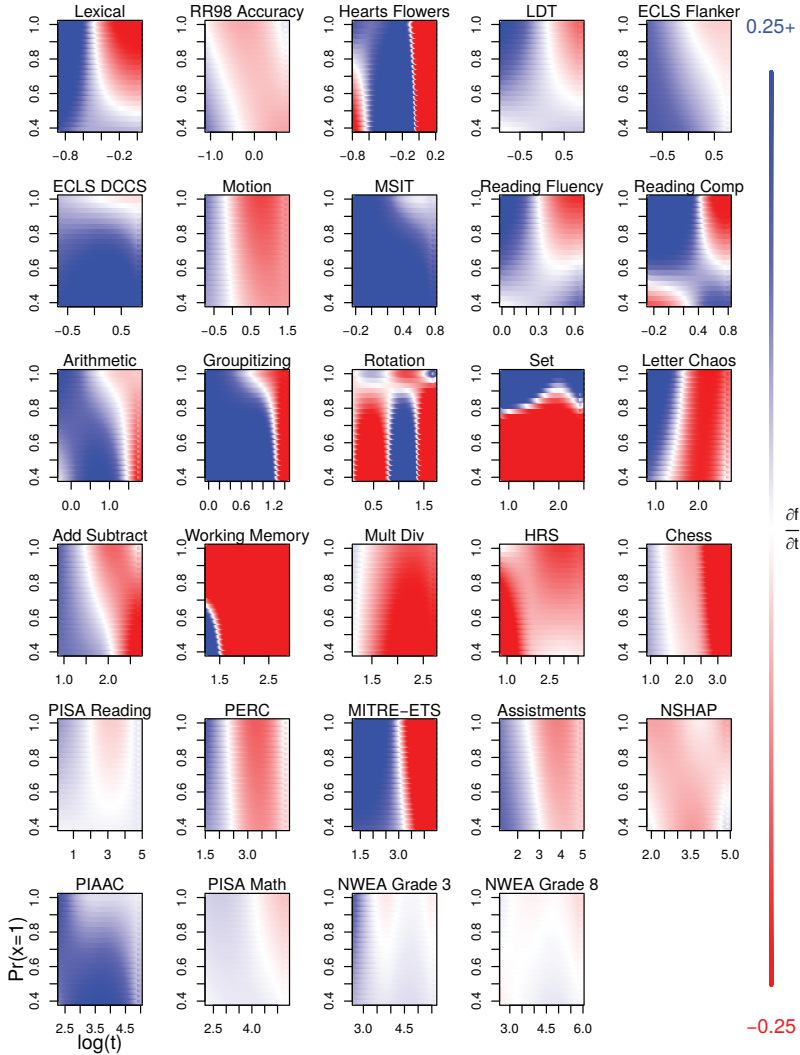


FIGURE 4. Estimated change in accuracy as a function of both response time (RT; x-axis) and p_0 (y-axis). Colors can be interpreted based on legend on right. Blue indicates points where a marginal increase in time spent by a respondent on an item is expected to increase accuracy; red indicates points where the opposite is true. A lack of color represents a point with no estimated association between marginal increase in RT and accuracy.

shows RT for the test. The y-axis shows the p_0 of the interaction; a value of, for example, 0.7 means that an individual responding to a given item is projected by the Rasch model to have a 70% probability of getting the item correct. At a given

point in each panel of the figure, the color represents $\frac{\partial f}{\partial \log t}$. Areas in blue correspond to $\frac{\partial f}{\partial \log t} > 0$, suggesting that a marginal increase in time for an interaction of the given difficulty is associated with increased accuracy (i.e., the SAT seems to be operant). Areas in red correspond to $\frac{\partial f}{\partial \log t} < 0$; in such areas, marginal increases in time are associated with decreases in accuracy. If we consider a vertical strip, a change in color suggests sensitivity in the time/accuracy relationship as a function of p_0 . Likewise, when we consider a horizontal strip a change in color suggests sensitivity in the time/accuracy relationship to the baseline duration of the response.

We start with the data sets consisting of rapid tasks. Results are fairly heterogeneous. One fairly universal finding (rotation and set being exceptions) is that, across values of p_0 , shorter responses are those that are likely to benefit from some increase in accuracy if they are marginally longer (i.e., the left side of each panel tends to be blue); this is perhaps due to marginally longer responses being less due to rapid guessing. The boundary between blue and red also tends to slope from upper left to bottom right such that, for a constant RT, marginal increases are more likely to be in the blue as opposed to the red if they represent more challenging interactions. Consider the Add Subtract data set. If $\log(t) = 1.8$ and $p_0 \approx 0.5$, we observe $\frac{\partial f}{\partial \log t} > 0$, while if $p_0 \approx 0.8$, we observe $\frac{\partial f}{\partial \log t} < 0$.

With less rapid tasks, many of the same patterns appear. In particular, we observe larger blue regions on the left and boundaries between blue and red regions tend to be negatively sloped. However, there are also cases where the partial derivative is uniformly positive (e.g., PIAAC) or negative (e.g., HRS). All told, these analyses suggest that whether the SAT holds may vary both across the nature of the task but also as a function of the precise conditions within the set of tasks in a given data set.

3.3. Item-Level Heterogeneity

Using a modified approach (e.g., Equation 4), we focus on SAT curves for individual items. We focus on the marginal effect of time net of p_0 . Results are shown in Table 2 focusing on only those items that have at least 100 responses. Given that each data set contained numerous items, we identified those items showing positive/negative marginal associations with time based on the estimates of β_1 that were significant after adjusting (via Bonferroni correction) for multiple testing of all items within data set.

In general, associations tended to be positive or null. However, note that, for example, the chess data that had a relatively large proportion of items show a negative association and nearly all data had at least some items that showed negative associations; we speculate on the reasons for such negative associations

TABLE 2.
Item-Level Analysis for Those Items With > 100 Responses

Data	<i>N</i> Items	% ($\beta_1 > 0$)	% ($\beta_1 < 0$)	$r(\beta_1, \delta_i)$	Confident Interval- Lower	Confident Interval- Upper
Lexical	15	40	13	0.22	−0.33	0.66
RR98 accuracy	32	0	0	−0.11	−0.44	0.25
Hearts flowers	8	12	12	0.86	0.39	0.97
Lexical Decision Task	495	1	0	−0.14	−0.22	−0.05
ECLS flanker	20	70	10	0.78	0.52	0.91
ECLS DCCS	30	40	0	0.87	0.74	0.94
Motion	30	13	10	−0.48	−0.71	−0.14
Multi-Source Interference Test	24	50	0	0.70	0.41	0.86
Reading fluency	292	10	4	−0.13	−0.24	−0.01
Reading comp	408	11	2	−0.40	−0.48	−0.32
Arithmetic	170	31	1	0.25	0.11	0.39
Groupitizing	88	59	0	0.29	0.08	0.47
Rotation	10	0	0	−0.04	−0.65	0.61
Set	10	0	80	−0.41	−0.82	0.30
Letter chaos	10	20	0	0.30	−0.40	0.78
Add subtract	60	38	7	−0.11	−0.35	0.15
Working memory	4	0	75	0.83	−0.64	1.00
Mult. div.	60	3	63	−0.05	−0.30	0.20
Health and Retirement Study	20	5	65	0.17	−0.30	0.57
Chess	80	5	26	−0.03	−0.25	0.19
PISA reading	218	39	16	−0.25	−0.37	−0.12
PERC	15	13	40	−0.26	−0.68	0.29
MITRE-ETS	95	13	2	−0.63	−0.73	−0.49
Assistments	604	0	1	0.10	0.02	0.18
National Social Life, Health, and Aging Project	13	8	54	0.21	−0.38	0.68
Programme for the International Assessment of Adult Competencies	104	71	0	0.53	0.37	0.65
PISA math	60	28	15	0.05	−0.20	0.30
NWEA Grade 3	3,694	3	0	−0.09	−0.13	−0.06
NWEA Grade 8	3,331	3	2	−0.02	−0.06	0.01

Note. The percentage of items showing positive or negative coefficients of $\log(t)$ predicting accuracy (e.g., estimates of β_1 from Equation 4) is those that remain after Bonferroni correction. Only significant correlations between difficulty and β_1 are shown. PISA = Programme for International Student Assessment; ECLS = Early Childhood Longitudinal Studies; DCCS = dimensional change card sort; PERC = Persistence, effort, resilience and challenge-seeking task.

TABLE 3.
Person-Level Associations Between Ability (θ), Speed (τ), and Variation in Speed (σ_{rank})

Data	$r(\theta, \tau)$	$\mathbb{E}(\sigma_{\text{rank}})$	$r(\theta, \sigma_{\text{rank}})$	Confident Interval- Lower	Confident Interval- Upper
Lexical	.15	.25	.06	-.14	.26
RR98 Accuracy	.17	.25	-.14	-.47	.24
Hearts flowers	-.20	.25	-.36	-.46	-.25
Lexical Decision Task	-.05	.22	-.61	-.72	-.47
ECLS flanker	-.05	.19	-.18	-.20	-.16
ECLS DCCS	-.11	.23	-.22	-.24	-.20
Motion	.06	.26	-.43	-.57	-.26
Multi-Source Interference Test	-.23	.24	-.27	-.34	-.20
Reading fluency	-.03	.21	-.12	-.15	-.09
Reading comp	-.28	.22	-.20	-.23	-.17
Arithmetic	.19	.22	-.56	-.61	-.52
Groupitizing	-.46	.24	-.41	-.48	-.33
Rotation	.12	.22	-.05	-.25	.15
Set	.16	.25	-.11	-.21	-.01
Letter chaos	-.08	.22	-.20	-.32	-.07
Add subtract	.08	.23	-.13	-.14	-.11
Working memory	.31	.23	-.08	-.22	.06
Mult. div.	.05	.24	-.13	-.15	-.11
Health and Retirement Study	.41	.24	-.06	-.10	-.02
Chess	.44	.24	-.01	-.13	.11
PISA reading	-.23	.25	-.28	-.29	-.27
PERC	-.43	.24	-.20	-.24	-.15
MITRE-ETS	-.62	.22	-.23	-.29	-.16
Assistments	-.36	.24	-.27	-.31	-.23
National Social Life, Health, and Aging Project	.30	.24	-.07	-.11	-.03
Programme for the International Assessment of Adult Competencies	-.33	.23	-.48	-.51	-.45
PISA math	-.15	.24	-.06	-.08	-.05
NWEA Grade 3	.02	.24	-.14	-.15	-.14
NWEA Grade 8	.04	.23	-.21	-.22	-.20

Note. PISA = Programme for International Student Assessment; ECLS = Early Childhood Longitudinal Studies; DCCS = dimensional change card sort; PERC = Persistence, effort, resilience and challenge-seeking task.

Speed–Accuracy Trade-Off?

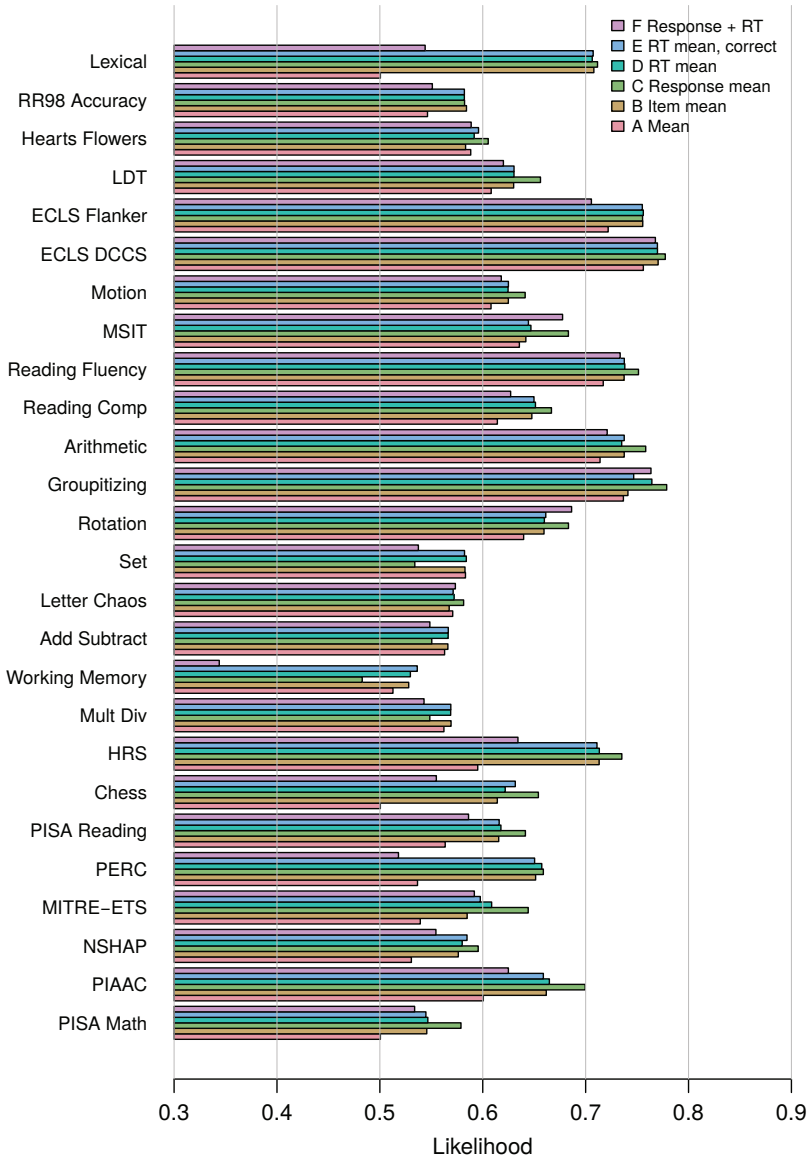


FIGURE 5. Comparison of out-of-sample predictions (via $\exp(\bar{\ell})$; see Section 2.2.4) in 10% holdout. Predictions are made based on predictors shown in the legend. (A) is based on the overall mean accuracy in the data (see Figure 2). (B) is based on the mean accuracy for each item. (C) is based on the mean accuracy for each person. (D and E) are based on the mean standardized response time for each person (with E focusing just on correct responses). (F) combines C and D.

in Section 4. We also investigated correlations between item difficulty and the marginal time/accuracy associations. Such associations varied widely across the data sets.

3.4. Person-Level Heterogeneity

We next analyze person-level speed via Equation 5. Results are shown in Table 3. We first consider correlations between the estimates of ability and speed. Correlations vary widely. In some cases, more able respondents are also faster (e.g., chess); in other cases, the opposite is true (e.g., the PIAAC and PISA).

We next consider within-person variation in speed during the test. We observed variation in speed—as indexed by changes in a respondent’s rank ordering of RT across items—that was fairly consistent across all the data sets although the Early Childhood Longitudinal Studies (ECLS) flanker tasks showed the least amount of within-person variation. This quantity has an interesting pattern of association with ability. Across nearly all data sets (lexical being the exception), respondents with larger estimates of θ showed less variation in speed. Although this association was not always significant, we think it suggestive of a potentially important insight regarding fluctuations in respondent speed across responses and resulting estimates of ability based on the collected responses.

3.5. Predictive Power of RT

Finally, we examine the predictive power of RT as compared to alternative predictors. Recall that all predictions of out-of-sample data are based on quantities computed in a training sample; fit is evaluated via $\exp(\bar{\ell}_{ij})$, where ℓ_{ij} is as in Equation 6. Results are shown in Figure 5. We focus here on three comparisons (denoted via letters assigned in Section 2.2.4 and referenced in Figure 5 legend). We ask how prediction changes when we exchange person-level response accuracy for person-level RT (C vs. D), exchange RT information for RT based only on correct items (D vs. E), and combine accuracy and RT information (F vs. C/D).

With respect to the first comparison (C vs. D), we generally make better predictions based on accuracy rather than RT. There are exceptions (ECLS flanker, set, add subtract, working memory, and mult. div.); we emphasize that, especially for data containing more complex tasks that take longer than 10 second, we are better able to predict novel responses using accuracy rather than RT. With respect to the second comparison (D vs. E), differences were quite small. In only two cases were differences larger than 0.01; in both cases (groupitizing and MITRE-ETS), prediction was superior when using all RT information. With respect to the third comparison (F vs. C/D), we generally find that prediction using both RT and accuracy is generally inferior to models based on just a single predictor (RT or accuracy). Similarly, results from analyses in the SI suggest that

using RT from an individual item response tends to degrade prediction as compared to predicting based on p_0 alone. In sum, these analyses—especially when combined with results from SI 3.4—suggest that RT may not be an especially useful predictor of accuracy in many cases. This could be due, in part, to the fact that additional time on an item may predict both positive and negative changes in accuracy (i.e., Figure 3).

4. Discussion

We use the standardized analysis of 29 RT data sets to study the interplay between speed and accuracy in nonexperimental settings. In nonexperimental settings, marginal increases in time do not necessarily lead to increased accuracy. In some cases, we observed patterns consistent with those predicted by the SAT but, in other cases, we do not. Accuracy occasionally declined with increased RT but more frequently showed an inconsistent relationship with increased RTs. In many cases, we observe a curvilinear relationship as anticipated by previous work (Chen et al., 2018). Further, there may be additional heterogeneity within a set of tasks when we stratify by the underlying difficulty (i.e., p_0) of the interaction. We also note that we saw few cases where the within-person independence of speed and accuracy (as indicated by a flat line in Figure 3) required by some models (e.g., van der Linden, 2007) held. While substantial experimental evidence (Heitz, 2014) indicates that artificial manipulation of time pressure has an effect on accuracy, our findings suggest that other factors may be at work in observational data and generally tend to reduce the role of the SAT as a sufficient first-order explanation for observed behavior.

When we observe results inconsistent with the SAT, what might explain such behavior? The tasks we consider are varied and may not allow for a single explanation but several psychological phenomena may be relevant. One possibility is that reductions in accuracy for long responses are associated with a decline in goal-oriented action (i.e., mind wandering; Smallwood and Schooler, 2015), but we note two challenges to articulation of precise theoretical mechanisms. The first is the wide range of tasks considered here. The second is that we have attempted to conduct analysis within item and person. For example, specific item features may induce attentional capture (Simons, 2000), thus leading to decreased accuracy for longer RTs for those items. However, our analysis is not focusing on between-item differences, thus reducing item- or person-specific features as potential explanatory factors.

Focusing on respondents, we observe inconsistent relationships between respondent speed and ability. While faster respondents are not necessarily more able, we do observe a consistent relationship between variation in respondent speed across items and their ability; respondents who receive lower estimates of ability tend to vary their speed more. Such variation in speed could be a

phenotype worth further study. Previous work suggests that such variation tends to predict cognitive aging in older samples (Lövdén et al., 2007).

Indeed, one substantively interesting case wherein the SAT does not hold involves older respondents (i.e., the HRS and NSHAP). In these data, additional time predicts a decrease in accuracy. We suspect that this finding has to do with both the nature of cognition in older respondents and the tasks in question. With respect to the age of the respondents, they may be experiencing cognitive aging (Tucker-Drob, 2019), an age-related decline in cognitive functioning. For respondents experiencing cognitive aging, it is possible that a within-person reduction in response speed isn't associated with deliberation and increased accuracy but, rather, confusion and decreased accuracy. Our findings can be read alongside others suggesting a change in the SAT (Heitz, 2014; Salthouse, 1979) as respondents age.

Turning to items, we identify those for which longer RT predicts increased accuracy—as anticipated by the SAT—as well as others for which the opposite is true. This inconsistency across items is one reason that RT is of limited predictive value. This limited predictive utility is also apparent in Figure 3 as curves showing association between time and accuracy are either relatively flat or otherwise not monotonic in many cases. Generally, RT is typically less useful than accuracy in predicting out-of-sample responses. That said, we would also advocate for extensive interrogation of, for example, the assumption that within-person variation in speed is unassociated with accuracy. Such an assumption is key to some models (van der Linden, 2007); our findings suggest that such a relationship may be complex and context-dependent but are largely inconsistent with the notion that such variation can be entirely ignored in attempts to better understand accuracy.

Our work connects with other recent studies of the SAT. Recent work suggests combining speed and accuracy into a single metric (Hughes et al., 2014; Liesefeld & Janczyk, 2019; Vandierendonck, 2017, 2018). Our findings suggest that there may be between-task heterogeneity that necessitates caution in development of such metrics. Our results also suggest, dovetailing with others, that accuracy is integral for understanding individual differences (i.e., RT alone may be insufficient; Draheim et al., 2019; Draheim et al., 2020). Our approach—emphasizing multiple data sets and nonparametric models—could also be incorporated into further tests of newly developed models (e.g., Kang et al., 2021). Future work could also examine the degree to which our approach could be used as a test of whether models that make restrictions on the SAT are appropriate; for example, identification of a fairly flat curve using the approach of Figure 3 could be a positive sign that the hierarchical model (van der Linden, 2007) could be used.

We also argue that our work documents a range of empirical phenomena that may exist in observational settings. These findings suggest that RT data are rich and may offer wide-ranging information about respondent behavior and the

functioning of a measure. Future work should focus on an exploration of such riches. In a given data set, we'd advocate for substantial investigatory work prior to the application of models based on relatively strong assumptions, given that our results suggest many different potential behaviors, several of which may violate necessary assumptions.

We acknowledge limitations. Other features of data collection may be relevant. We have not addressed ordering effects (Debeer & Janssen, 2013; Domingue et al., 2021; Vida et al., 2021). There are presumably motivational differences across the data sets that we do not measure and cannot study. There is evidence to suggest that emotional states—e.g., worry (Hallion et al., 2020)—that may vary as a function of motivational differences and/or testing pressure may affect the SAT.

We also note assumptions required by our analytic approach. The Rasch model that we use is relatively restrictive and unlikely to capture all of the features of the relevant item response functions; this may induce bias in Figure 4 if estimates of p_0 are distorted. We did consider the 2PL in supplemental analyses, but future work could further investigate whether still other item response models may offer different perspectives on these issues. There are also cases where our ability to identify items (e.g., working memory) is relatively weak in the sense that we are classifying a relatively broad class of tasks as a single item. In other cases (e.g., assistments), the assumption of a static ability may be inappropriate. We think that the potential insights from a common analysis applied to a broad variety of data sets offer great value, but findings should be interpreted in light of these limitations.

The heterogeneity of our findings suggests that there are many occasions wherein additional RT is not necessarily associated with an increase in accuracy. We argue that this suggests a need to be vary about the assumption that the SAT is a viable first-order descriptor of behavior in data wherein time pressure is not being explicitly manipulated, especially for challenging cognitive tasks. It may indeed be useful in describing behavior in some settings, but this assumption requires empirical verification. In observational settings, people vary their speed for a variety of reasons that diverge from the reasons that people vary their speed in the context of experimental SAT studies. When one manipulates time pressure with appropriate cognitive tasks, one observes the SAT. However, in broader settings, people are making decisions that affect speed and accuracy for lots of reasons, not all of which lead to results anticipated by the SAT.

Acknowledgments

The authors acknowledge the support of the iLEAD Consortium's investigators: Melina Uncapher, Adam Gazzaley, Joaquin Anguera, Silvia Bunge, Fumiko Hoeft, Bruce McCandliss, Jyoti Mishra, and Miriam Rosenberg-Lee.

Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This work was supported in part by the Institute of Education Sciences (R305B140009) and a gift from an anonymous donor. The Health and Retirement Study is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

ORCID iDs

Benjamin W. Domingue  <https://orcid.org/0000-0002-3894-9049>

Matthieu Brinkhuis  <https://orcid.org/0000-0003-1054-6683>

James Soland  <https://orcid.org/0000-0001-8895-2871>

Notes

1. In some cases (e.g., the hearts and flowers data), time pressure is manipulated across blocks. We examine this variation in the SI but focus here on a single block with constant time pressure. In other cases (e.g., the reading fluency and comp data), the test as a whole was timed, but there was not intentional variation of the time pressure across tasks.
2. In a few cases (e.g., National Social Life, Health, and Aging Project), we dichotomized polytomously scored responses so as to increase the number of available items.
3. An analytic plan was registered on June 1, 2020, https://osf.io/eqrd6/?view_only=ae099af11ed54c09b9fd0844a2f93a7a. We do not describe this as a preregistration as it was registered following preliminary analysis of some data. Further, as described in the SI, we have made some (relatively modest) adjustments to this analytic plan.
4. As used here, B-splines are a map from \mathbb{R}^1 to \mathbb{R}^K , where K is specified by the user; our use is similar to previous work (Domingue et al., 2021). To implement this mapping, we use $K = 4$ and the defaults in the `bs` function in R (i.e., splines are cubic). Illustrations of spline transformations can be seen in, for example, Figure 5.20 of Friedman et al. (2001) or Figure 1 of Woods and Thissen (2006).
5. We note one important limitations of this analysis. Data collected in an adaptive fashion lead to potential concentration of respondents into certain items.
6. Note that we omit both the NWEA and assistments data from this analysis, given the fact that the first data are adaptive and the second data may have dynamics in ability that are poorly captured by our assumption of a constant θ_j .

7. So as to make comparisons between relatively similar bits of information, we focus on predictions based on quantities computed in relatively comparable manners instead of focusing on, for example, the item response theory–based probability p_0 .
8. The analyses presented in the SI are the ones proposed in the original registration.
9. We focus here on $\log t$, but results are similar when we consider results in seconds, see SI.

References

- Bergé, L. (2018). *Efficient estimation of maximum likelihood models with multiple fixed-effects: The r package fenmlm* (tech. rep.). Center for Research in Economic Analysis, University of Luxembourg.
- Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in psychology, 9*, 1525.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software, 48*(6), 1–29.
- Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence, 69*, 16–23.
- Cheung, Y. B. (2007). A modified least–squares regression approach to the estimation of risk difference. *American Journal of Epidemiology, 166*(11), 1337–1344.
- Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*(11), 2037–2078.
- Debeer, D., & Janssen, R. (2013). Modeling item–position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164–185.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software, 39*(12), 1–28.
- De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102.
- Dennis, I., & Evans, J. S. B. (1996). The speed–error trade-off problem in psychometric testing. *British Journal of Psychology, 87*(1), 105–129.
- Domingue, B. W., Kanopka, K., Stenhaug, B., Soland, J., Kuhfeld, M., Wise, S., & Piech, C. (2021). Variation in respondent speed and its implications: Evidence from an adaptive testing scenario. *Journal of Educational Measurement, 58*(3), 335–363.
- Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin, 145*(5), 508.
- Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2020). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General, 150*(2), 242–275.

- Fife, D. A., & Rodgers, J. L. (2021). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the “replication crisis.” *American Psychologist*. doi: <https://doi.org/10.1037/amp0000886>
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics.
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608.
- Hallion, L. S., Kusmierski, S. N., & Caulfield, M. K. (2020). Worry alters speed-accuracy tradeoffs but does not impair sustained attention. *Behaviour Research and Therapy, 128*, 103597.
- Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, 150.
- Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T., & Bunting, M. F. (2014). Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods, 46*(3), 702–721.
- Jaccard, J., & Brinberg, M. (2021). Monte Carlo simulations using extant data to mimic populations: Applications to the modified linear probability model and logistic regression. *Psychological Methods, 26*(4), 450–465.
- Kang, I., De Boeck, P., & Ratcliff, R. (2021). Modeling conditional dependence of response accuracy and response time with the diffusion item response theory model. doi: <https://doi.org/10.1007/s11336-021-09819-5>
- Lappin, J. S., & Disch, K. (1972). The latency operating characteristic: Ii. effects of visual stimulus intensity on choice reaction time. *Journal of Experimental Psychology, 93*(2), 367.
- Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods, 51*(1), 40–60.
- Lövdén, M., Li, S.-C., Shing, Y. L., & Lindenberger, U. (2007). Within-person trial-to-trial variability pre-cedes and predicts cognitive decline in old and very old age: Longitudinal data from the berlin aging study. *Neuropsychologia, 45*(12), 2827–2838.
- Maris, G., & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615–633.
- Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 205–228.
- Molenaar, D., Tuerlinckx, F., van der Maas, H. L., & Et, al. (2015). Fitting diffusion item response theory models for responses and response times using the r package diffirt. *Journal of Statistical Software, 66*(4), 1–34.
- Pachella, R. G., Fisher, D. F., & Karsh, R. (1968). Absolute judgments in speeded tasks: Quantification of the trade-off between speed and accuracy. *Psychonomic Science, 12*(6), 225–226.
- Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision, 5*(5), 1–1.

- Pew, R. W. (1969). The speed–accuracy operating characteristic. *Acta Psychologica*, 30, 16–26.
- Ranger, J., Kuhn, J., & Gaviria, J.-L. (2015). A race model for responses and response times in tests. *Psychometrika*, 80(3), 791–810.
- Ranger, J., Kuhn, J., & Pohl, S. (2021). Effects of motivation on the accuracy and speed of responding in tests: The speed-accuracy tradeoff revisited. *Measurement: Interdisciplinary Research and Perspectives*, 19(1), 15–38.
- Rasch, G. (1993). *Probabilistic models for some intelligence and attainment tests*. ERIC.
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281.
- Salthouse, T. A. (1979). Adult age and the speed-accuracy trade-off. *Ergonomics*, 22(7), 811–821.
- Schouten, J., & Bekker, J. (1967). Reaction time and accuracy. *Acta Psychologica*, 27, 143–153.
- Simons, D. J. (2000). Attentional capture and inattention blindness. *Trends in Cognitive Sciences*, 4(4), 147–155.
- Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66, 487–518.
- Su, S., & Davison, M. L. (2019). Improving the predictive validity of reading comprehension using response times of correct item responses. *Applied Measurement in Education*, 32(2), 166–182.
- Tucker–Drob, E. M. (2019). Cognitive aging and dementia: A life–span perspective. *Annual Review of Developmental Psychology*, 1, 177–196.
- Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3), 287.
- Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, 49(2), 653–673.
- Vandierendonck, A. (2018). Further tests of the utility of integrated speed–accuracy measures in task switching. *Journal of Cognition*, 1(1): 8.
- van Rijn, P. W., & Ali, U. S. (2018). A generalized speed–accuracy response model for dichotomous items. *Psychometrika*, 83(1), 109–131.
- Vida, L. J., Brinkhuis, M. J. S., & Bolsinova, M. (2021). Speeding up without loss of accuracy: Item position effects on performance in university exams. *Proceedings of the 14th International Conference on Educational Data Mining*. EDM 2021, Educational Data Mining.
- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68(3), 456–477.
- Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, 41(1), 67–85.
- Wise, S. L. (2015). Response time as an indicator of test taker speed: Assumptions meet reality. *Measurement: Interdisciplinary Research and Perspectives*, 13(3–4), 186–188.
- Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, 71(2), 281–301.

Authors

BENJAMIN W. DOMINGUE is an assistant professor at the Graduate School of Education at Stanford University. He is interested in psychometrics and quantitative methods.

KLINT KANOPKA is a PhD student at the Graduate School of Education at Stanford University. He is interested in psychometrics and machine learning.

BEN STENHAUG received a PhD from the Graduate School of Education at Stanford University. He is interested in psychometrics.

MICHAEL J. SULIK is a developmental psychologist who studies how children's experiences in the family and at school influence the development of self-regulation and social-emotional learning and how these skills contribute to mental health and academic success.

TANESIA BEVERLY received a PhD from the University of Connecticut. Her research focuses on how test takers interact with digital testing platforms, developing noncognitive assessments, and technical and practical issues in large-scale testing.

MATTHIEU BRINKHUIS is a program director at Utrecht University. His research interest is in bringing together psychometrics, applied data science, and learning technology to answer societal challenges.

RUHAN CIRCI is a senior quantitative researcher at AIR. In this role, she participates in the National Assessment of Educational Progress research study design, analysis, quality control, and report writing activities.

JESSICA FAUL is a research associate professor at the University of Michigan's Survey Research Center. Her work focuses on socioeconomic predictors of health and health disparities across the life course.

DANDAN LIAO, PhD, is a senior psychometrician at Cambium Assessment, Inc., formerly AIR Assessment. She leads the operational and research work for the Next Generation Science Standards Assessments in three states. Her research interests include standard and extended IRT models, local dependence in large-scale assessments, and process data.

BRUCE MCCANDLISS, PhD, is the head of the Educational Neuroscience Initiative at Stanford University where he is a professor in the Graduate School of Education and the Department of Psychology (by courtesy). His research uses the tools of developmental cognitive neuroscience to study individual differences and educational transformations in key cognitive skills, such as attention, literacy, and mathematics.

JELENA OBRADOVIĆ is an associate professor in the Developmental and Psychological Sciences program at the Stanford Graduate School of Education, where she directs Stanford Project for Adaptation and Resilience in Kids Lab (<https://sparklab.stanford.edu/>). Her research focuses on identifying caregiving and educational practices that promote children's self-regulation, learning, and well-being and on understanding how executive functions and stress physiology support child development. She also works

on developing pragmatic, scalable, and culturally relevant assessments of children’s skills, behaviors, and experiences.

CHRIS PIECH is an assistant professor of Computer Science Education at Stanford University. His research is in machine learning to understand human learning.

TENELLE PORTER is a postdoctoral researcher at the University of Pennsylvania and an incoming assistant professor at Ball State University.

PROJECT iLEAD CONSORTIUM is a multiuniversity NSF Science of Learning network, partnering with investigators at Stanford, UC Berkeley, UCSF, UC San Diego, Rutgers, and UC Davis to investigate how executive function contributes to academic achievement in middle childhood.

JAMES SOLAND is an assistant professor of quantitative methods at the University of Virginia and an affiliated research fellow at NWEA, an assessment nonprofit. His research is situated at the intersection of educational measurement, practice, and policy. Particular areas of emphasis include understanding how measurement decisions impact the estimates of treatment effects and psychological/social–emotional growth, as well as detecting and quantifying test/survey disengagement.

JON WEEKS is a senior measurement scientist at ETS. His work focuses on multiple forms of test linking, including equating and scaling.

STEVEN L. WISE is a senior research fellow at NWEA. He has published extensively during the past three decades in applied measurement, with particular emphases in computer-based testing and the psychology of test taking. In recent years, his research has focused primarily on practical methods for effectively dealing with the measurement challenges posed by test-taker disengagement on achievement tests.

JASON YEATMAN is an assistant professor in the Graduate School of Education and Division of Developmental and Behavioral Pediatrics at Stanford University. As the director of the Brain Development and Education Lab, the overarching goal of his research is to understand the mechanisms that underlie the process of learning to read, how these mechanisms differ in children with dyslexia, and to design literacy intervention programs that are effective across a wide spectrum of learning differences.

Manuscript received August 11, 2021

Revision received March 25, 2022

Accepted April 15, 2022