Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Variation in respondent speed and its implications: Evidence from an adaptive testing scenario

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Domingue, Benjamin	Stanford Graduate School of Education	
Kanopka, Klint	Stanford Graduate School of Education	
Stenhaug, Ben	Stanford Graduate School of Education	
Soland, James	NWEA	
Kuhfeld, Megan	NWEA	
Wise, Steve	NWEA	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2021

Check type of content being submitted and complete one of the following in the box below:

- ☑ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☐ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

Journal of Educational Measurement, 58(3), 335–363.

DOI or URL to published work (if available) https://onlinelibrary.wiley.com/doi/abs/10.1111/jedm.12291

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** Institute of Education Scie through **[Grant number]** R305B1400 to **Institution]** Stanford Junio. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Institute or the U.S. Department of Education.

Variation in respondent speed and its implications: Evidence from an adaptive testing scenario

Benjamin W. Domingue[1,a], Klint Kanopka[1], Ben Stenhaug[1], James Soland[2,3], Megan Kuhfeld[2], Steve Wise[2], and Chris Piech[4]

[1]Stanford Graduate School of Education

[2]NWEA

[3]University of Virginia

[4]Department of Computer Science, Stanford University

[a]bdomingu@stanford.edu

Abstract

The more frequent collection of response time data is leading to an increased need for an understanding of how such data can be included in measurement models. Models for response time have been advanced, but relatively limited large-scale empirical investigations have been conducted. We take advantage of a large dataset from the adaptive NWEA MAP Growth Reading Assessment to shed light on emergent features of response time behavior. We identify two behaviors in particular. The first, response acceleration, is a reduction in response time for responses that occur later in the assessment. We note that such reductions are heterogeneous as a function of estimated ability (lower ability estimates are associated with larger increases in acceleration) and that reductions in response time lead to lower accuracy relative to expectation for lower ability students. The second is within-person variation in the association between time usage and accuracy. Idiosyncratic within-person changes in response time have inconsistent implications for accuracy; in some cases additional response time predicts higher accuracy but in many cases additional response time predicts declines in accuracy. These findings have implications for models that incorporate response time and accuracy. Our approach may be useful in other studies of adaptive testing data.

Acknowledgements: This work was supported in part by the Institute of Education Sciences (R305B140009) and a gift from an anonymous donor.

Introduction

Response time is an important type of process data collected in many measurement settings. There are multiple rationales as to why such data may be of interest. Response time data may improve estimation of the relevant parameters of the item response function (van der Linden, 2007), may help to detect disengagement (Wise & Kuhfeld, 2021), and may be useful in further tuning adaptive algorithms (Fan, Wang, Chang, & Douglas, 2012). At the core of these rationales is the notion that such data may help us better understand response processes and, consequently, improve related measurement instruments. Given the increasing availability of response time data, there has been recent work on the development of statistical models for the joint distribution of response time and accuracy (van der Linden, 2007; Ranger, Kuhn, & Gaviria, 2015; Ratcliff, Smith, Brown, & McKoon, 2016; van Rijn & Ali, 2018).

While this initial work has offered many insights, we argue that there is still a need for large-scale empirical examinations of the full range of complexity that may underlie interplay between speed—as indexed by response time—and accuracy. Descriptive empirical work is largely missing from the literature yet is necessary to further inform subsequent development of statistical models. For example, an assumption that respondent speed is constant may be acceptable if there is relatively limited variation in speed or, alternatively, if variation in speed is unrelated to accuracy. In particular, there is relatively limited empirical data on the effect of variation in a person's speed during a test on performance. Here, we focused on using item responses from a computer adaptive test taken by a large number of students to better understand interplay between speed and accuracy. Our approach reveals important features of respondent behavior that point to shortcomings of existing approaches for incorporating response time while also suggesting opportunities for future work.

In contrast to much previous analysis of response time data coming from linear testing, we utilized data from adaptive testing. Given that adaptive testing is likely to be a common feature when item responses are collected in digital environments, the lack of research based on this type of data constitutes a major gap.[1] The key challenge in working with such data is that students are exposed to systematically different sets of items as a function of their ability, making it difficult to disentangle information about speed and accuracy. This feature of adaptive testing complicates methods for estimating speed-related features of items and respondents, given that missingness is not at random. Methods for analyzing response time data in the context of adaptation are sorely needed (and are being developed, see Kang, Zheng, & Chang, 2020); the approach we describe here could be valuable for analysis of other data from adaptive tests as well. However,

[1]There is a literature on using response time to refine adaptive testing (van der Linden & van Krimpen-Stoop, 2003; Fan et al., 2012), but this work often uses time as a supplement to observed responses so as to better select subsequent items and generally improve efficiency of the adaptive algorithm. We have a different aim: to characterize interplay between response time and accuracy.

analysis of adaptive testing also offers opportunities. For example, since items are used at different points in the particular sequence of items faced by individual students, one can potentially disentangle sequence effects from item effects.

In this paper, we utilize methods for examining variation in response time and, subsequently, variation in the degree to which response time is associated with accuracy. We use these methods first to illustrate a substantial decline in time usage later in the assessment. We then offer evidence that respondents who receive lower ability estimates tend to reduce time spent on items more than do respondents who receive higher ability estimates. Focusing next on within-person variation in time usage, we illustrate heterogeneity in the degree to which increased time spent on an item predicts increased accuracy. Indeed, in the data considered here, marginal within-person increases in time are generally associated with declines in accuracy. These empirical results may offer guidance for future conceptual work on response time data and, even if these results do not generalize to other assessment scenarios, the approach we utilize here may be useful.

Interplay of speed and accuracy

Response time has been of longstanding interest; there are comprehensive overviews of this literature (De Boeck & Jeon, 2019; Heitz, 2014). Here, we contrast two kinds of approaches for understanding interplay between time and accuracy. We begin with a discussion of statistical approaches that have focused on the joint distribution of response time and accuracy. We then discuss approaches based on conceptual models from psychology. In both cases, we emphasize the implications of changes in speed for accuracy.

The statistical approaches

Statistical approaches to understanding interplay of speed and accuracy focus on $h(x, t)$, the joint distribution of response accuracy (x) and response time (t). Some research has focused on models for relatively specialized scenarios. For example, Ratcliff et al. (2016) posited a specific data-generating model widely used in settings wherein "items" are exchangeable trials that take small amounts of time (typically less than 5 seconds), although extensions to other settings have also been developed (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). In contrast, van Rijn and Ali (2018) focused on scenarios wherein scoring emphasizes response time: rapid responses are explicitly weighted more heavily than slower responses. Here we focused on a broadly applicable model for the joint distribution (van der Linden, 2007) that has been widely used with data from educational measures.

The hierarchical model (van der Linden, 2007) posits that the joint distribution of dichotomously scored

item responses (x) and the associated response times (t) is

$$h(x,t|\theta,\tau,\delta,\gamma) = \prod_i \prod_j f(x_{ji}|\theta_j,\delta_i) g(t_{ji}|\tau_j,\gamma_i), \tag{1}$$

where i and j index parameters associated with items and persons respectively. There is great flexibility in the choice of $f()$ and $g()$. For responses, $f()$ is typically taken to be a standard item response theory (IRT) model, for example, the Rasch model

$$f(x_{ji}|\theta_j,\delta_i) = \Pr(x_{ji}=1|\theta_j) = \sigma(\theta_j - \delta_i). \tag{2}$$

Here, $\sigma(z) = \frac{1}{1+e^{-z}}$ (the standard logistic sigmoid), θ_j represents a person's ability, and δ_i represents an item's difficulty. For response times, a standard assumption is that the distribution of response times is lognormal (although other assumptions are possible):

$$\log(t_{ji}) \sim N\left(\gamma_i - \tau_j, \sigma_t^2\right), \tag{3}$$

where τ_j represents a person's speed and γ_i represents an item's time demand.

The hierarchical model relies upon specific assumptions regarding speed and ability. First, at the person level, speed and ability are assumed to be constant (note that τ_j and θ_j are indexed only by person, j). This need not be the case. In fact, some evidence (Wise, 2015b) has pointed towards respondent speed changing over the course of the assessment; our analysis focused on this issue as well. Second, this model makes strong assumptions about conditional independence. These include the typical conditional independence assumptions in standard IRT models for accuracy (e.g., responses are independent conditional on the relevant item and person parameters). We focus here on the novel components: (a) responses are independent of person- and item-level speed parameters conditional on θ_j and δ_i; and (b) response times are independent of the person- and item-level accuracy parameters conditional on τ_j and γ_i. These assumptions collectively imply that speed is constant within person and that between-person associations between speed and ability are captured by associations between θ and τ (i.e., $m(\theta,\tau)$ and $n(\gamma,\delta)$ are joint densities describing the relationship between the person- and item-level parameters respectively). Note that, at the population level, the model is agnostic about the nature of $m(\theta,\tau)$ (e.g., θ and τ may be positively or negatively correlated).

The conceptual approach

An alternative approach to understanding interplay of speed and accuracy is based on an important insight from experimental psychology. It is typically assumed that requiring a person, via an experimental intervention, to increase their speed (and thus reduce their time spent on an item) will result in a decrease in accuracy. This "speed-accuracy tradeoff" (Heitz, 2014) is interpreted as a purely within-person phenomenon (i.e., it does not have implications for $m(\theta, \tau)$ in the hierarchical model). We might anticipate that, even when changes in speed are not induced via experimental interventions, more deliberative work on the part of a respondent might be both slower and more accurate. If respondent speed varies during an assessment in this way, the speed-accuracy tradeoff may then have implications for scenarios wherein an individual's speed is not experimentally manipulated.

A related yet conceptually distinct strand of work (Molenaar, Bolsinova, & Vermunt, 2018; Molenaar & de Boeck, 2018; Molenaar, Rózsa, & Bolsinova, 2019; Molenaar, Tuerlinckx, & van der Maas, 2015) considers responses as coming from a mixture of response processes. In particular, responses are generated by either "fast" or "slow" processes and item parameters may vary as a function of the generating process. In some cases (Coomans, Hofman, Brinkhuis, van der Maas, & Maris, 2016), there is an additional supposition that fast responses are less likely to be correct.

As stated above, the assumptions of the hierarchical model imply static speed and ability; and thus no within-person change in position on the speed-accuracy curve dictating performance on a test. An individual's response time is dependent on a latent and fixed individual speed and any observed within-person variation in response time is orthogonal to accuracy. The speed-accuracy tradeoff focuses largely on experimental conditions wherein speed is directly manipulated. Many educational measurements may be positioned between the two approaches contrasted in this section. Respondent speed is not directly manipulated, but a variety of forces—time limits, disengagement, fatigue—may raise questions about the strong assumption of constant respondent speed. Empirical evidence pertaining to the effects that idiosyncratic within-person variation in response time—i.e., a respondent spending relatively more time on one item and relatively less on another—may have on accuracy are thus of great interest, especially given the ways in which such effects tend to be critical in statistical models for speed and accuracy. Our work provides such evidence.

Methods

NWEA MAP Growth Reading Assessments

We used data from 150 assessments delivered as a part of the MAP Growth suite from NWEA.[2] These assessments focus on student reading abilities (comprehension, vocabulary, and understanding of genres and texts) and are generally low stakes for students.[3] Assessments were administered at three time points per year during the 2014–2015 through 2017–2018 academic years across Grades 3–8 in two states. Assessments drew from the same item pool and were built to conform to similar standards. We thus grouped data from the assessments by grade. The vast majority of items in the pool were multiple choice items with four response options. The assessments did not have time limits for individual items or for the test overall (although, in some cases, schools may have enforced time limits due to logistical constraints). The lack of time limits suggests that the test was not "speeded" in the classic sense (Gulliksen, 1950).

Descriptive statistics showing the number of students and items, as well as the number of item responses, are shown in Table 1. For each grade, there were roughly 1.2 million students and nearly 50 million item responses. Tests were administered in a computer-adaptive setting; students responded to, on average, 39 items. Note that each test contained thousands of unique items—over 6,000 in each grade—due to NWEA's deployment of a large item bank.

Table 1: NWEA MAP Growth Reading sample sizes by grade

Grade	# Students	# Items	# Responses
3	1 270 377	6237	49 842 299
4	1 258 982	6267	49 272 026
5	1 256 432	6952	49 121 700
6	1 198 564	7951	46 278 802
7	1 177 322	7926	45 308 627
8	1 164 994	7918	44 868 453

NWEA uses the Rasch model for scaling. While the Rasch model is restrictive, NWEA conducted extensive fit analyses (see Section 5.6.1 of NWEA, 2019) to assemble the item bank deployed here. Along with item responses and response times, we utilized NWEA-estimated item difficulties to construct retrospective

[2] `https://www.nwea.org/the-map-suite/`

[3] The MAP is used by educators for a variety of purposes, e.g., informally, to assess student growth over time or, more formally, to assess the effectiveness of schools and/or teachers. The NWEA discusses uses for the MAP Growth suite here: `https://www.nwea.org/blog/2018/new-resource-guidance-on-administering-map-growth/`.

estimates of the probability of a correct response (i.e., p_0 in Equation 4) for a given person-item interaction. We use "retrospective" to emphasize the fact that all of a person's responses on the adaptive test were observed (cf., in cases where provisional ability estimates were used to inform selection of subsequent items, we would say they are "prospective").

We now characterize the response time data. Figure 1 Panel A shows associations between item difficulty and mean response time. Items tended to take less than 3 minutes (i.e., most average item response times were less than $\log 180 = 5.2$ seconds) of time on average, although some items took longer. As may be anticipated, harder items took longer. Rapid guesses usually reflect respondent behaviors that are not informative about ability (Wise & Kuhfeld, 2020) and that may otherwise lead to score distortion (Wise & Kingsbury, 2016; Wise & Kuhfeld, 2021). Thus, we conducted analyses after removing item responses classified as rapid guesses using thresholds previously identified (and used operationally) by NWEA (Soland, Wise, & Gao, 2019). We also only used responses that took less than 30 minutes of time as longer response times were likely due to a student taking a break from the assessment.

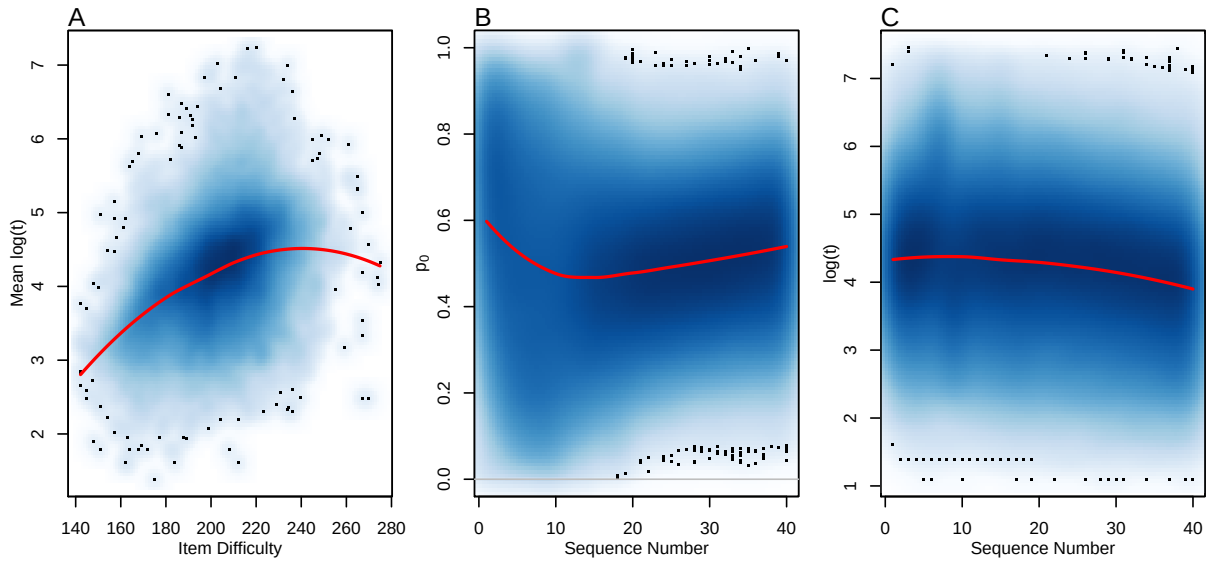

Figure 1: Data Description. Shaded regions represent density by color with darker shades of blue representing increasing density (using the `smoothScatter` function in R); isolated points represent outliers. Red lines represent LOESS fits to a subsample of the data. Panel A. Item-level associations between difficulty and mean $\log t$. Item difficulty is in RIT units, defined as $200 + 10 \cdot \theta$ where θ is on the logit scale (Thum & Hauser, 2015). Panel B. Response-level probabilities of a correct response (i.e., p_0) as a function of sequence number. Panel C. Response-level $\log t$ as a function of sequence number.

The adaptive algorithm of the MAP Growth Assessment is based on a phased structure. In the first phase (items 1–6), items are chosen to maximize information using the Maximum Fisher's information approach (Kingsbury & Zara, 1989) while also limiting exposure. The first item administered to a student is based on

a starting ability value that is half a logit *below* their ability estimate from their last test (with a system in place that borrows information from other subjects if necessary). In the second phase (roughly items 7–14), students are given items being field tested or that involve a common stimulus (items 7–10 and 15–18 were based on a common stimulus in some cases; note that response times for these items were collected in the same manner as for other items). Phase 3 (items 15–39) focuses on balancing the test per the underlying blueprint alongside maximizing information; Phase 4 (items 40–44) focuses on balancing information across the test blueprint's content areas.

Figure 1 Panel B illustrates the effect of adaptive testing on the probability of a correct response (using Equation 4). In the first phase, model-based probabilities were on average around 0.6. These fell to 0.5 but then slowly drifted upwards. Probabilities were also more widely dispersed at the beginning of the test as compared to the end. However, even late in the sequence (we use "sequence" to refer to the location where we observed an item-person interaction), there was substantial variability in the model-based probability. We emphasize this variability in the model-based probability across the sequence given that many of our models depend on relatively substantial variation in this quantity. Restrictions in this variability limit our ability to understand behavior for extreme probabilities (e.g., below 0.1 or above 0.9) later in the sequence, but we otherwise have relatively rich data in this regard. We also observed (Panel C) a decline in $\log t$ later in the sequence. This is early evidence for what we call "response acceleration," defined in more detail below.

Our modeling approach

We aimed to model response time and responses net of accuracy as predicted by the operational IRT model. To do so, we first constructed the IRT-based probability of correct response via the Rasch model:

$$p_{0,ji} = \Pr(x_{ji} = 1) = \sigma(\theta_{ji} - \delta_i). \tag{4}$$

We used these probabilities of response accuracy, which are of focal interest in adaptive testing, as predictors in our models. To compute $p_{0,ji}$, we used the estimate of item difficulty, δ_i, provided by NWEA. We estimated abilities, θ_{ji}, in a manner designed to remove information from item i in computations of $p_{0,ji}$, as such information leads to potentially inflated associations between x_{ji} and $p_{0,ji}$. We computed them as follows: for each response x_{ji}, we omitted that response from the individual's response string and used the remaining responses to estimate θ_{ji} via maximum likelihood.[4] Our logic is similar to that motivating the removal of items from the total score in analysis of item-total score correlations (Guilford, 1953; Henrysson,

[4]Correlations between p_0 values computed based on the θ_j ability estimates provided by NWEA and values based on our θ_{ji} were above 0.98 in all grades. We use the estimates of θ_j produced by NWEA for descriptive purposes where a single ability is needed for an individual.

1963) and the removal of focal items from matching criteria in studies of differential item functioning (Tan, Xiang, Dorans, & Qu, 2010).

Before further discussing our approach, we emphasize several of its key features. Given our interest in exploratory analyses, we used splines in order to leverage the large sample sizes so as to flexibly model nonlinearities in associations between quantities in a way that does not depend upon functional form assumptions. Splines have been used to relax functional form assumptions in item response functions (Winsberg, Thissen, & Wainer, 1984) and to flexibly model response time distributions (Bloxom, 1985). We utilized aspects of both of these approaches and considered spline representations of various quantities (in particular, p_0 and $\log t$). Several approaches to spline construction are available; we use B-splines (Woods & Thissen, 2006; Woods, 2006), which are piecewise polynomials. Additional detail is in Appendix A.

In contrast to the hierarchical model, wherein person and item effects are treated as random, we treated them as fixed (estimation via Bergé, 2018). This approach utilizes within-unit variation (Clarke, Crawford, Steele, & Vignoles, 2010). Consider the role of person fixed effects. By including fixed effects, we were able to analyze only the variation in the outcome that remained after we had adjusted for the mean of the person (note the equivalence of the fixed effect and differenced approaches in Section 2.2 of Clarke et al., 2010). Given that we were studying within-person variation, we eliminated as potential confounders any invariant features of the person, such as baseline processing speed or ability.

Inclusion of item fixed effects similarly controls for invariant features of an item, such as its baseline temporal demand. This analysis thus allowed us to focus specifically on whether a change in item location was associated with a change in the time respondents used to produce a response or whether a change in the response time for a person was associated with a change in accuracy. A consequence of this approach is that it cannot be used to study, for example, person-level features. We thus did not consider whether ability was associated with other quantities of interest (e.g., speed). Given our inclusion of fixed effects, we have used different notation for the fixed elements as compared to the random elements discussed above (e.g., compare τ_j in Equation 1 to ψ_j in Equation 5).

We first studied time usage and examined the consistency of time usage across the test. We argue that the time demand of an item should be fixed and that variation in response time is thus due to variation in respondent speed. We thus first standardized $\log t$ within item and then computed item-level means based on the item's position in the sequence. We then took averages over the values for each sequence position. After this descriptive analysis, we turned to a formal model of change in speed across the sequence, n_{ji} (i.e., if the first response in the assessment collected for person j were for item i, we would write $n_{ji} = 1$). The adaptive nature of the test—in particular, the fact that items appear in multiple positions in the sequence—is essential for inclusion of n_{ji}. Following recent work (Bolsinova, de Boeck, & Tijmstra, 2017), we also

allowed for response-specific time usage and included x_{ji} as a predictor. We focused on the following model for response times:

$$\log t_{ji} \sim N\left(\beta n_{ji} + \Phi(x_{ji}, p_{0,ji}) + \psi_j + \omega_i, \sigma_t^2\right), \tag{5}$$

where $\Phi(x_{ji}, p_{0,ji})$ is as shown in Appendix A. Our interest is in β, which reflects change in expected time as a function of an item's location in the sequence. Given that we modelled $\log t$, if β is estimated from Equation 5 then $\exp(\beta)$ gives the ratio $\mathbb{E}\left(\frac{t_{j,i+1}}{t_{j,i}}\right)$; that is, $\exp(\beta)$ tells us about the ratio of the expected time on the next item relative to the current item. This approach assumes that response times are lognormal. While most items had response time distributions more consistent with the lognormal than with Weibull or gamma alternatives, we also have evidence from tests based on the Shapiro-Wilkes statistic that the distributions were not fully consistent with the lognormal distribution (see Appendix B). We thus conducted sensitivity analyses to ensure that our results were not driven by this assumption.

We then examined potential implications of the variation in time usage for observed accuracy. To do this, we first estimated time spent on an item net of expectation given the item's position and item/person fixed effects. We did this by deriving residuals from

$$\log t_{ji} \sim N\left(\beta n_{ji} + \psi_j + \omega_i, \sigma_t^2\right). \tag{6}$$

These residuals reflect time spent on a response net of the position in which the item occurred—which is important given the pronounced change in time usage across the sequence—as well as person-specific speed and item-specific time demands. Note that these residuals are similar to z_{ji} in Bolsinova, Tijmstra, and Molenaar (2017). We then used the residuals, denoted Δ_{ji}, to predict accuracy in an attempt to understand how time usage may be differentially associated with accuracy as a function of observed respondent ability:

$$x_{ji} \sim N\left(\nu_1 \Delta_{ji} + \nu_2 n_{ji} + \nu_3 p_{0,ji} + \mu_j, \sigma_x^2\right). \tag{7}$$

In particular, ν_1 can be interpreted as the change in expected probability associated with the change in time based on Δ_{ji}. When $\Delta_{ji} = 1$ this implies $\frac{t_{ji}}{\mathbb{E}(t_{ji})} = e$ given that $\Delta_{ji} = \log t_{ji} - \log \mathbb{E}(t_{ji}) = \log(\frac{t_{ji}}{\mathbb{E}(t_{ji})})$. If, for example, we anticipate a respondent taking 30 seconds on an item, ν_1 describes our expectation with respect to a change in accuracy had they taken $30e \approx 82$ seconds. We estimated Equation 7 separately in data stratified by ability vigintile with the goal of identifying variation in ν_1 across vigintile. We discuss additional caveats related to estimation of such a model below, after we introduce Equation 8.

We used Equation 7 to study variation in accuracy as a function of systematic decline in time usage across the test. We now turn to a different question about accuracy in order to probe the implications of variation

in time usage for accuracy within respondent. To do this, we built upon the p_0 quantity derived from the basic Rasch model in Equation 4 to ask if other features—specifically, response time and sequence—predict deviations from p_0. We consider

$$x_{ji} \sim N\left(\Gamma_{ji}(n_{ji}, p_{0,ji}, \log t_{ji}) + \mu_j, \sigma_x^2\right), \quad [5] \tag{8}$$

where $\Gamma(n_{ji}, p_{0,ji}, \log t_{ji})$ is as shown in Appendix A. This approach was designed to produce time-accuracy curves; that is, curves that illuminate the within-person relationship between time usage and how it is associated with changing accuracy. We did this by calculating values of Γ for different combinations of its parameters.

We made two specific assumptions in Equation 8. First, we relied on a linear probability model. Though this is clearly a departure from convention, we think it justifiable given our prior that the departures from p_0 should be relatively small (i.e., $\Gamma_{ji} + \mu_j \approx p_0$). Use of the linear probability model was motivated by our desire to include the μ_j terms, which help to ensure that we were focusing on within-person variation in time usage. Second, we used p_0 from the Rasch model. Even after the fit analyses conducted by NWEA, the Rasch model may rely on overly strong assumptions regarding item functioning. We address the robustness of our findings with respect to these two assumptions in Appendix C.

To further unpack the impact on accuracy of marginal changes in time, we also considered the partial derivative with respect to time of the expectation in Equation 8. We estimated $\frac{\partial \Gamma}{\partial \log t}$ where Γ is as in Equation 8. We utilized estimates of this quantity to study variation in whether accuracy is increasing or decreasing in $n \times p_0 \times \log t$ space.

Results

Respondents accelerate during the course of testing; lower ability respondents accelerate more.

To explore the stability of time usage over the test, we focused on mean response time (standardized within item) across the sequence. Results are shown in Figure 2 Panel A. The clear trend is substantial decline across the sequence; that is, response times accelerate as the assessment progresses. Early in the sequence, responses took 0.1–0.3 standard deviations above average for an item; later in the sequence, this fell to 0.2 standard deviations below average. Given that the items themselves only changed their position in

[5] Note that, in contrast to Equation 5, we have not relied here on the assumption that t_{ji} is lognormal. Rather, we have rescaled t to minimize the effect of the skew of response times.

the sequence, the likely implication is that respondents' behavior changed across the sequence. As NWEA assessments do not have time limits, this behavior is not due to speededness. There is also a systematic deviation related to the sequence position of the common stimulus items (due seemingly to large amounts of time spent on the first item in the bundle, note the increase in Figure 2 Panel A at the beginning of each gray bar). We address issues associated with common stimulus items below in several sensitivity analyses.

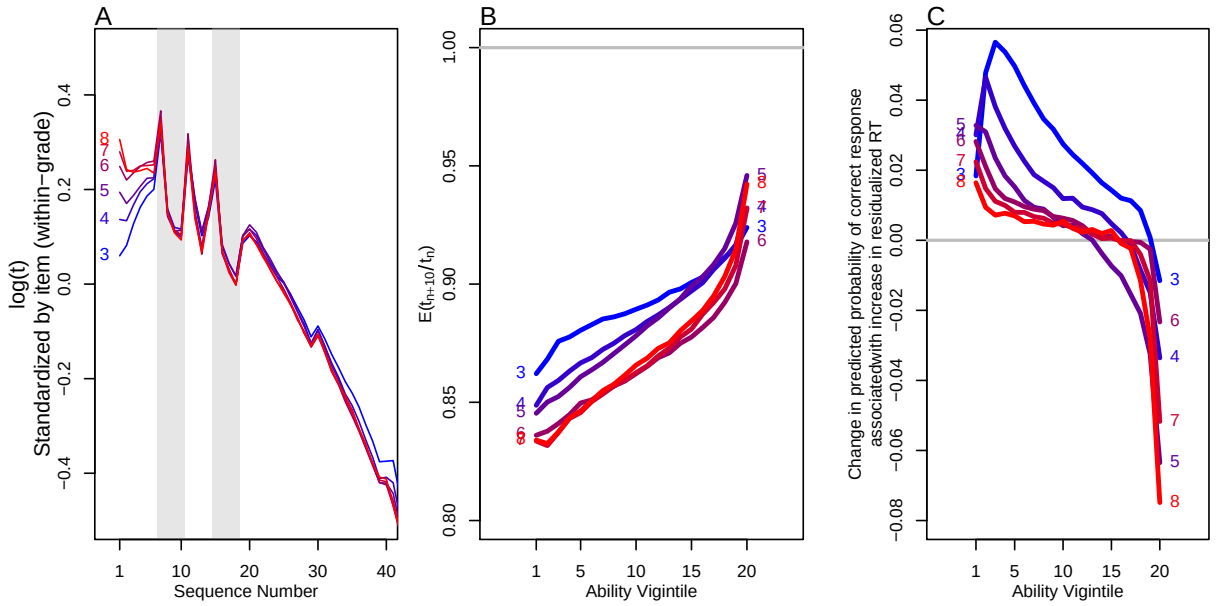

Figure 2: Analysis of response acceleration (different grades represented by colored lines). Panel A: Change in item-level mean response time (standardized) across all items as a function of sequence. Gray bars emphasize responses for common stimulus items. Panel B: Heterogeneity in response acceleration (by grade) as a function of estimated respondent ability; y-axis shows expected ratio of response time for a response that occurs 10 positions later in the sequence (see discussion of ν_1 in Equation 7). Panel C: Change in accuracy (by grade) associated with residualized response time.

The results in Panel A could be due to a changing sample of respondents who encounter the items at different positions. We now turn to an approach—Equation 5—designed to reduce such confounding by including fixed effects for item and person. Grade-level estimates of β are shown in Table 2 alongside the ratio of expected time for an item were it in a later position (in particular, 10 positions later) relative to the current position. Our data suggests that, relative to a given position in the sequence, a response to an item 10 positions later took only 88% as long (averaged across grades). Across the entire assessment, this led to a substantial decline in response time. For example, our data show that a response to an item in 30th position took ($100 \cdot 0.88^3 =$) 68% as long as a response to an item in first position. This is not a feature of adaptive testing: NWEA does not offer students items that take less time later in the test. Our approach holds items and persons fixed and controls for the key source of variation in the sequence introduced by adaptation (the IRT-based probability of a correct response). Results were comparable when we removed responses from

common stimulus items (see right-hand columns of Table 2).

Table 2: Estimates of response acceleration

	All items		No common stimulus items	
Grade	$\beta^{\dagger}$	$\mathbb{E}\left(\frac{t_{n+10}}{t_n}\right)$	$\beta^{\dagger}$	$\mathbb{E}\left(\frac{t_{n+10}}{t_n}\right)$
3	-0.0117	0.89	-0.0113	0.89
4	-0.0124	0.88	-0.0120	0.89
5	-0.0123	0.88	-0.0120	0.89
6	-0.0144	0.87	-0.0140	0.87
7	-0.0141	0.87	-0.0139	0.87
8	-0.0138	0.87	-0.0135	0.87

† Z statistics for all tests are greater than 900 in magnitude.

Given the pronounced response acceleration, we asked whether there is heterogeneity in response acceleration as a function of estimated respondent ability. Figure 2 (Panel B) shows estimates of response acceleration along the lines of those shown in Table 2 but now based on analysis of data stratified by vigintile of respondents' ability (using the θ_j ability provided by NWEA). We observed a patterning of response acceleration as a function of respondent ability. Respondents in the bottom vigintiles of ability showed a larger amount of response acceleration than did respondents in the top vigintiles. Compared to an initial response, a response 10 positions later from a respondent in the bottom vigintiles is expected to take $\approx 85\%$ the time of the initial response. In contrast, the response from a respondent in the top vigintiles took $> 90\%$ the time of the initial response. A sensitivity analysis probing the effects of the lognormal assumption is in Appendix B; results suggest similar trends.

Having documented variation in response acceleration, we next asked whether such variation might have implications for observed student performance. To study associations between time usage and accuracy, we first estimated stratified models wherein $\log(t)$ is first residualized (Equation 6) and then used to predict accuracy (Equation 7). If the estimated coefficient of Δ_{ji} from Equation 7 is greater than zero, this suggests that extra time spent on a response tended to lead to a gain in the marginal probability of a correct response. Results are in Figure 2 (Panel C). Respondents with low estimated abilities saw such a gain in terms of additional correct responses from marginal increases in time spent on items: when they spent more time on items than expected given where the item occurred in the sequence, they tended to respond more accurately. At the lowest vigintiles, an increase in time spent on an item might increase the probability of success by 2–4 percentile points (pp). Students in higher vigintiles tended to exhibit lower levels of increased

accuracy for extra time spent; gains vanished completely for those students in the top vigintiles.

Note the specific combination of findings from this set of analyses. Respondents with relatively low estimated abilities spent dramatically less time on later items. Also, they seemed to benefit more, in terms of increased accuracy, from more time spent on items. In tandem, these results have important implications for interpreting ability estimates. We further unpack these in the discussion. We now turn to an analysis that extends the logic of this question to examine the impacts of within-person variation in time usage more generally. We transition from asking about the potential heterogeneity in how time is associated with accuracy across different types of respondents to asking about heterogeneity in the relationship between time and accuracy within respondent.

Accuracy is affected (consistently) by item position and (inconsistently) by response time.

How might accuracy be affected by item position and response time net of expectation given by p_0? We have attempted to track such changes in accuracy using Equation 8. In modeling variation net of p_0, we effectively asked about variation in accuracy that may have been informed by features of the assessment—e.g., sequence effects, time usage on items—that were not accounted for in the basic IRT model (i.e., Equation 4). We began with results from Grade 3 (Figure 3 Panel A) focused on time. This panel shows time-accuracy curves that do not allow for sequence effects (i.e., n_{ji} was not included in the simplified version of Equation 8 used here; we also omitted interaction terms). The line types correspond to estimated accuracy for responses with model-based probabilities 0.2 below the overall mean, at the mean, and 0.2 above the mean for all responses in the grade. Solid and dotted lines thus indicate responses to relatively challenging and relatively easy items for a given person; dashed lines indicate responses to items with relatively middling difficulty for that person. In general, we observed increases in time translating into modest increases in the probability of correct responses (net of $p_{0,ji}$). Under this model, we expect responses that take ≈ 150 seconds to have a probability of accuracy that is 2pp higher than responses taking ≈ 20 seconds.

Given the pronounced role of sequence explored above, might the shape of these curves be affected by item position? To answer this question, we re-estimated Equation 8 now including n_{ji} and interaction terms (Figure 3 Panel B). Information on model fit is in Appendix C; inclusion of these additional terms improves model fit. We noted two phenomena. First, returns to accuracy due to an increase in time are inconsistent across p_0. Consider the lines of a similar type, all of which are based on the same underlying p_0. For responses with higher p_0 (solid lines), increases in time corresponded to relatively similar increases in accuracy (i.e., the solid lines are approximately parallel) irrespective of item position. In contrast, for responses with lower

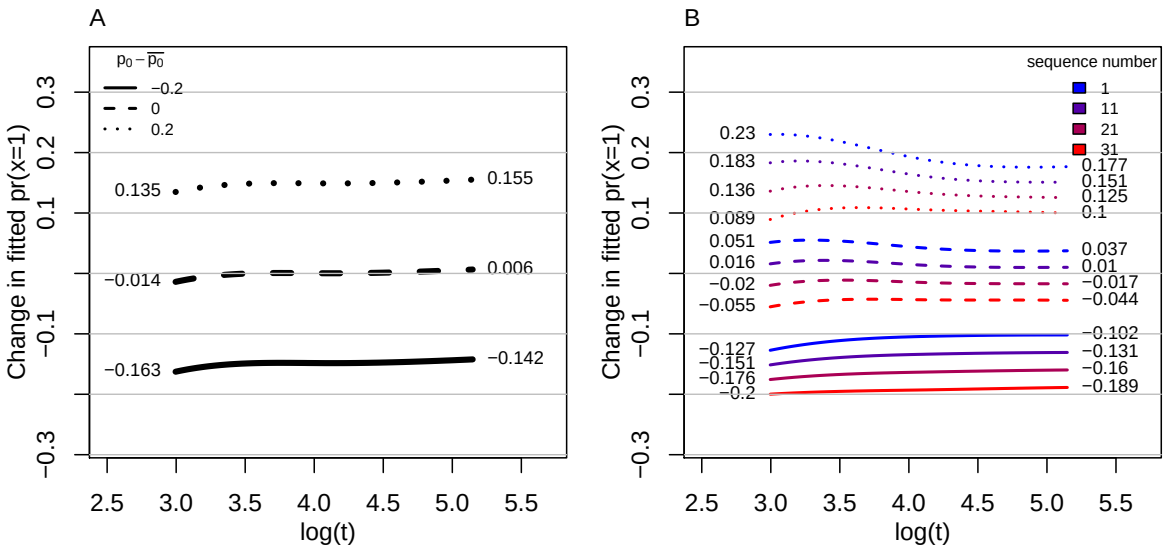

Figure 3: Within-person change in probability of a correct response as a function of response time for all Grade 3 responses. Estimates were derived from models that include item and person fixed effects. Line type (solid, dashed, dotted) represents responses with different IRT-based probabilities (i.e., p_0). Panel A: Black lines show main effect of time. Panel B: Colored lines allow for heterogeneity as a function of sequence effects.

p_0 (dotted lines), additional time early in the test (blue line) was associated with a decrease in accuracy and relatively little change in accuracy later in the test (red line). Second, the consistently large effect of sequence is quite clear. Vertical differences in the lines are associated with differences due to the sequence number, n. Responses earlier in the sequence had a higher probability of being correct than did responses later in the sequence (i.e., blue lines are consistently above red lines) net of person and item fixed effects and overall IRT-based probability. To take an extreme example, while a rapid response to an easy item early in the test had an associated increase in accuracy of 0.23 above the mean p_0, that same response time later in the test had an associated increase of only 0.09, a change of 14pp. Differences between early and late responses were smaller for responses that take more time (0.18 compared to 0.1, a change of only 8pp).

We next considered similar models across the full range of grades (Figure 4). As students age, the key change was a general tendency for the time-accuracy curves to be relatively flat or decreasing such that longer time spent on an item predicted either no change in accuracy or reduced accuracy. Consider Grade 8. Increased time was consistently associated with a decrease in accuracy irrespective of response position or challenge (i.e., p_0). However, the main effect of item position was still pronounced: responses later in the sequence continued to be associated with greatly reduced probabilities of a correct response.

We address the potential sensitivity of these findings to alternative specifications in Appendix C. Focusing on Grade 8, we considered several alternatives. First, we considered responses to items that are not based

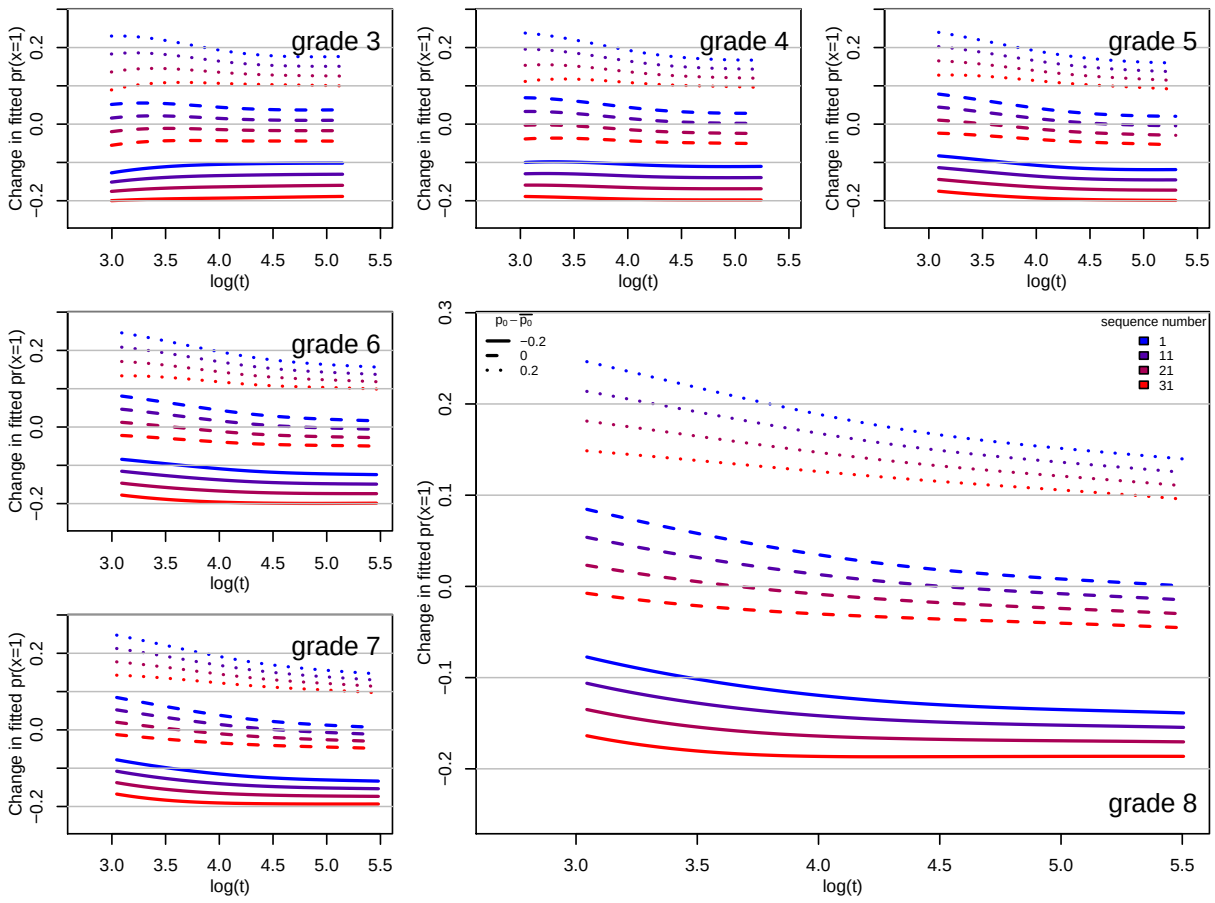

Figure 4: Within-person change in probability of a correct response as a function of the response time (separately by sequence number and IRT-based probability). Estimates were derived from models that include item and person fixed effects.

on a common stimulus. Second, we considered specification of Equation 8 based on a traditional logistic framework rather than our linear probability model. Third, we considered estimates of p_0 using the 2PL rather than the Rasch model. Results across all three alternative specifications were qualitatively similar to our main results. To illustrate the implications of our use of splines and θ_{ji}, we also considered results based on linear specifications and the ability estimates provided by NWEA; see Appendix C for additional details.

To draw attention to variation in whether changes were associated with increases or decreases in accuracy, we considered $\frac{\partial \Gamma}{\partial \log t}$. Estimates are in Figure 5, where blue represents regions wherein extra time spent on an item predicted an increase in accuracy, $\frac{\partial \Gamma}{\partial \log t} > 0$. Red regions, on the other hand, are those wherein an increase in time spent on an item predicted a decrease in accuracy, $\frac{\partial \Gamma}{\partial \log t} < 0$. We illustrate variation in these changes between -0.05 and 0.05; these can be interpreted as the change in probability associated with a one unit change in $\log t$ (based on a local linear approximation to Γ). Consider Grade 3. The fact that lines

in Figure 4 show both positive and negative slopes is apparent in Figure 5 given the existence of both blue and red regions. We can see that increases in time were most associated with declines in accuracy *earlier* in the test (i.e., there is relatively more red in the top row of panels). *Later* in the test (i.e., the bottom row of panels), we can observe a decline in the fraction of the panel shaded in red.

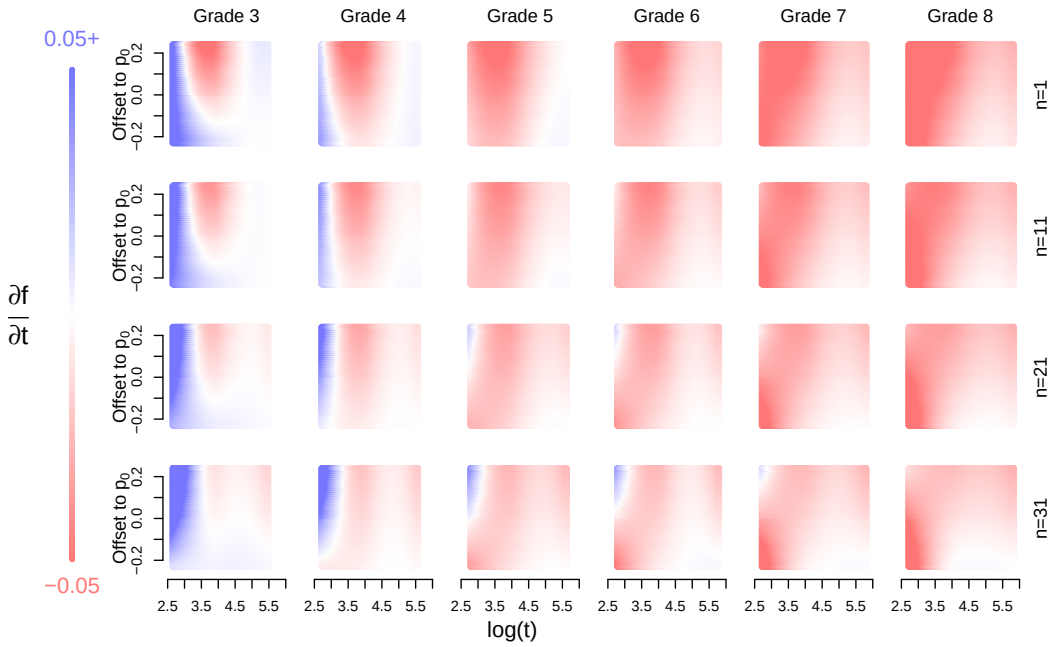

Figure 5: Changing within-person associations between marginal increases in time and accuracy. Partial derivatives ($\frac{\partial \Gamma}{\partial \log t}$, see Equation 8) indicating the effect of changes in time on accuracy. Blue regions indicate that additional time predicts increased accuracy; red regions indicate that additional time predicts decreased accuracy (net of p_0). Results are shown for Grades 3–8 and for different values of n.

Discussion

We used a large collection of item responses and response times collected in a typical reading measure for US students in Grades 3–8 to document several important features related to interplay between response time and accuracy. Using models that adjust for both the IRT-based probability of a correct response and person/item fixed effects, we have shown that response time is strongly sensitive to item position (i.e., response acceleration). We then explored variation in accuracy as a function of response time but allowing for heterogeneity across item position and challenge. We now discuss these findings and their connections to the wider literature.

Response acceleration has been previously observed (Wise, 2015b) and, more broadly, is related to issues of long-standing interest in educational measurement, for example the literature on item position effects

(Yen, 1980; Kingston & Dorans, 1984; Sykes & Fitzpatrick, 1992). We found item position effects to be such that items were more challenging when they were given later in the assessment, a phenomenon that has been known for some time (Yen, 1980) and is also of more recent interest (Debeer & Janssen, 2013; Nagy, Nagengast, Becker, Rose, & Frey, 2018; Weirich, Hecht, Penk, Roppelt, & Böhme, 2017; Hartig & Buchholz, 2012). There has also been recent interest in person- and item-level variation in such item position effects (Debeer & Janssen, 2013; Nagy et al., 2018; Weirich et al., 2017). Our findings are synergistic with this new area of work in supporting analysis of such differences and also offer novel information via the consideration of response time. We identified response acceleration that was heterogeneous across respondents. Respondents who received relatively low ability estimates tended to accelerate the most; this acceleration seems associated with a reduction in their performance. Future work could further examine variation in both item- and person-level parameters across the assessment and perhaps use response time as a feature for understanding such variation.

Response acceleration also resembles another feature of psychological measurement of long-standing interest in educational measurement: speededness (Gulliksen, 1950). We argue that response acceleration and speededness are distinct phenomena. Response acceleration is a behavior of the *respondents* while speededness is a property of the *measure* (or the conditions under which the measure is administered). Speededness occurs when a measure has time limits (or scoring rules) such that respondents must work quickly. When speededness is knowingly implemented in psychological measurement, speed explicitly becomes an important part of the construct. However, the MAP Assessment does not have a time limit. The response acceleration observed here is a phenomenon driven by individual choices; persons may engage in more or less acceleration. In our analysis, an individual's tendency to accelerate seems to have become implicitly embedded in the construct we measured given the findings of Figure 2 (Panels B & C).

This raises two questions: Why does acceleration happen and what does it mean for interpreting the resulting ability estimates? With respect to why, we propose two clear hypotheses but acknowledge that it is challenging to distinguish between them. The first is a diminishment of respondent effort. Recall that we have removed responses that are especially rapid, but this does not exclude the possibility that respondents generally "cared less" later in the test. The second is respondent fatigue. Some previous work (Ackerman & Kanfer, 2009; Jensen, Berry, & Kummer, 2013) has argued that fatigue effects are fairly minimal (note that both of these studies focused on university students; such findings may not generalize to our population). Given the minimal fatigue effects previously observed, this may seem to suggest effort diminishment as the more likely alternative. However, it is possible that the adaptive nature of the MAP exacerbates any fatigue effects since respondents are consistently being given items that require hard work. That is, when $p_0 \approx 0.5$, students may be having to constantly give maximal effort (thus resulting in increased fatigue). In general,

questions along these lines are a promising opportunity for future research wherein response time is further incorporated into work on item position effects.

Turning to the issue of interpretation, one key question is whether it is appropriate for ability estimates to have the observed property that students with lower ability estimates accelerate more than students with higher ability estimates. We considered this question from a test validity perspective (Wise, 2015a). One view is that this property is potentially a source of construct-irrelevant variance in the sense that it is largely due to a decline in effortful behavior on the part of the respondent. One salient distinction between our work and previous studies focused on noneffort (Wise & DeMars, 2010) is that we have attempted to remove the effects of rapid guessing behavior. While we have identified an increase in speed during the test, we do not view this increase as a cessation of meaningful test-taking activity (although it may certainly indicate diminished effort). Future work that, for example, compares "as-is" ability estimates to those that attempt to adjust for differences in response acceleration (e.g., asking whether such estimates differ in their ability to predict behavior at the next testing period) may help to resolve this question.

Following our analysis of response acceleration, our second key empirical observation was that both sequence and time contain predictive information about accuracy net of p_0. As the black lines in Figure 3 indicate, we generally found that an increase in response time was associated with a modest improvement in accuracy net of p_0, the IRT-based probability of a correct response. Returning to the speed-accuracy tradeoff discussed earlier, this finding was consistent with the general intuition underlying that conceptual model. However, the incorporation of item position into our model suggested a more complicated story, as depicted in Figure 5. If, to a first approximation, the concept of speed-accuracy was driving response processes, we would have observed far more blue regions than we did. Whether additional time predicted an increase or decrease in accuracy depended on both sequence location and challenge. Note that an analysis considering time linearly (as opposed to our nonlinear treatment via splines) may not be able to detect the range of variation observed here. The relationship between speed and accuracy also seemed to vary across respondent age (as proxied by grade).

These results suggest that the speed-accuracy tradeoff may not accurately characterize behavior in some settings (i.e., many regions of Figure 5 are red). In this sense, our findings are consistent with an alternative approach to studying this issue, namely response moderation models (Bolsinova, Tijmstra, & Molenaar, 2017). These models extend the typical item response function by suggesting that, for example,

$$\Pr(x_{ji} = 1|\theta_j, z_{ji}) = \sigma(\alpha_i\theta_j + \beta_i + \eta_{ji}z_{ji}), \tag{9}$$

where z_{ji} is a residualized indicator of response time (i.e., Was this response relatively fast or slow? See

also Equations 6 and 7). In Bolsinova, Tijmstra, and Molenaar (2017), different functional forms of η_{ji} were considered. Our results would be inconsistent with at least the simpler models that focus on z being item or person specific—their Model 1A ($\eta_{ji} = \lambda_i$) and Model 1B ($\eta_{ji} = \omega_j$)—given that our results suggest substantial heterogeneity within person and item. However, more recent work suggested curvilinear relationships between response time and accuracy (Bolsinova & Molenaar, 2018). In particular, their results suggested that responses are more likely to be accurate as they go from very fast to more moderately paced but that such increases in accuracy may stop or even reverse as response times go from average to long (see also Chen, De Boeck, Grady, Yang, and Waldschmidt (2018)). Such findings—although based on a different approach and without analysis of item position effects—echo ours.

We emphasize the salience of two issues for future research. First, adaptive testing data offer analytic challenges when it comes to studying response time. The study of response accuracy is relatively easy given that adaptive tests rely on precalibrated item parameters related to accuracy. No such information, however, is available related to the temporal demands of items. Given that items are given to nonrandom samples of students, estimation of such parameters is a challenge (to say nothing of the challenges associated with estimating something like the hierarchical model with sparse data of the type created by adaptation). We rely extensively on p_0 and item/person fixed effects in an attempt to surmount these issues. Such an approach could perhaps be utilized to conduct exploratory analyses whose results could inform subsequent modeling decisions (e.g., Should respondent speed be modeled as invariant?).

Second, our findings suggest that exploration of variation in individual speed and its potential effects over the course of an assessment is worthwhile. Empirical evidence could then be used to inform the choice of formal models for responses and response times. For example, in our case, item position would presumably need to be accounted for by any joint model $h(x, t)$. Note that, given the results shown in Figure 2 (Panels B–C), simply adjusting for item position may be insufficient since there was also noteworthy respondent-related variation in the degree to which position seemed to affect speed and accuracy. A related question is whether a model for $h(x, t)$ needs to allow for within-person variation in speed to affect accuracy. For the data discussed here, different approaches may be appropriate. The slopes of the curves in Figure 4 for Grade 3 are relatively flat; perhaps excluding such effects there is warranted. In contrast, the slopes in Grade 8 suggest rather large changes in accuracy as a function of within-person variation in speed; such changes may need to be accounted for in formal models of $h(x, t)$.[6]

Our findings have several implications for operational practice particularly in terms of how response time is handled. First, studies of fatigue and position effects should be prioritized. Second, the findings of

[6]We also note that our analytic approach recovers the anticipated "flat" within-person relationship between speed and accuracy in data simulated from the hierarchical model. In our view, this is further evidence of this approach's potential benefit. Code to demonstrate this is available at `https://gist.github.com/ben-domingue/14d3942c4194dd84cc8b3410e2dbeee5`.

such studies may need to be accounted for in subsequent interpretations of ability (i.e., How impacted by response acceleration are ability estimates? What does this tell us about the measured construct?). Third, given that our results suggest the potential for substantial heterogeneity in how response time is associated with accuracy, response time should be included in formal models of accuracy (or joint models of time and accuracy) only with great care.

These findings may not generalize to other settings. One limitation here pertains to the fact that the MAP Growth Assessment is relatively low stakes (as compared to a state-mandated assessment, for example). In particular, the response acceleration observed here might be more extreme than that observed in higher-stakes settings. A second limitation pertains to the Rasch model. We have used that model here given that it is the basis for the NWEA's work but note that others have studied item position effects in the context of more elaborate models—e.g., changes in item discrimination across the sequence (Nagy et al., 2018)—that we may not have been well positioned to detect with this data. That said, we do offer evidence that key findings are not sensitive to alternative specifications (e.g., the 2PL) of the p_0 quantity. We would further note that these results may not generalize to data collected from linear tests or that adaptation may enhance the role of certain processes (i.e., perhaps response acceleration is more pronounced when respondents are given a sequence of items with $p_0 \approx 0.5$ and wherein they cannot navigate between items).

Given the more frequent use of digital interfaces for the purposes of measurement, we anticipate a continued increase in the volume of response time data and its salience. As digital assessments collect more response time data, we think it prudent to be open to a need for changes in the assumptions brought to the analysis of such data. Our research has suggested that conceptually appealing ideas—e.g., all else equal, faster responses may be somewhat hurried and thus less likely to be correct—may not be appropriate in all cases. For process data such as response time to positively impact measurement techniques, we think that deeper empirical investigations along the lines of the one conducted here are needed. Such investigations will be useful in emphasizing the data features that are most relevant in a given scenario. Where such investigations suggest a clean alignment with existing theory, all the better. A failure to find such alignment offers possibilities for advancing our theoretical understanding and will suggest future directions for statistical modeling.

Appendices

Appendix A: Complete model specifications

We offer additional detail on Equations 5 and 8. In both cases, we used B-splines to allow for nonlinearities in key predictors. As noted in the main text, B-splines are piecewise polynomials.[7] We denoted the spline-based representation of a predictor z as $b(z)_k, k \in \{1, 2, 3, 4\}$. We also allowed for interactions among predictors.

We begin with Φ from Equation 5. We used B-splines to model p_0 as time may vary nonlinearly as a function of p_0. We defined

$$\Phi(x_{ji}, p_{0,ji}) = \alpha_1 x_{ji} + \sum_{k=1}^{4} \gamma_{0,k} b(p_0)_k + x_{ji} \sum_{k=1}^{4} \gamma_{1,k} b(p_0)_k, \tag{10}$$

where subscripted α and γ coefficients are parameters to be estimated.

Turning to Γ in Equation 8, we allowed for nonlinear effects in time by transforming $\log t$ via B-splines. We defined

$$\begin{aligned}
\Gamma(n_{ji}, p_{0,ji}, \log t_{ji}) = \alpha_1 n_{ji} + \alpha_2 p_{0,ji} + \alpha_3 n_{ji} \cdot p_{0,ji} \\
+ \sum_{k=1}^{4} \gamma_{0,k} b(\log t)_k \\
+ n_{ji} \sum_{k=1}^{4} \gamma_{1,k} b(\log t)_k \\
+ p_{0,ji} \sum_{k=1}^{4} \gamma_{2,k} b(\log t)_k \\
+ p_{0,ji} \cdot n_{ji} \sum_{k=1}^{4} \gamma_{3,k} b(\log t)_k.
\end{aligned} \tag{11}$$

Again, subscripted α and γ coefficients are parameters to be estimated.

Appendix B: Prediction of response time

Response time distributions

Given that there are different distributional choices for modeling response time, we considered the response time distributions for each item, within each grade. We first considered whether the lognormal, Weibull, or gamma distribution was preferred using Kolmogorov-Smirnov distance (computed via Delignette-Muller &

[7] B-splines allow for a flexible representation of a variable in a higher dimensional space. The user must specify this dimension; we denoted it as K. We chose a $K = 4$ representation so that each original value is represented as a 4-tuple. Illustrations of these transformations can be seen in, for example, Figure 5.20 of Friedman, Hastie, and Tibshirani (2001) or Figure 1 of Woods (2006). We used the defaults in the `bs` function (R Core Team, 2019) to construct these transformations (in particular, we used cubic splines) and conducted sensitivity analyses regarding our choice of $K = 4$.

Dutang, 2015). Results are in Table A1. In over 85% of cases (i.e., items in a grade), the Kolmogorov-Smirnov test preferred the lognormal distribution. More recent work (Sinharay & van Rijn, 2020) has suggested that the Shapiro-Wilkes test is a superior alternative. Despite the fact that the lognormal test was preferred to the two alternatives based on the Kolmogorov-Snirnov distance, we also tended to reject the null that the response time distributions were in fact lognormal. When the lognormal was the preferred distribution among the three, the average p-value for those items ranged between 0.005 and 0.01. The average p-value for all items was roughly an order of magnitude lower.

Table A1: Analysis of response time distributions

	Proportion[a]			Average p-value[b]	
Grade	Weibull	Gamma	lognormal	lognormal items	All items
3	0.020	0.123	0.857	4.657e-03	1.775e-04
4	0.020	0.105	0.875	5.236e-03	1.591e-04
5	0.020	0.096	0.883	4.371e-03	1.971e-04
6	0.034	0.094	0.872	1.034e-02	9.317e-04
7	0.029	0.089	0.882	1.071e-02	3.358e-03
8	0.028	0.084	0.887	1.035e-02	1.592e-03

[a] Preferred distribution based on Kolmogorov-Smirnov (Delignette-Muller & Dutang, 2015).
[b] p-value based on Shapiro-Wilkes test for each item. We separately considered all items and those items whose preferred distribution was the lognormal. For items with more than 5000 responses in a grade, we considered a random sample of 5000 responses.

Response acceleration focusing on rank-ordered response times

Figure A1 contains results based on an analysis similar to that in Figure 2 (Panel B, based on Equation 5) but wherein we attempted to minimize the effects of two issues. First, we removed responses to common stimulus items given the patterns shown by those items in Figure 2 (Panel A). Second, we conducted analyses on the item-specific quantiles of each response in order to minimize the effect of the lognormal assumption. Results based on these data still showed that lower ability students—as judged by the official NWEA ability estimate—tended to decrease time spent on later items more than did higher ability students.

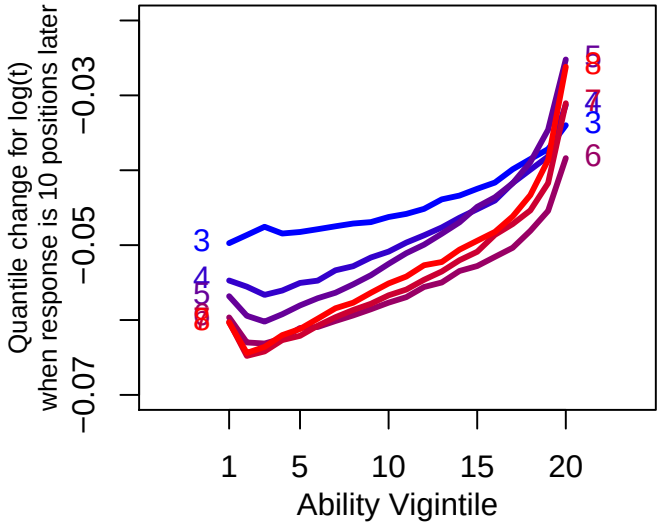

Figure A1: Estimates of response acceleration after removing responses to common stimulus items and where we modeled the change in the response time distribution (i.e., quantiles) rather than $\log t$.

Appendix C: Alternative specifications in prediction of item responses

We first offer additional information about estimates derived from Equation 8 as well as estimates of the error variance (i.e., σ_x^2 from Equation 8) in Table A2. We can compare fit in models with and without n_{ji} and interactions (i.e., a comparison of the panels of Figure 3); fit is improved in the expanded models. For example we can construct a likelihood ratio test for Grade 3 based on

$$\chi^2(-2(-34331095 + 34166779), 14) = \chi^2(328632, 14).$$

The probability associated with this test statistic is very small (i.e., $< 1^{-10}$). For each grade, models were based on nearly 50 million responses.

These estimates may be sensitive to several specific analytic choices. In Figure A2—which shows results comparable to those of Figure 3 but focusing on Grade 8 (the large number of items for this grade supported our strategy for estimating the 2PL)—we considered sensitivity of our findings to alternative specifications. The top right panel of Figure A2 contains results identical to those for Grade 8 in Figure 4 of the main text

for the purposes of comparison.

NWEA ability

We first note that results were relatively sensitive to whether ability is computed with or without x_{ji}. Our main analyses utilized ability estimates θ_{ji} computed without x_{ji}; for each response x_{ji}, we estimated an ability using all responses for a student other than x_{ji}. In Panel B of Figure A2, we used the NWEA ability estimate that includes all item responses for a student (i.e., the ability estimate is constant for a student). Resulting estimates showed some degree of sensitivity to this difference in abilities. Note, for example, the differences in the dotted lines associated with the easiest items. For the hardest items (solid lines), resulting changes in probability were more widely spaced when we used the θ_j ability estimates from NWEA as compared to our θ_{ji} estimates. Future work along the lines of, for example, Bolsinova, Tijmstra, & Molenaar (2017), may wish to consider similar sensitivity analyses.

Analysis of item responses without common stimulus items

We considered results based on item responses that did not include the common stimulus items (those in positions 7–10 or 15–18). Results (Panel C of Figure A2) were qualitatively quite similar to those in Panel A.

Sensitivity to choice of linear probability model

We considered results that focused on a logistic rather than a linear probability model. Specifically, we supposed that

$$\Pr(x_{ji} = 1) = \sigma\left(\Gamma(n_{ji}, p_{0,ji}, \log t_{ji})\right), \tag{12}$$

where Γ is as in Equation 8. Note that, for purposes of computational feasibility, we did not include person fixed effects here. Findings (Panel D of Figure A2) were again quite similar to those of the analysis presented in the main text.

Analysis based on 2PL model

We considered variation in estimates of p_0 (Equation 4) based on an alternative item response model (the 2PL), where

$$p_0 = \sigma(\alpha_i(\theta_j - \delta_i)). \tag{13}$$

However, estimation of this model was nontrivial given the adaptive nature of the test (recall that, in the Rasch-based analysis, we used the NWEA estimates of δ_i that had been obtained via independent analysis of different data). To estimate the 2PL, we used data from eighth grade students. We focused on students with at least 10 responses and items with at least 1000 responses. We then constructed a subset of roughly 1000 items by sequentially removing the 10 items with the least data so as to produce a set of item responses that had relatively high rates of completion relative to one another. We then estimated the 2PL based on a random sample of 100,000 students from this data.[8] For Grade 8, the correlation between the Rasch (via NWEA) and 2PL difficulty estimates was 0.77.

Results (Panel E of Figure A2) were again fairly similar to those in Panel A. One relatively small difference that we note so as to help emphasize the overall similarity is among responses with relatively high p_0 (solid lines). For those responses, the 2PL results suggested curves that are relatively flat or increasing as a function of time while the Rasch results suggested largely declining curves.

Analysis based on nonsplined response time

We considered results based on a linear treatment of logged response time (i.e., we did not utilize B-splines). Results (Panel F of Figure A2) continued to look qualitatively similar to Panel A. However, such results did not allow us to detect the potential flattening of the response time-accuracy curve, hence our preference for the spline-based approach in the main text.

Choice of spline basis

We considered a version of Figure 3 again based on Grade 8 wherein we varied the degree of the spline basis (i.e., the value of K). Results are in Figure A3. Qualitative patterns were consistent irrespective of the choice of spline basis. However, the additional flexibility offered by the higher degree is apparent in the bottom two panels: note the fluctuations in the solid curves around $\log t = 4.4$ for $K = 6$.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181.

[8] We used a lognormal prior with parameters 0 and 0.5 in the `mirt` specification (Chalmers, 2012) for the difficulty parameter.

Table A2: Fit statistics for estimates of Equation 8

	Without n_{ji}, main effects only[a]				With n_{ji} and interactions[b]			
Grade	# Responses	σ_x^2	log likelihood	AIC	# Responses	σ_x^2	log likelihood	AIC
3	49842299	0.238	-34331094.607	71202953.214	49842299	0.237	-34166779.121	70874350.243
4	49272026	0.239	-34017909.817	70553793.634	49272026	0.238	-33872772.088	70263546.175
5	49121700	0.238	-33854575.432	70222024.865	49121700	0.237	-33730942.210	69974786.419
6	46278802	0.239	-31941419.843	66279977.687	46278802	0.238	-31831406.136	66059978.272
7	45308627	0.238	-31197594.273	64749842.545	45308627	0.237	-31112097.423	64578876.847
8	44868453	0.237	-30798382.603	63926763.207	44868453	0.236	-30729392.239	63788810.478

[a] i.e., Figure 3 Panel A
[b] i.e., Figure 3 Panel B

Bergé, L. (2018). *Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm.* CREA Discussion Paper Series, 18-13. Center for Research in Economic Analysis, University of Luxembourg.

Bloxom, B. (1985). A constrained spline estimator of a hazard function. *Psychometrika*, *50*(3), 301–321.

Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*(4), 1126–1148.

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology*, *9*, 1525.

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 257–279.

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.

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. F. (2010). The choice between fixed and random effects models: Some considerations for educational research. *IZA Discussion Paper*, *5287*. Retrieved from `http://ftp.iza.org/dp5287.pdf`

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PLOS ONE*, *11*(5).

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in psychology*, *10*, 102.

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(4), 1–34.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*(5), 655–670.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning.* New York: Springer.

Guilford, J. P. (1953). The correlation of an item with a composite of the remaining items in a test. *Educational and Psychological Measurement*, *13*(1), 87–93.

Gulliksen, H. (1950). The reliability of speeded tests. *Psychometrika*, *15*(3), 259–269.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual

persistence. *Psychological Test and Assessment Modeling*, *54*(4), 418–431.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 150.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika*, *28*(2), 211–218.

Jensen, J. L., Berry, D. A., & Kummer, T. A. (2013). Investigating the effects of exam length on performance and cognitive fatigue. *PLOS ONE*, *8*(8).

Kang, H.-A., Zheng, Y., & Chang, H.-H. (2020). Online calibration of a joint model of item responses and response times in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *45*(2), 175–208.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359–375.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*(2), 147–154.

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., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297.

Molenaar, D., Rózsa, S., & Bolsinova, M. (2019). A heteroscedastic hidden Markov mixture model for responses and categorized response times. *Behavior Research Methods*, *51*(2), 676–696.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74.

Nagy, G., Nagengast, B., Becker, M., Rose, N., & Frey, A. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling*, *60*(2), 165–187.

NWEA. (2019). *MAP® Growth technical report.* Author Portland, OR.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ranger, J., Kuhn, J.-T., & Gaviria, J.-L. (2015). A race model for responses and response times in tests. *Psychometrika*, *80*(3), 791–810.

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.

Sinharay, S., & van Rijn, P. W. (2020). Assessing fit of the lognormal model for response times. *Journal of Educational and Behavioral Statistics*, *45*(5), 534–568.

Soland, J., Wise, S. L., & Gao, L. (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, *32*(2), 151–165.

Sykes, R. C., & Fitzpatrick, A. R. (1992). The stability of IRT b values. *Journal of Educational Measurement*, *29*(3), 201–211.

Tan, X., Xiang, B., Dorans, N. J., & Qu, Y. (2010). The value of the studied item in the matching criterion in differential item functioning (dif) analysis. *ETS Research Report Series*, *2010*(1), i–27.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth.* Portland, OR: NWEA.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287.

van der Linden, W. J., & van Krimpen-Stoop, E. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265.

van der Maas, H., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356.

van Rijn, P. W., & Ali, U. S. (2018). A generalized speed–accuracy response model for dichotomous items. *Psychometrika*, *83*(1), 109–131.

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115–129.

Winsberg, S., Thissen, D., & Wainer, H. (1984). Fitting item characteristic curves with spline functions. *ETS Research Report Series*, *1984*(2), i–14.

Wise, S. L. (2015a). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, *28*(3), 237–252.

Wise, S. L. (2015b). Response time as an indicator of test taker speed: Assumptions meet reality. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 186–188.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27–41.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86–105.

Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. Margolis & R. Feinberg (Eds.), *Integrating timing considerations to improve*

testing practices (pp. 150–164). New York, NY: Routledge.

Wise, S. L., & Kuhfeld, M. R. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130–149.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological methods*, *11*(3), 253–270.

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.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297–311.

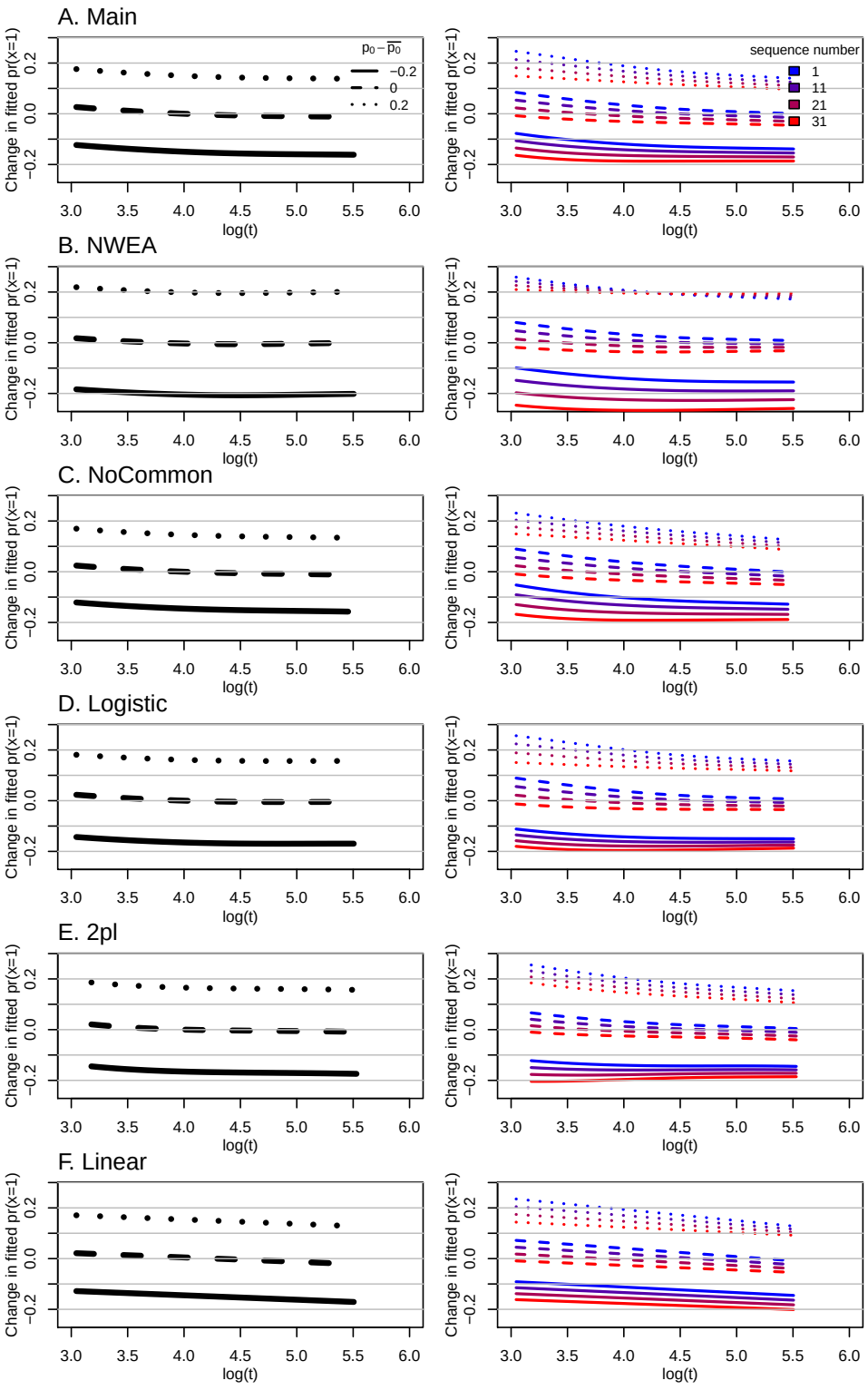

Figure A2: Comparison of (A) main results to those based on original NWEA ability estimates (B) and alternative specifications (C–F) for Grade 8. See Appendix C for details.

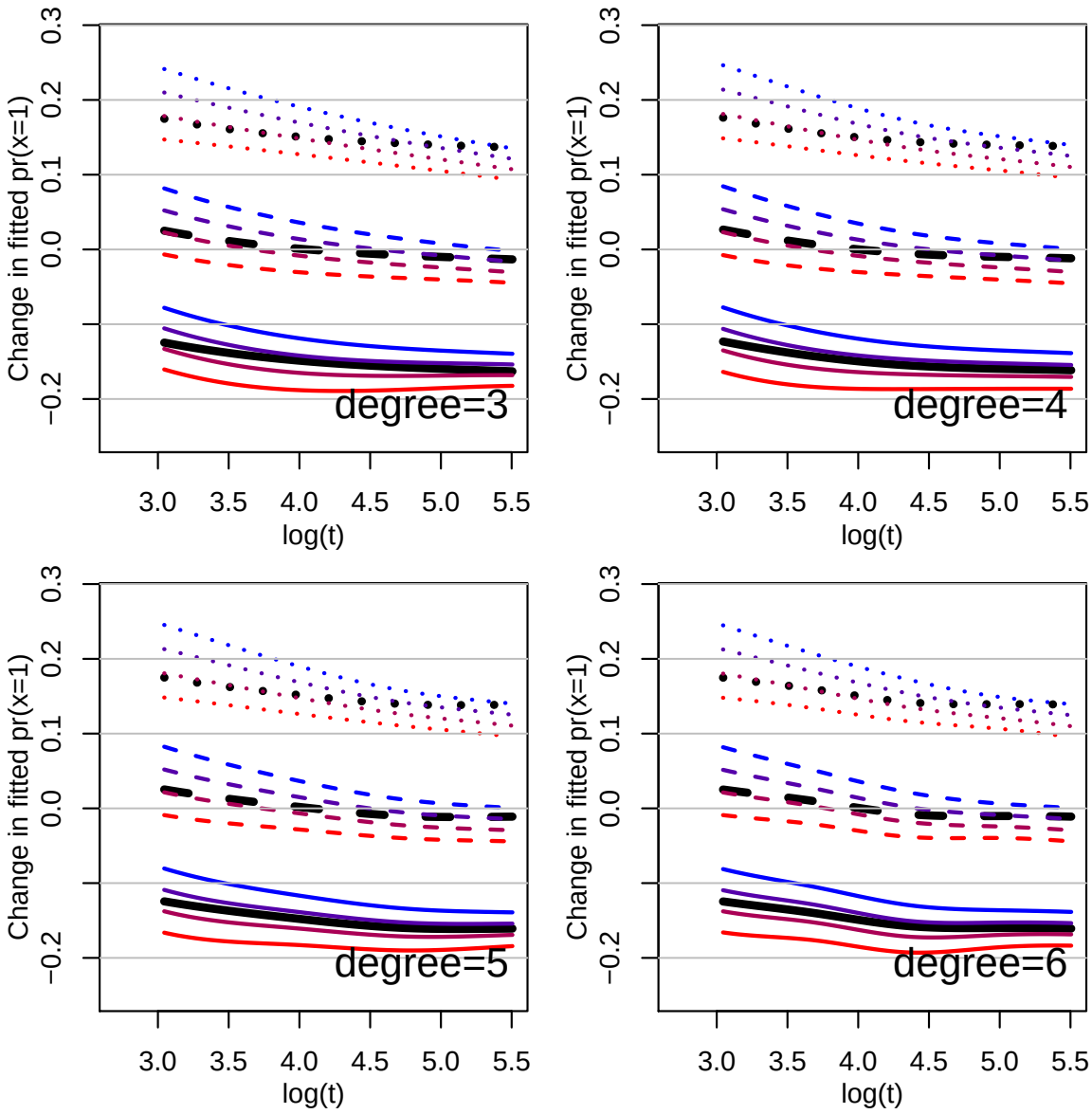

Figure A3: Sensitivity of prediction of accuracy to degree of B-spline basis (i.e., K) for Grade 8 data. Predicted probability of a correct response as a function of the response time (separately by sequence number and IRT-based probability).