Psychological Test and Assessment Modeling, Volume 58, 2016 (4), 671-701

Using response time data to inform the coding of omitted responses

Jonathan P. Weeks[1], *Matthias von Davier & Kentaro Yamamoto*

Abstract

Examinees may omit responses on a test for a variety of reasons, such as low ability, low motivation, lack of attention, or running out of time. Some decision must be made about how to treat these missing responses for the purpose of scoring and/or scaling the test, particularly if there is an indication that missingness is not skill related. The most common approaches are to treat the responses as either not reached/administered or incorrect. Depending on the total number of missing values, coding all omitted responses as incorrect is likely to introduce negative bias into estimates of item difficulty and examinee ability. On the other hand, if omitted responses are coded as not reached and excluded from the likelihood function, the precision of estimates of item and person parameters will be reduced. This study examines the use of response time information collected in many computer-based assessments to inform the coding of omitted responses. Empirical data from the Programme for the International Assessment of Adult Competencies (PIAAC) literacy and numeracy cognitive tests are used to identify item-specific timing thresholds via several logistic regression models that predict the propensity of responding rather than produce a missing data point. These thresholds can be used to inform the decision about whether an omitted response should be treated as not administered or as incorrect. The results suggest that for many items the timing thresholds (20 to 30 seconds on average) at a high expected probability level of observing a response are notably higher than thresholds used in the evaluation of rapid guessing of responses (e.g., 5 seconds).

Key words: Response time data, omitted reponses, timing tresholds

[1] *Correspondence concerning this article should be addressed to:* Jonathan P. Weeks, PhD, Educational Testing Service, 225 Phillips Boulevard, 08628, Princeton, New Jersey, United States; email: jweeks@ets.org

In many assessments there is a high likelihood that some examinees will omit at least one answer for one reason or another. This type of nonresponse may or may not be ability related. While low ability with respect to the measured construct may play a role, other reasons, such as low motivation, lack of attention, or running out of time may be likely possibilities. If the missing data are ignorable (i.e., missing at random or missing completely at random), estimates of item parameters and examinee ability in a latent variable model will be unbiased, but if they are not ignorable, the treatment of these values can introduce systematic error into parameter estimates (Rubin, 1976). When analyzing responses from test administrations in which nonresponse data are more than rarely occurring exceptions, some principled way of treating these data is required. This is true in operational analyses using either classical test theory (which typically requires complete data without missingness) or modern test theory such as item response theory (IRT; Lord & Novick, 1968) which, in principle, can handle data that are missing completely at random or missing at random. For this paper we primarily address nonresponse data treatments in the context of IRT or related methods. The goal of this study is to examine whether the coding of omitted responses based on response time information from a computer-based assessment in a low-stakes context can improve results compared to ad hoc methods (e.g., treating omitted responses as incorrect by default) typically applied in estimates of item/ability parameters. This goal is accomplished using empirical data from the Programme for the International Assessment of Adult Competencies (PIAAC) literacy and numeracy cognitive tests.

Background

Terminology

Before proceeding, it is important to clearly define the different types of nonresponse seen in large scale assessment data. We use the term "nonresponse" to refer to any value in a dataset of item responses that, after scoring, does not correspond to a correct or incorrect response code (or by extension for polytomous items, responses that do not correspond to a score category that influences an examinee's estimate of ability). In more simple terms, if an individual does not provide an answer to a given item, it is considered a nonresponse. If an examinee has no opportunity to respond to the item, either by design or because the individual did not see the item, we refer to these as not administered[2] and not reached items respectively as "missing" responses. On the other hand, we use the term "omit" to refer to nonresponse values in cases where the examinee saw the item (or is believed to have seen the item) but no response was given. The reason for this distinction is that missing and omitted responses are treated differently for the purpose of item response modeling and/or scoring. Not reached items, not administered items, and omitted item responses all warrant a different treatment: An individual who never saw an

[2] In the case of responses that are missing by design, missing values can be treated as ignorable (Rubin, 1976).

item by design cannot be expected to respond, obviously. Similarly, an examinee who did not reach the last 2-3 items because of time constraints also had no chance to produce a response and may or may not have gotten the items correct. On the other hand, an individual who saw an item and decided not to provide a response may have done so due to an understanding that the item is too difficult, or due to other reasons such as a lack of motivation, or an intent to come back to this item later that was never acted upon.

Treatment of nonresponse data

Typically, nonresponse data are treated in one of two ways for the purpose of item analysis and scoring: 1) the values are coded as not administered and excluded from the estimation of item and/or ability parameters or 2) the values are coded as omits and scored as incorrect or partially correct. The former approach is generally applied for missing responses that appear sequentially, usually at the end of a test or test section. If the examinee did not see these items, the missingness is usually assumed to be not indicative of respondent proficiency (in explicitly speeded tests, however, this may not be the case). Conversely, missing values in the middle of a string of observed responses are commonly coded as omits and scored as incorrect. When an examinee provides a response to previous and subsequent items, it is assumed that the individual had time to consider the (omitted) item(s) but chose not to answer them.

Much of the uncertainty in the coding of nonresponse data is tied to paper-and-pencil tests because there is rarely any additional information to determine whether an item without a response was omitted due to skill-related reasons. Imagine, for example, one test taker who goes systematically through all items and completes them in sequential order versus another test taker who browses to the end of the test booklet and then works on items for which a correct response is easily generated and skipping items that seem to require a lot of work. Since neither of these test-taking approaches is directly observable, applying a single treatment rule to the nonresponse data is likely to introduce some bias into estimates of the item parameters and/or ability estimates. Computer-based assessments, on the other hand, can often provide information such as timing and keystroke or process data in conjunction with the item response data. Hence, the coding of nonresponses in computer-based testing does not need to rely as heavily on assumptions about the missing data mechanism. In particular, timing data can provide immediate insight into whether an item was presented; that is, items observed by an examinee should have an associated response time; whereas, the response time should be missing or equal to zero for items that were not administered. Given the availability of this additional information, decisions about the treatment of "not administered" responses should be readily resolved (e.g., excluding them from the estimation); however, decisions regarding the treatment of omitted responses are not so clear.

The question of interest for this study is whether one should change how omitted responses are treated in the estimation of item and ability parameters, as a function of response time. By extension, this raises the question as to which responses should be recoded and what values should be assigned? Consider three examinees with no responses to a given item: One individual has no associated response time (or 0-second exposure) for the item, another moves to the next item after only 4 seconds, and the other advances after 45 seconds without responding. It is obvious that the first examinee did not see the item; thus, the response should be coded as "not administered" irrespective of where the item appears in the test. While this response is likely to occur for an item near the end of the test, it is possible that in the middle of the test the item was not presented, either because of routing rules, a software glitch, some intervention of the test proctor, or other circumstances. In this case the examinee should not be penalized because the item was not seen. For the other two examinees, it is useful to consider whether the time spent on the item was sufficient to read the prompt and answer the question. More information is needed about the item (e.g., difficulty, complexity, average response time) and possibly the examinees (e.g., ability level, motivation) in order to sufficiently address this question. However, if both responses are coded as omits, we implicitly assume the same missing data mechanism applies to both examinees (and all additional examinees with omitted responses). If this assumption does not hold, a strong potential exists for bias to be introduced into estimates of the item and ability parameters (Glas & Pimentel, 2008; Moustaki & Knott, 2000; O'Muircheartaigh & Moustaki, 1999; Rose et al., 2010).

As an alternative to coding all of these nonresponse values as omits (and eventually as incorrect), it may be possible to identify a single threshold, or a set of item-specific thresholds, for the minimum amount of time needed to provide a valid (not random) response. For examinees with response times below this threshold, one might argue that the missing response is more consistent with the not reached/not administered characterization (Wise & Kong, 2005). At what point then can one reasonably argue that a sufficient amount of time has been spent on an item for an omitted response to be recoded as incorrect? This question has been considered primarily in the context of rapid-guessing behavior for unmotivated examinees (Kong, Wise, & Bhola, 2007; Schnipke & Scrams, 1997; Wise & Kong, 2005) and to some extent in the context of speeded tests (cf. Lee & Chen, 2011) as well as through the use of item response models that incorporate timing information (cf. Maris & van der Maas, 2012; van der Linden, 2007). To be clear, our interest in this study is in the coding of the responses prior to conducting any IRT modeling; therefore, we do not examine approaches that incorporate response times into the latent variable model. In any case, one should expect the propensity of omitted responses to be only weakly correlated with the construct of interest; hence, any bias in estimates of the item/ability parameters associated with excluding these responses from the likelihood function should be small. On the other hand, for omitted responses with associated response times that are longer (i.e., above some threshold), there is likely a stronger justification for treating an omitted response as incorrect, although disentangling the causal mechanism for these responses is more challenging.

Impact of omit coding in large-scale assessments

There are a variety of reasons why examinees may omit responses on a test. Among these are low skill levels, low motivation, lack of attention, or running out of time. The tendency to produce omitted responses may also be associated with item type, response format, and whether the test has high or low stakes for participants. In a high-stakes context, the implications of omitting responses may be more consequential (e.g., receiving a lower score on a college entrance exam) and the presence of nonresponses are more likely to be associated with speededness (e.g., Yamamoto & Everson, 1997). For low-stakes tests (e.g., large-scale surveys where no individual scores are returned/reported), the bigger issue is likely low motivation (DeMars, 2000; Wise, Bhola, & Yang, 2006; Wise & DeMars, 2005), although speededness may also be a problem (Wise & Kong, 2005). In both cases (high and low stakes) it can be shown that responding randomly in speeded situations has a detrimental effect on estimates of item and person parameters (cf. Bolt, Cohen, & Wollack, 2002; Wise & DeMars, 2006; Yamamoto & Everson, 1997). One should expect a similar detrimental effect on estimates of item and person parameters if omitted responses are coded as incorrect.

Depending on the number of omitted responses, coding these responses as incorrect is likely to introduce negative bias into estimates of item difficulty (e.g. Rose, von Davier & Xu, 2010); the magnitude of this bias will increase as the number of omitted responses coded as incorrect increases. A concomitant decrease in estimates of examinee ability and group means is also likely in this case. On the other hand, if omitted responses are coded as not reached and excluded from the estimation, estimates of item and person parameters should have less potential bias, although this assumes that there are no systematic reasons for the omitted responses. Note that the associated standard errors will be larger relative to estimates with omitted responses coded as incorrect, simply because the number of responses included in the determination of the estimate is smaller when ignoring nonresponses.

In the case of large-scale assessments, particularly when the emphasis on reporting results at a group level (e.g., at the country or state level) rather than at the examinee level, the assessments are likely to be perceived as inconsequential and hence low-stakes by many test takers. At the same time, the results may have higher associated stakes for, say, country or state leaders due to perceptions about the group ranking. If the proportion of omitted responses differs substantively across groups and these responses are coded as incorrect, estimates of group performance will likely be worse than expected. The reverse may be less likely but is still possible: if respondents in some country were instructed to finish the test no matter what and always give a response, this may lead to many random answers and hence a higher proportion of incorrect responses compared other countries where respondents are encouraged to omit responses if unsure. The challenge is determining how best to minimize potential bias. Decisions about how to code omitted responses are one way to address this issue.

Nonresponse coding in large-scale assessments

For this study, we use data from the Programme for the International Assessment of Adult Competencies (PIAAC). The survey includes cognitive items in the domains of literacy, numeracy, and problem solving; it is also intended to gather data about how adults use these skills at home and at work. As currently implemented, cognitive responses with no associated response time or response times less than five seconds are coded as not reached; whereas, nonresponse data with associated response times greater than or equal to five seconds are coded as omits in the database and treated as incorrect for the purpose of item parameter estimation and scoring. The data for other large-scale assessments tend to be coded in different manners, the National Assessment of Educational Progress (NAEP), for instance, uses partially correct coding (proportional to the guessing probability) for omitted responses, and the Programme for International Student Assessment (PISA) applies yet another set of scoring rules.

PISA is an interesting case with respect to the coding of omitted responses. PISA is a cognitive test of (primarily) reading, math, and science that allows for comparisons of country performance in each of these domains. In PISA administrations from 2000 to 2012, all not reached responses were treated as missing for the purpose of item parameter estimation; however, when estimating plausible values to generate country means and other group statistics, all not reached responses were treated as incorrect. The primary impact is that the estimation conventions used until 2012 results in a discrepancy between expected response probabilities and estimates of country performance. Still, the important take-away is that the treatment of omitted responses does not always have a clear solution. Our goal is to examine whether response time data can be used to provide a more evidence-based rationale for treating omitted responses as either incorrect versus missing.

Research questions

1. Is it possible to identify a threshold for computer-administered items that provides a defensible demarcation between the response-time distributions of omitted and observed item responses?
2. Is there a single response time threshold that can be applied to all items, or, are there item-specific thresholds? Further, do item thresholds apply across countries or are response time thresholds country-specific?
3. Is a response time threshold of five seconds (typically applied when recoding rapid responses) defensible given the empirical results?

Methods and results

The analyses for this study were conducted in two phases. The first phase focused on a descriptive examination of the distributions of response times (total time on the item) for correct, incorrect, and omitted[3] responses across PIAAC literacy and numeracy items. The second phase employs a model-based approach to identify item-specific thresholds for response times that could be used to provide an alternative coding scheme for omitted responses. Note that the term "response time" is used regardless of whether a valid response was given or an omission was recorded. Also note that an omission requires a response in the sense of hitting the “next” button in order to advance to the next item.

[3] Responses with missing associated response times were not included in the analyses.

Data

The data for this study include item responses and response time information for the sample of examinees from the 2012 PIAAC literacy and numeracy cognitive tests. There are 49 literacy items and 49 numeracy items. These items are all computer based and do not incorporate data from examinees who took a partially parallel set of items in a paper-based format (because the paper-based items do not have associated timing data). The data include 260,179 examinees from 22 countries. On average there are 11,826 examinees per country, with a minimum and maximum of 5,095 and 33,987 examinees, respectively. Response times greater than three standard deviations above the mean time for each item were treated as outliers and removed from the data prior to conducting any of the analyses.

Phase 1: Methods

The first phase of this study is intended to be descriptive in nature. We began by examining the frequency of missing responses and the relationship between nonresponse data, response times, and examinee ability. We then examined the response time distributions for each item for each response category (omit, incorrect, correct), across countries, in order to identify patterns. In particular we considered rapid response times, the separation in times between response categories, and the relationship between response times and item difficulty.

Phase 1: Results

As a first step, we examined the proportion of missing responses and the correlation at the examinee level between the number of omitted responses, total test time, and an estimate of examinee ability[4] in both literacy and numeracy. The percentage of omitted responses for a given literacy item ranged from 2.6% to 18.3% with a mean of 8.5%. The percentage of omitted responses for a given numeracy item ranged from 1.5% to 15.4% with a mean of 6.2%. In literacy, 64% of examinees did not have any omitted responses, 28% had fewer than six omitted responses, and 8% had six or more omitted responses with a maximum of 19 omitted responses. In numeracy, 70% of examinees did not have any omitted responses, 27% had fewer than six omitted responses, and 3% had six or more omitted responses with a maximum of 19 omitted responses.

Table 1 presents a summary of the country-level correlations between the number of omitted responses, total test time, and examinee ability. For both literacy and numeracy, the number of omitted responses is weakly to moderately negatively correlated with both

[4] Plausible values for examinee ability (von Davier, Gonzalez, & Mislevy, 2009) were available with the data for literacy and numeracy respectively. The first plausible value was used. Note: All missing responses in an examinee's response string were excluded from the likelihood function for the purpose of estimating both item parameters and plausible values of ability.

total time and ability across countries. That is, examinees taking more time overall and higher ability examinees tend to have fewer omitted responses. Conversely, the correlations between total time and examinee ability are low to moderate and positive; higher ability examinees tend to take more time overall. This is not an unexpected result because the items of PIAAC are challenging, whereas for easy items, response time is typically negatively related to proficiency (i.e., high-ability respondents are faster than low-ability respondents on easy tasks, while the reverse is true for challenging tasks). If omission propensity were perfectly (negatively) correlated with ability, it would be appropriate to recode all omitted responses as incorrect; however, because this is not the case, recoding omitted responses as incorrect has the potential to introduce negative bias. As such, any bias introduced by coding omitted responses as incorrect is likely to have a greater impact on low-ability examinees (pushing their estimates of ability even lower) relative to higher ability examinees. The magnitude and direction of these correlations is quite consistent across countries.

As a next step in our analysis we produced graphical displays of the distributions of response times for omitted, incorrect, and correct responses for each country (for each item individually). Figure 1 illustrates smoothed distributions for a single literacy item to illustrate the separation between the response time distributions for omitted versus observed responses. The horizontal axis corresponds to the item response time in seconds, and each line, for each response category, corresponds to a different country. The majority of the omitted responses, across countries, have associated response times that are very short; whereas the time associated with observed responses tends to be longer and more variable.

Table 1:
Summary of Country-Level Correlations Between Number Omitted, Total Time, and Examinee Ability

Unit	Statistic	Cor (number omit, total time)	Cor (number omit, ability)	Cor (total time, ability)
Literacy	Min	-.56	-.55	.05
	Max	-.20	.08	.37
	Mean	-.45	-.37	.20
	SD	.08	.13	.10
Numeracy	Min	-.39	-.48	.00
	Max	-.14	-.06	.30
	Mean	-.31	-.32	.13
	SD	.06	.12	.08

Visually, considerable overlap is seen in the response times associated with incorrect and correct responses, while the distributions of response times associated with omitted responses are noticeably different from those for the observed responses. The differences between these distributions shown in Figure 1 hold with some variation across all literacy and numeracy items. That is, omitted responses are consistently associated with faster average times than observed responses; however, the separation between response time patterns for incorrect and correct responses and the corresponding notion of a threshold for omitted responses is a bit more variable. For some items there is a clear delineation among all three categories based on response time, whereas for other items there is no apparent demarcation. Figures 2 and 3 show the distributions of log-transformed response times – aggregated across countries – for all literacy and numeracy items respectively. The items are ordered, based on P+ (proportion correct) values, from least difficult (top) to most difficult (bottom) and reference lines (on the log-transformed scale) for 5 and 20 seconds are included. The items are sorted to provide some indication of the relationship between item difficulty and response times. More formally, the correlation between P+ values and mean response times is *r* = -.56 for literacy and *r* = -.69 for numeracy. In other words, more difficult items tend to be associated with longer response times. Tables 2 and 3 present the median and 90th percentile (P90) time for each item for each response category in literacy and numeracy as well as the corresponding proportion of omits and correct responses.

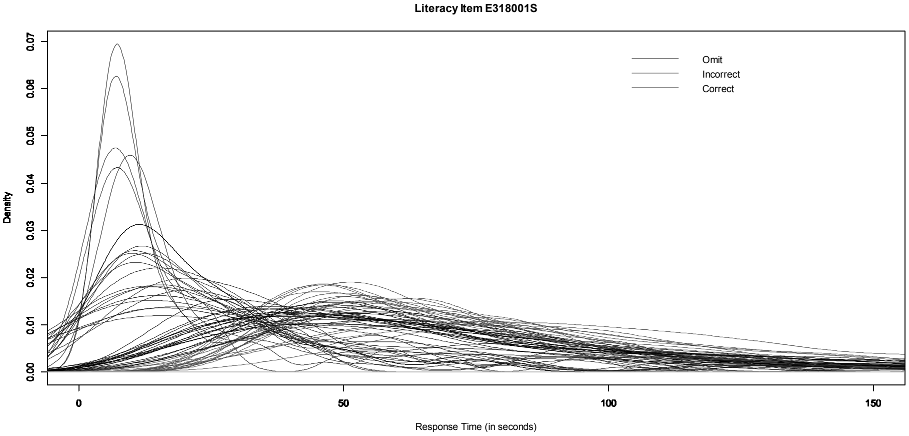


Figure 1:
Item response time distributions for a single PIAAC literacy item, by country, by response category

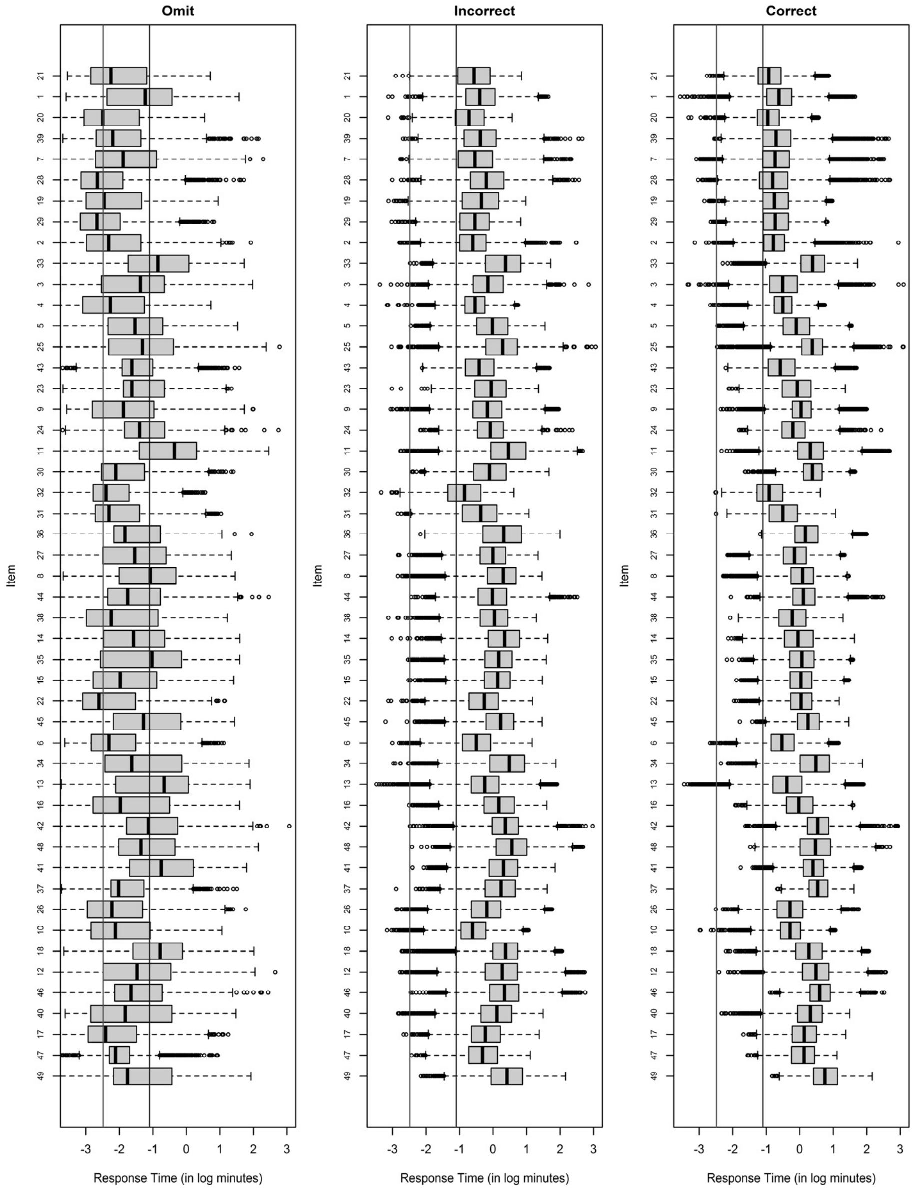


Figure 2:
Item response time distributions, by item, across countries – literacy items

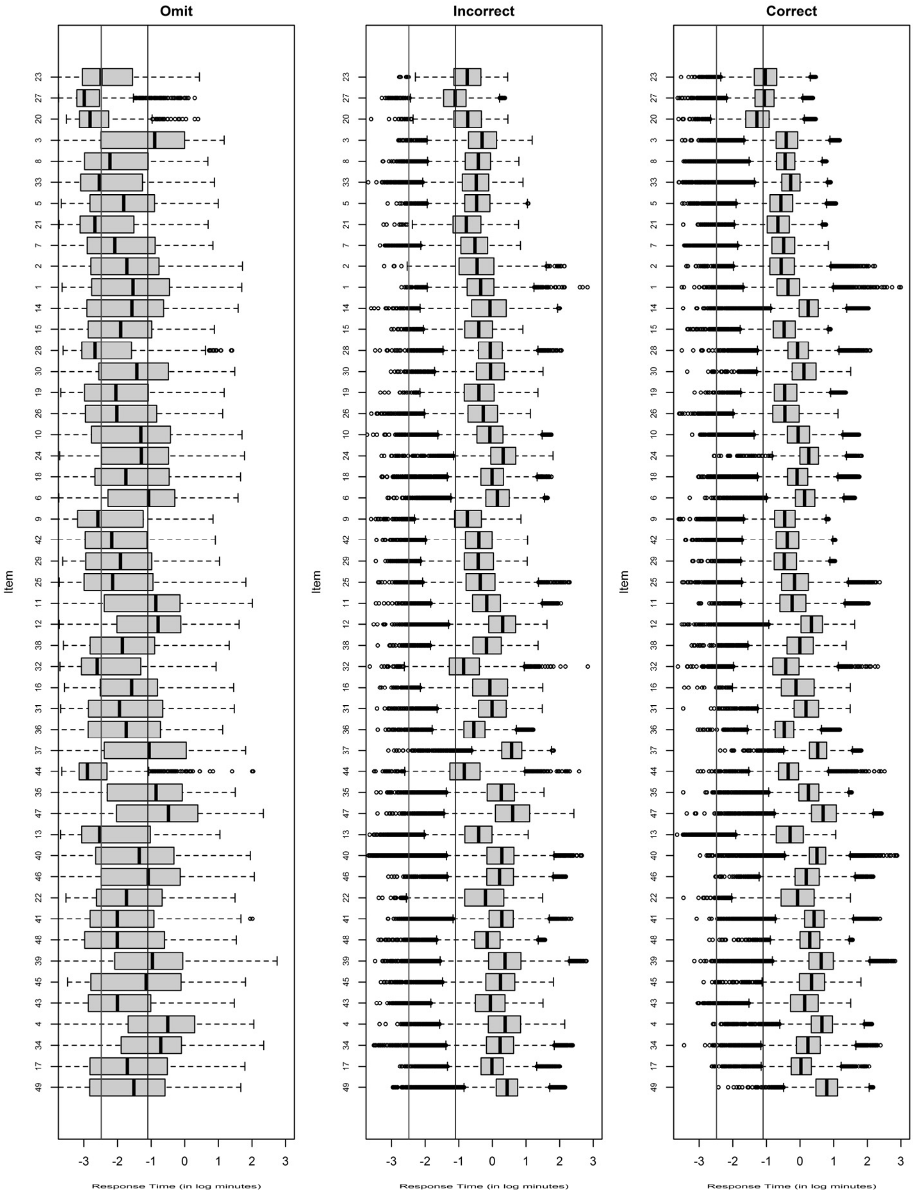


Figure 3:
Item response time distributions, by item, across countries – numeracy items

Table 2:
Response Time Summary (In Seconds) by Item by Response Category – Literacy

Item	Literacy							
	Proportion Omit	Proportion Correct	Omit		Incorrect		Correct	
			Median	P90	Median	P90	Median	P90
21	0.05	0.96	6	39	34	79	24	50
1	0.03	0.91	18	66	40	100	32	68
20	0.10	0.91	5	36	29	68	23	46
39	0.04	0.89	7	37	41	103	30	70
7	0.03	0.86	9	55	35	103	29	70
28	0.08	0.83	4	26	49	132	27	66
19	0.12	0.83	5	54	42	105	28	65
29	0.08	0.82	4	21	35	78	29	62
2	0.06	0.82	6	37	33	72	27	53
33	0.15	0.81	26	122	87	194	88	174
3	0.03	0.80	15	51	52	123	36	87
4	0.04	0.78	6	35	35	64	36	61
5	0.05	0.76	13	56	59	140	54	114
25	0.12	0.74	16	82	80	173	87	157
43	0.09	0.71	12	44	40	93	34	77
23	0.09	0.71	12	69	57	127	56	116
9	0.08	0.68	9	53	51	118	62	117
24	0.15	0.66	15	57	55	116	49	99
11	0.11	0.65	42	132	95	243	82	178
30	0.06	0.65	7	36	54	132	88	154
32	0.08	0.64	5	23	26	62	24	54
31	0.09	0.63	6	39	42	97	36	83
36	0.07	0.60	10	50	83	213	71	150
27	0.08	0.60	13	63	60	122	51	101
8	0.13	0.58	20	83	81	161	65	122
44	0.04	0.58	10	63	59	136	67	133
38	0.07	0.58	6	53	63	127	48	106
14	0.06	0.58	12	62	85	190	57	137
35	0.09	0.57	22	86	71	150	64	129
15	0.09	0.57	8	55	69	138	62	118
22	0.06	0.52	4	33	46	101	62	114
45	0.07	0.51	17	95	76	154	77	145
6	0.06	0.50	6	30	36	83	35	72

34	0.18	0.50	12	119	98	216	97	201
13	0.03	0.49	31	109	47	110	41	98
16	0.10	0.48	8	77	71	170	58	133
42	0.10	0.48	19	83	86	181	103	191
48	0.08	0.39	15	91	105	237	96	219
41	0.09	0.39	28	138	82	186	89	167
37	0.12	0.35	8	34	76	163	103	183
26	0.13	0.33	7	36	50	107	45	91
10	0.06	0.32	7	42	33	70	45	84
18	0.16	0.30	28	95	87	176	79	164
12	0.11	0.28	14	74	79	178	98	197
46	0.09	0.27	12	59	85	186	109	206
40	0.08	0.25	10	79	67	147	83	158
17	0.09	0.22	5	35	48	113	69	136
47	0.10	0.20	7	24	44	98	69	121
49	0.11	0.16	10	89	91	212	127	260
Min	0.03	0.16	4	21	26	62	23	46
Max	0.18	0.96	42	138	105	243	127	260
Mean	0.08	0.58	12	62	60	136	60	122
SD	0.04	0.21	8	29	21	47	27	51

Note: P90 denotes the 90th percentile

Table 3:
Response Time Summary (In Seconds) by Item by Response Category – Numeracy

Item			Numeracy					
	Proportion Omit	Proportion Correct	Omit		Incorrect		Correct	
			Median	P90	Median	P90	Median	P90
23	0.07	0.96	5	27	28	59	21	42
27	0.02	0.92	3	13	20	40	21	37
20	0.03	0.91	4	14	29	61	17	35
3	0.04	0.89	25	98	44	108	40	83
8	0.05	0.87	6	39	39	79	39	68
33	0.03	0.85	5	44	37	74	45	78
5	0.02	0.85	10	47	38	82	34	67
21	0.04	0.83	4	33	28	64	31	61
7	0.05	0.83	8	52	36	72	37	70
2	0.03	0.83	11	55	38	101	34	74
1	0.01	0.82	13	70	43	92	42	83
14	0.03	0.79	12	65	56	136	76	140
15	0.05	0.79	9	47	40	86	37	72
28	0.04	0.78	4	35	56	117	56	106
30	0.08	0.77	14	69	57	127	67	139
19	0.07	0.77	8	46	40	96	38	77
26	0.05	0.76	8	58	46	104	38	87
10	0.03	0.76	16	80	56	120	57	110
24	0.11	0.75	16	65	83	169	78	140
18	0.05	0.74	10	77	59	116	55	104
6	0.05	0.74	20	76	70	142	69	126
9	0.02	0.73	5	40	29	65	38	70
42	0.03	0.72	7	40	40	84	41	80
29	0.07	0.71	9	44	39	94	38	78
25	0.04	0.71	7	52	42	98	51	113
11	0.06	0.69	25	84	51	114	48	106
12	0.06	0.68	27	92	82	171	84	160
38	0.10	0.67	9	49	50	114	60	128
32	0.02	0.65	4	44	26	63	39	86
16	0.11	0.63	12	49	56	140	53	137
31	0.12	0.61	9	63	60	132	72	146
36	0.05	0.61	11	46	35	66	38	66
37	0.10	0.57	21	102	106	190	102	172

44	0.03	0.57	3	19	26	63	42	79
35	0.06	0.56	26	99	78	163	77	141
47	0.11	0.56	37	170	110	269	119	259
13	0.02	0.53	5	53	40	85	45	92
40	0.08	0.52	16	85	79	157	100	166
46	0.08	0.49	20	98	74	166	72	156
22	0.14	0.49	11	76	49	135	56	138
41	0.12	0.49	8	51	80	152	91	162
48	0.08	0.46	8	66	51	109	80	145
39	0.12	0.46	23	124	87	215	113	229
45	0.09	0.46	19	98	76	173	85	180
43	0.06	0.45	8	45	57	128	69	147
4	0.11	0.42	36	144	88	205	115	212
34	0.15	0.39	29	93	76	163	76	156
17	0.03	0.37	11	68	59	114	62	116
49	0.06	0.21	13	61	93	172	133	241
Min	0.01	0.21	3	13	20	40	17	35
Max	0.15	0.96	37	170	110	269	133	259
Mean	0.06	0.67	13	65	55	119	60	118
SD	0.04	0.17	9	31	22	47	27	52

Note: P90 denotes the 90th percentile

For both literacy and numeracy, the majority of observed responses have associated response times greater than 20 seconds; however, for omitted responses, there are a number of items where the bulk of the response time distribution falls below 20 seconds. It is also important to note that there does not appear to be a single, consistent response time demarcation between omitted and observed responses. This finding suggests that when making decisions about how to code omitted responses, item-specific thresholds may be defensible. Some obvious differences are present between the response times for observed and omitted responses, within and between items, but it is also clear that some observed responses are fast responses. For example, for the numeracy items, an appreciable number of examinees provided a response in less than 5 seconds.

Phase 2: Methods

A key goal of Phase 1 was to provide some justification for the specification of the models used in Phase 2 to identify response time thresholds. With respect to decisions regarding the model specification, we are interested in (a) whether a single threshold can be identified and applied to all items or if item-specific thresholds are justified, and (b) whether the identified threshold(s) can be applied across all countries or if country-specific thresholds are justified. To address these considerations, we started with a variance components analysis to identify the proportion of variability in response times explained by respondent time spent within and between items, and items crossed with countries. For this examination, the outcome variable y_{ijc} is the response time (log-transformed) for item *i*, nested within examinee *j*, and administered in country, *c*. The same items are administered in all countries. As such, the items and countries are fully crossed. Examinees, on the other hand, do not necessarily receive all of the same items (due to a planned incomplete design, not reached items).To characterize the variance components associated with item response times, we fit the following mixed-effects model:

$$\mathrm{y}_{ijc} = \sum_{i} (\beta_{0i} + \xi_{ij})\mathrm{Item}_i + \sum_{i}\sum_{c} \beta_{1ic}\mathrm{Item}_i \times \mathrm{Country}_c + \varepsilon_{ijc} \,. \tag{1}$$

In this equation, β_{0i} is the fixed effect for each item, which serves as baseline item mean time. To account for examinee-level variability, ξ_{ij} is included as a random effect by item. Because countries are fully crossed with items, the fixed effect β_{1ic} is included as an item-by-country interaction. The residual ε_{ijc} is normally distributed as $\varepsilon_{ijc} \sim \mathrm{N}\left(0, \sigma^2\right)$, and the random effect for examinees is normally distributed as $\xi_{ij} \sim \mathrm{N}\left(0, \psi^2\right)$. Items and countries were dummy coded with the first item/country treated as the reference group.

In the second part of the Phase 2 analysis we employed a model-based approach to identify item-specific thresholds for response times that could be used to provide an alternative coding scheme for omitted responses. We denoted the response to a multiple-choice item *i* by examinee *j* as $x_{ij} =$ 1 for a correct response, $x_{ij} = 0$ for an incorrect response, and $x_{ij} = 9$ for an omitted response. A multinomial logistic model could be used to identify the response time thresholds for the various response categories; however, because omitted responses are typically coded as incorrect (versus a combination of correct and incorrect), we chose to focus on the response time distinction between omitted and incorrect responses only. This decision was also supported by findings from Phase 1, which pointed to a substantial overlap in the response time distributions for correct and incorrect responses. As such, we defined the outcome variable y_{ij} as

$$y_{ij} = \begin{cases} 1 & \text{if } x_{ij} = 0 \\ 0 & \text{if } x_{ij} = 9 \end{cases}.$$

To identify response time thresholds, we considered three binary logistic regression models. It is important to note that the model parameters were estimated separately for each item in each content domain rather than dummy coding the items and estimating all of the coefficients simultaneously. The three models were specified as follows:

Model 1:

$$\Pr(Y_{ij} = 1 \mid \log[\text{Time}_{ij}]) = \frac{\exp\left(\beta_{0i} + \beta_{1i}\log\left[\text{Time}_{ij}\right]\right)}{1+\exp\left(\beta_{0i} + \beta_{1i}\log\left[\text{Time}_{ij}\right]\right)} \tag{2}$$

Model 2:

$$\begin{aligned} &\Pr\left(Y_{ij} = 1 \| \log[\text{Time}_{ij}\right], \log\left[\text{TotalTime}_{j}\right]) \\ &= \frac{\exp\left(\beta_{0i} + \beta_{1i}\log\left[\text{Time}_{ij}\right] + \beta_{2i}\log\left[\text{TotalTime}_{j}\right]\right)}{1+\exp\left(\beta_{0i} + \beta_{1i}\log\left[\text{Time}_{ij}\right] + \beta_{2i}\log\left[\text{TotalTime}_{j}\right]\right)} \end{aligned} \tag{3}$$

Model 3:

$$\begin{aligned} &\Pr\left(Y_{ij} = 1 \| \log[\text{Time}_{ij}\right], \log\left[\text{TotalTime}_{j}\right], \text{Ability}_{j}) \\ &= \frac{\exp\left(\beta_{0i} + \beta_{1i}\log\left[\text{Time}_{ij}\right] + \beta_{2i}\log\left[\text{TotalTime}_{j}\right] + \beta_{3i}\text{Ability}_{j}\right)}{1+\exp\left(\beta_{0i} + \beta_{1i}\log\left[\text{Time}_{ij}\right] + \beta_{2i}\log\left[\text{TotalTime}_{j}\right] + \beta_{3i}\text{Ability}_{j}\right)} \end{aligned} \tag{4}$$

The outcome of interest in all three models is the expected probability of an incorrect response conditional on some combination of item response time, total examinee time, and examinee ability in a given domain. For Model 1, the only predictor is the log-transformed response time for the given item in a given content domain. For Model 2, the log-transformed total test time for the examinee in the given domain is added as a predictor, and for Model 3 the ability estimate for the examinee in the given domain is added as well. No fixed effects for countries were specified in these models. This decision is premised on the finding from the variance decomposition (see below) that little variability in the response times is explained by between-country differences.

In the specification of Model 3 it is useful to consider the specific estimate included to represent examinee ability. If IRT-based point estimates of ability are used, particularly for examinees with omitted responses, the uncertainty around the ability estimate as well as potential bias in these estimates may be inadequately accounted for in the identification of a response time threshold. In an attempt to alleviate this issue, we opted to use plausible values (e.g., Mislevy, 1991; von Davier, Gonzalez, & Mislevy, 2009) obtained from the population model for PIAAC (for details on the model used, see, for example, von Davier & Sinharay, 2013). Stated differently, plausible values are imputations based on item responses to all domains and include background data as covariates, and hence are likely a more adequate basis for estimating how examinee ability and response times are related to omitted versus incorrect responses than, say, an MLE or EAP point estimate of ability.

Phase 2: Results

Variance decomposition

As a first step in the model-based approach to identifying timing thresholds, we examined the proportion of variability in response times explained by items, examinees, and countries for literacy and numeracy respectively. Table 4 presents the variance components and the corresponding proportion of the variance explained by each of the factors. In both literacy and numeracy, the largest proportion of the variability in response times is accounted for by the examinee. This provides some justification for incorporating timing information into the item response model; however, given that the choice of item response model has already been established for PIAAC, these types of models are not examined further.

The second largest proportion of variability in response times occurs within items. This is followed by item-by-country interactions. The variability explained by the items suggests that item-specific response time thresholds should be considered. This finding is consistent with our conclusion based on Phase 1 analyses above. On the other hand, given the small proportion of variability explained by item-by-country interactions (which is desirable from a policy perspective), country interactions do not appear to be necessary in the model-based identification of thresholds. Stated differently, there is not much support for identifying item-by-country specific thresholds.

Table 4:
Response Time (In Seconds) Variance Components

Factor	Literacy		Numeracy	
	Variance	% Explained	Variance	% Explained
Item	0.077	16.8	0.106	24.1
Item x country	0.021	2.9	0.016	2.4
Examinee	0.243	34.6	0.213	32.7
Residual	0.382	45.8	0.333	40.8

Model-based thresholds

The findings from Phase 1 and the variance decomposition suggest that there is some distinction between the time required to produce an observed response versus that for omitting the response. We used three models to identify these thresholds. Based on a likelihood ratio test, Model 3 fit the data significantly better than Model 2 for all but four items in literacy and five items in numeracy (with a Bonferroni correction applied). This is not surprising given the large proportion of variability explained by examinees in the variance decomposition; however, estimates of examinee ability are not generally known until after the scaling and population modeling have been completed for a test. For this reason, the results from this model are used as a reference for evaluating the thresholds identified by the other models. Model 2 generally fit better than Model 1, although there were no significant differences between the models for 16 literacy items and 15 numeracy items (with a Bonferroni correction applied). This suggests that the inclusion of total time may add some value over simply the item time in the identification of response time thresholds.

The regression coefficients in these models characterize the change in the log-odds of an incorrect response, yet these coefficients need to be transformed into a meaningful threshold. We opted to identify the item-specific time where the expected probability of an incorrect response is .5, .7, .8, and .9. For Model 1, we simply solved for the threshold at each of the expected probability levels. For Models 2 and 3, which include continuous predictors, we identified three levels of the predictors. For Model 2, thresholds are reported for total times at the first quintile, median, and third quintile (low, moderate, and high total time, respectively). For Model 3, thresholds are reported for the median total time and plausible values at the first quintile, median, and third quintile (low, moderate, and high ability, respectively). Tables 5 and 6 provide summaries of the item-specific thresholds for literacy and numeracy respectively. The values in the undefined column are the number of items (out of 49) that had response times of less than 0.01 seconds.

Figure 4 illustrates the fit of Model 1 relative to the empirical data for a typical item and for the worst fitting item in literacy and numeracy (plots for all of items are included in the appendix). In Figure 4, the top panels present the results for the literacy items while the bottom panels present results for the numeracy items. The left panels correspond to the typical item and the right panels correspond to the worst fitting items. The horizontal axis is the response time (in seconds). Note that the indicators are not spaced at equal intervals. The interval was set based on log-transformed times. The set of grey lines in the background are the proportion of incorrect responses (out of the total incorrect and omitted responses per item) for each country where the data are binned by response time quintiles across countries – the variability in country data for the worst fitting items is due to small numbers of examinees in the given response time ranges. The thicker lines with the indicator points correspond to the overall proportion of incorrect responses at each point across countries while the smooth ogive is the model-based expected probability. The dashed vertical line corresponds to the response time threshold at the .8 expected probability level. In both content domains the typical item is fit well by the model and the fit for the worst fitting items is acceptable.

Table 5:
Summary of Literacy Response Time Thresholds (In Seconds)

Model	Factor	Expected probability	Response time threshold (in seconds)							
			Min	Q1	Median	Q3	Max	Mean	*SD*	Undefined
1		.5	0.0	9.9	13.1	17.3	43.9	14.1	7.2	0
		.7	0.2	14.6	22.1	28.0	95.0	23.6	14.5	0
		.8	0.9	19.1	29.1	38.4	155.4	33.1	23.8	0
		.9	6.3	29.5	41.6	63.8	176.9	51.0	32.8	1
2	Low total time	.5	0.2	8.6	11.9	16.5	44.1	13.3	7.9	1
		.7	2.6	13.7	21.2	25.7	93.1	22.6	14.8	1
		.8	8.8	17.7	27.9	38.5	149.8	32.2	23.2	1
		.9	15.8	30.5	43.6	63.2	159.3	53.5	34.8	2
	Moderate total time	.5	0.1	7.8	11.8	17.4	45.2	13.2	8.6	1
		.7	0.9	13.5	18.5	26.2	95.4	22.1	16.0	1
		.8	4.4	17.2	26.2	33.9	153.6	31.3	24.8	1
		.9	13.6	28.0	39.4	59.4	168.3	49.2	32.0	2
	High total time	.5	0.0	7.0	11.3	18.2	46.2	13.1	9.3	1
		.7	0.3	12.0	17.3	26.9	97.5	22.0	17.2	1
		.8	1.8	16.9	24.0	33.9	157.0	30.9	26.5	1
		.9	11.8	25.6	36.3	55.5	176.6	47.5	34.7	2
3[a]	Low ability	.5	0.2	7.5	11.2	15.6	43.5	12.8	8.4	1
		.7	1.3	12.2	17.5	26.7	93.0	21.7	15.8	1
		.8	3.7	17.2	25.3	36.5	151.0	30.9	24.7	1
		.9	10.9	27.3	38.1	61.0	167.0	48.8	32.5	2
	Moderate ability	.5	0.1	5.5	8.9	12.9	37.1	10.0	7.2	1
		.7	0.7	9.5	13.2	22.4	79.4	17.1	13.6	1
		.8	2.2	12.7	18.6	29.3	128.9	24.3	21.3	1
		.9	7.2	19.6	29.8	48.5	153.4	37.6	28.1	2
	High ability	.5	0.0	3.9	6.4	10.7	32.1	8.2	6.4	1
		.7	0.3	6.9	10.9	19.6	68.8	13.8	12.0	1
		.8	1.0	9.3	14.2	24.1	111.7	19.7	18.8	1
		.9	4.9	14.2	24.2	40.9	142.1	30.1	25.1	2

Note. Undefined indicates threshold values less than 0.01 seconds. Q = quintile.

[a] Median total time used.

Table 6:
Summary of Numeracy Response Time Thresholds (In Seconds)

Model	Factor	Expected probability	Response time threshold (in seconds)							
			Min	Q1	Median	Q3	Max	Mean	*SD*	Undefined
1		.5	0.1	4.2	7.0	12.8	29.6	8.7	6.2	0
		.7	0.5	9.1	14.4	22.9	45.5	16.4	10.5	0
		.8	1.0	13.1	22.1	32.5	67.0	25.4	16.5	0
		.9	3.5	22.9	38.5	58.2	165.3	46.3	31.5	1
2	Low total time	.5	0.0	4.1	9.0	14.7	32.3	10.0	7.6	6
		.7	0.0	7.7	15.9	23.2	47.3	16.9	11.7	5
		.8	0.1	11.3	23.0	31.9	60.4	24.4	15.6	5
		.9	0.2	21.7	42.0	55.6	109.7	42.6	25.9	4
	Moderate total time	.5	0.2	3.9	9.8	16.3	36.9	11.2	8.8	7
		.7	0.1	7.9	17.2	27.8	54.1	18.9	13.7	6
		.8	0.0	10.7	23.8	35.5	69.1	26.1	18.5	4
		.9	0.0	21.3	42.3	62.1	134.0	45.5	30.9	4
	High total time	.5	0.2	3.6	11.1	18.3	42.0	12.4	10.2	7
		.7	0.0	7.6	18.0	31.3	61.5	20.7	16.0	6
		.8	0.0	8.3	23.5	42.5	83.2	28.2	22.1	4
		.9	0.1	20.4	44.0	68.3	161.9	50.1	36.2	5
3[a]	Low ability	.5	0.1	3.0	8.8	16.4	34.8	10.3	8.5	5
		.7	0.3	6.6	15.1	28.1	50.9	17.9	13.5	5
		.8	0.0	10.2	21.4	36.3	73.1	25.2	18.8	4
		.9	0.3	18.1	43.3	62.6	145.1	44.8	32.4	4
	Moderate ability	.5	0.0	1.8	6.9	12.5	27.6	8.3	7.2	5
		.7	0.1	4.2	12.5	20.8	40.4	14.3	11.5	5
		.8	0.2	6.9	18.4	31.7	62.5	20.6	15.9	5
		.9	0.1	13.0	35.0	49.8	124.2	35.7	27.6	4
	High ability	.5	0.0	1.2	5.6	10.2	22.3	6.8	6.2	5
		.7	0.1	3.1	10.3	16.6	34.4	11.9	10.0	5
		.8	0.1	4.9	15.9	25.9	54.3	17.0	13.9	5
		.9	0.0	8.6	28.8	42.0	107.8	29.4	24.3	4

Note. Undefined indicates threshold values less than 0.01 seconds. Q = quintile.

[a] Median total time used.

Discussion

Several conclusions can be drawn from the findings of Phase 2 analyses. First, estimated threshold times can vary considerably between items. For instance, for Model 1 at the .8 expected probability level, the values range from just 0.9 seconds to just over 2.5 minutes in literacy and from 1.0 seconds to just over 1 minute in numeracy. In the former case, and in similar cases, with the very low timing threshold, the slope in the regression is very flat; hence, it is not really possible to distinguish between omit and incorrect timing distributions. If we consider the thresholds at the 25th percentile (Q1), the thresholds for literacy and numeracy are at about 19 and 13 seconds respectively, which is notably greater than the 5-second threshold commonly used to identify rapid responses.

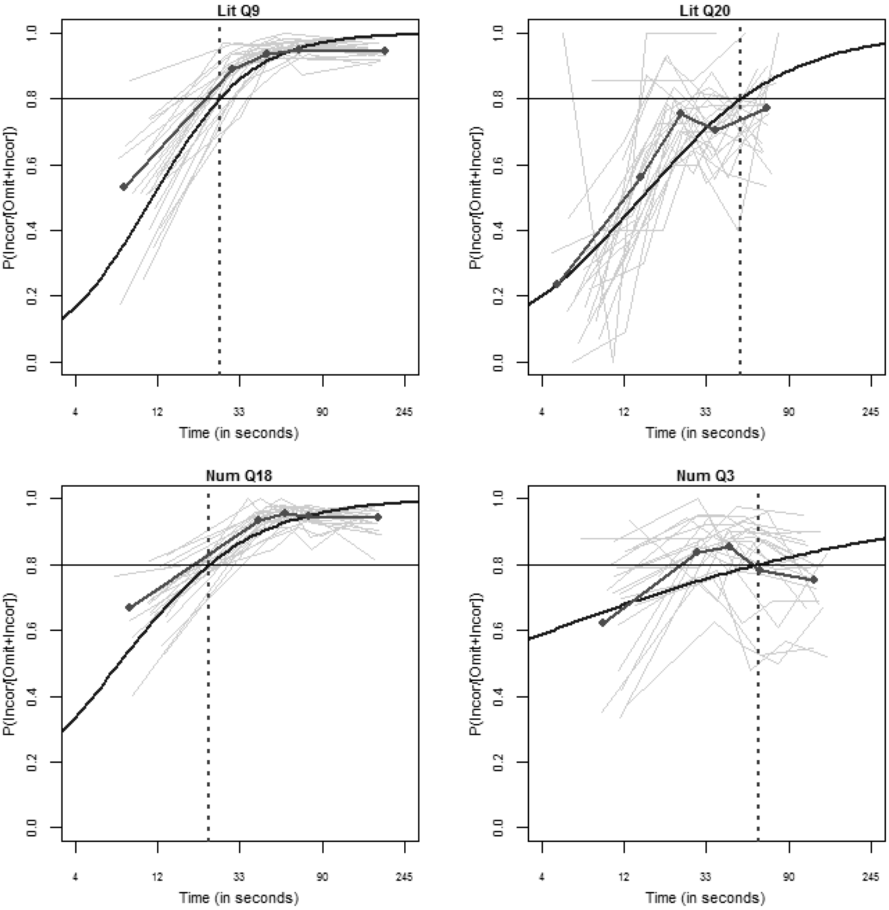


Figure 4:
Example Model 1 fit plots

When accounting for examinee ability and/or total time, the bottom end of the range changes slightly, but the range is fairly consistent. The standard deviation of threshold times (at the .8 level) in literacy is 23.8 seconds; in numeracy, the standard deviation is 16.5 seconds. These results suggest that the use of item-specific thresholds is defensible. Still, one could arguably use the item-specific threshold information to identify a single, high, time threshold (e.g., 30 to 40 seconds) that would ensure that most of the omitted responses that are recoded have a high expected probability of being incorrect (if only the predictors were observed).

Another clear (and expected) pattern in the results is the increase in threshold times associated with higher expected probabilities; this is consistent with having positive slope coefficients for all items in all of the models. In short, this confirms that the recorded times for omitted responses are in expectation shorter than the times associated with incorrect (and by extension correct) responses. Compared to Model 1, the mean threshold times are lower in literacy when examinee ability and/or total time are included in the model. In numeracy, the mean threshold times are lower when including both ability and total time; however, for Model 2, the thresholds are slightly higher for examinees with higher total times. In literacy, for Model 2, the thresholds consistently decrease as total times increase. In other words, examinees with shorter total times would need to spend a little more time on a given item before we would be willing to treat an omitted response as incorrect, at a given probability level, relative to an examinee who takes more time overall. For Model 3, the thresholds consistently decrease as examinee ability increases; the decrease in thresholds is more pronounced than in Model 2 (e.g., a decrease in means of 11.2 seconds relative to 1.3 seconds at the .8 expected probability level). In this case, higher ability examinees do not have to spend as much time on an item for us to believe that an omitted response should be coded as incorrect.

For Model 2 in numeracy, the thresholds consistently increase as total times increase. This pattern is in the opposite direction to that observed for literacy. That is, examinees who take more time overall would need to take a little more time on an item, relative to an examinee who took less time overall, for an omitted response to be treated as incorrect. This pattern may be due to the amount of attention given to numeracy as a whole. Examinees may consistently answer the items at a faster or slower pace; hence, the time spent on an item that results in an incorrect response is likely proportional to the time spent overall. On the other hand, in literacy, examinees who take more time overall may have more variability in the time spent on any given item. For Model 3, the pattern for numeracy is consistent with that for literacy; the difference in mean threshold times between the low- and high-ability groups is 8.2 seconds.

To illustrate the potential utility of a short time threshold (i.e., 5 seconds) we computed the expected probability associated with that time for each item under each model (see Tables 7 and 8). Consistent with the results presented above, using a threshold of 5 seconds does not tend to differentiate between omitted and incorrect responses with a high level of certainty; rather, the expected probabilities that an incorrect rather than omitted response was observed for many items under all three models is less than .5. The expected probability does improve slightly when incorporating examinee ability and/or total time, but the expected probabilities are still quite low. In short, using a 5-second rule is arguably too strict for a scoring rule (assigning incorrect responses too liberally) with respect to recoding omitted responses as incorrect. This threshold may be sufficient for identifying rapid guessing (e.g., the Wise & DeMars, 2005, approach); however, if a single response time threshold were to be specified, the time should be notably higher in order to minimize potential bias. On the other hand, it is important to note that a threshold of 5 seconds reduces bias relative to a rule that assigns all omitted responses to the incorrect response category.

Table 7:
Expected Probabilities for the 5-Second Response Time Threshold - Literacy

Model	Factor	Min	Q1	Median	Q3	Max	Mean	*SD*
1		0.016	0.085	0.152	0.257	0.891	0.194	0.161
2	Low total time	0.006	0.093	0.196	0.368	0.992	0.270	0.254
	Moderate total time	0.005	0.131	0.284	0.551	0.994	0.357	0.269
	High total time	0.019	0.204	0.488	0.706	0.995	0.478	0.283
3[a]	Low ability	0.006	0.119	0.235	0.466	0.992	0.308	0.261
	Moderate ability	0.005	0.131	0.284	0.551	0.994	0.357	0.279
	High ability	0.004	0.141	0.327	0.643	0.995	0.401	0.294

[a] Median total time used

Table 8:
Expected Probabilities for the 5-Second Response Time Threshold - Numeracy

Model	Factor	Min	Q1	Median	Q3	Max	Mean	*SD*
1		0.030	0.212	0.368	0.547	0.919	0.401	0.223
2	Low total time	0.008	0.177	0.319	0.699	0.997	0.432	0.355
	Moderate total time	0.005	0.124	0.338	0.866	0.998	0.470	0.369
	High total time	0.021	0.207	0.466	0.930	0.999	0.540	0.380
3[a]	Low ability	0.008	0.117	0.287	0.833	0.997	0.448	0.361
	Moderate ability	0.005	0.124	0.338	0.866	0.998	0.470	0.360
	High ability	0.004	0.135	0.405	0.910	0.999	0.490	0.355

[a] Median total time used

Limitations

The use of response time information to inform the coding of omitted responses appears to be a tractable solution in a fair number of cases; however, the use of response times does have its limitations. The key limitation is that the underlying causal mechanism for the omitted responses is unknown. Response times can be used to clearly articulate whether an item was presented, and rapid responses may be suggestive of lack of motivation, particularly for low-stakes tests, but given that there appears to be some relationship among omitted responses, response times, and examinee ability, we cannot assume that all of the omitted responses are missing at random. As such, it is unlikely that we will be able to completely eliminate bias in estimates of item and ability parameters. Still, there is reason to suspect that recoding some of the omitted responses as incorrect on the basis of response time is a valid approach. When a clear distinction can be made between omitted and incorrect responses with a high degree of probability, it seems reasonable to treat the omitted responses with corresponding times above the identified threshold as incorrect while leaving the omitted responses with times below this point as not reached/not administered even if the response mechanism is not fully revealed by looking at timing data alone. However, conservative rules may still be applied (setting a maximum value for any item-specific threshold or a minimum of the expected probability level) to ensure that no overcorrection is applied to the data.

For the present study, the emphasis on reducing potential bias in estimates of item parameters and country performance is considered in the context of low-stakes testing where the results for individual examinees are not reported and plausible values for examinee performance, rather than point estimates of ability, are used to obtain estimates of country proficiency. On the other hand, for high-stakes testing, there may be concerns about fairness in the recoding of item responses. If scores are to be reported for individual examinees, consistent scoring rules should be applied for all examinees. That is, while there may be a potential reduction in bias for some examinees, different estimates of performance could be obtained for other examinees depending on how quickly (or slowly) they respond to given items. If examinees are unaware that response times are a consideration for scoring, they may respond in a manner that does not benefit them (e.g., taking a longer time to try to figure out an item rather than skipping it altogether). But again, this is a greater concern in high-stakes testing where individual results are reported as opposed to low-stakes testing where the focus is on group performance.

Conclusion

In any large-scale assessment, there is a strong likelihood that some examinees will omit responses to one or more items. When these omitted responses are coded as incorrect, estimates of item and ability parameters in a latent variable model are likely to be biased. Using the response time information from computer-based tests may provide an alternative way to treat these responses for the purpose of parameter estimation and scoring, apart from model-based approaches that directly model the dependency between nonresponse propensity and skills. Examinees who move from one item to the next quickly, as evidenced by the timing information, may not have had sufficient time to respond to the item(s); hence, there may be value in treating these data points as not reached and excluding them from the likelihood function. Based on a descriptive analysis of the response time information, there appears to be some demarcation between omitted and observed responses, typically with omitted responses occurring with short response times. We employed a model-based approach to examine these demarcations more systematically and identify item-specific thresholds to distinguish between omitted and incorrect responses. We fit three binary logistic models to literacy and numeracy data from PIAAC. One model was specified using item response time as the only predictor, another included total test time as well, and the third added plausible values for examinee ability.

The model that included examinee ability fit the data best, followed by the model that included total time, and lastly by the model that only included the item time. In both content domains, the simplest model using only item response times generally produces thresholds that are not appreciably different from the thresholds identified when including examinee ability and/or total time; hence, the simpler model may be preferable in practical applications. In consideration of the other models, however, the thresholds tend to drop slightly relative to the simplest model, and when accounting examinee ability, the thresholds do decrease as examinee ability increases (i.e., higher ability students would generally have a lower threshold).

In literacy, the median threshold times, at an expected probability level of .8, tend to be around 30 seconds, whereas for numeracy the median threshold times are closer to 20 seconds. This is notably greater than the 5-second threshold commonly used to identify rapid responses; hence, he application of a 5-second rule for retaining omitted responses as not reached may be too liberally assigning incorrect responses for most items, particularly given that this threshold has an associated expected probability lower than chance of distinguishing between omitted and incorrect responses. Taken together, the results of this study suggest that the use of item-specific thresholds is defensible and this approach may provide a feasible mechanism for coding omitted responses in order to minimize potential bias in estimates of item parameters and examinee ability.

References

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39,* 331–348.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77.Glas, C. A. W., & Pimentel, J. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement, 68,* 907–922.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67,* 606–619.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359–379.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615–633.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177–196.

Moustaki, I., & Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of the Royal Statistical Society*, *Series A*, *163*(3), 445–459.

O'Muircheartaigh, C., & Moustaki, I. (1999). Symmetric pattern models: A latent variable approach to item nonresponse in attitude scales. *Journal of the Royal Statistical Society*, *Series A, 162*(2), 177–194.

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with IRT* (ETS-Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34,* 213–232.
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier and D. Hastedt (Eds.), *IERI monograph series: Issues and methodologies in large scale assessments* (Vol. 2, pp. 9–36). Princeton, NJ: IEA-ETS Research Institute.

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). London, UK: Chapman Hall/CRC Press.

Wise, S. L., Bhola, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21–30.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43,* 19–38.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 16,* 163–183.

Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89–99) New York, NY: Waxmann.

Appendix

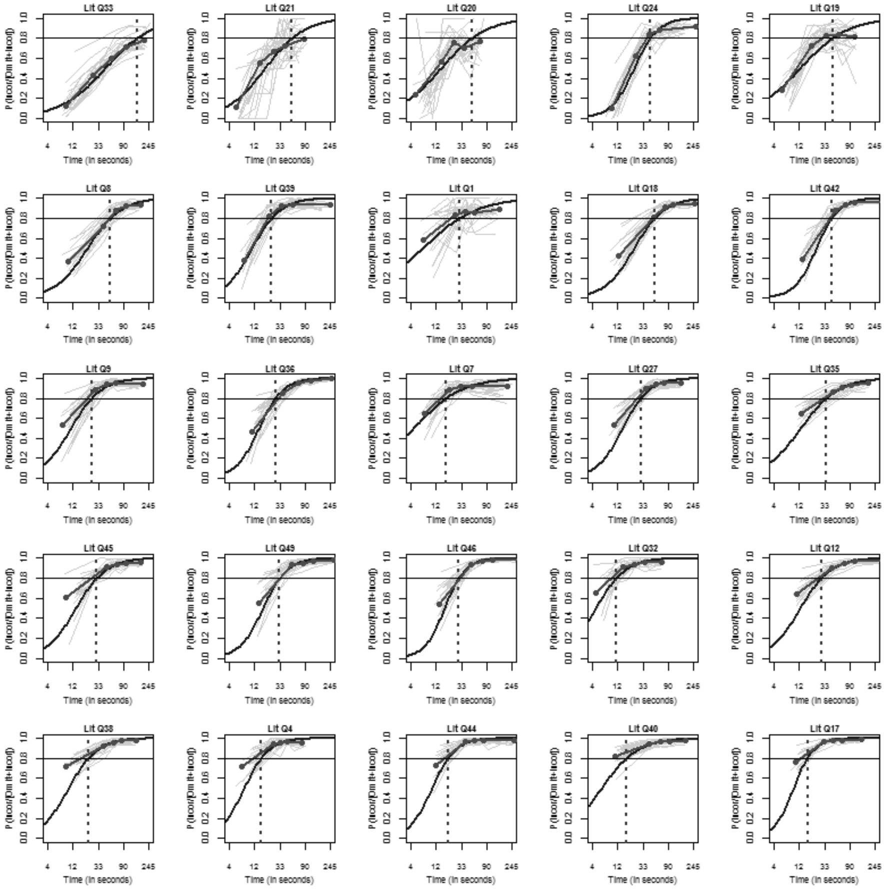
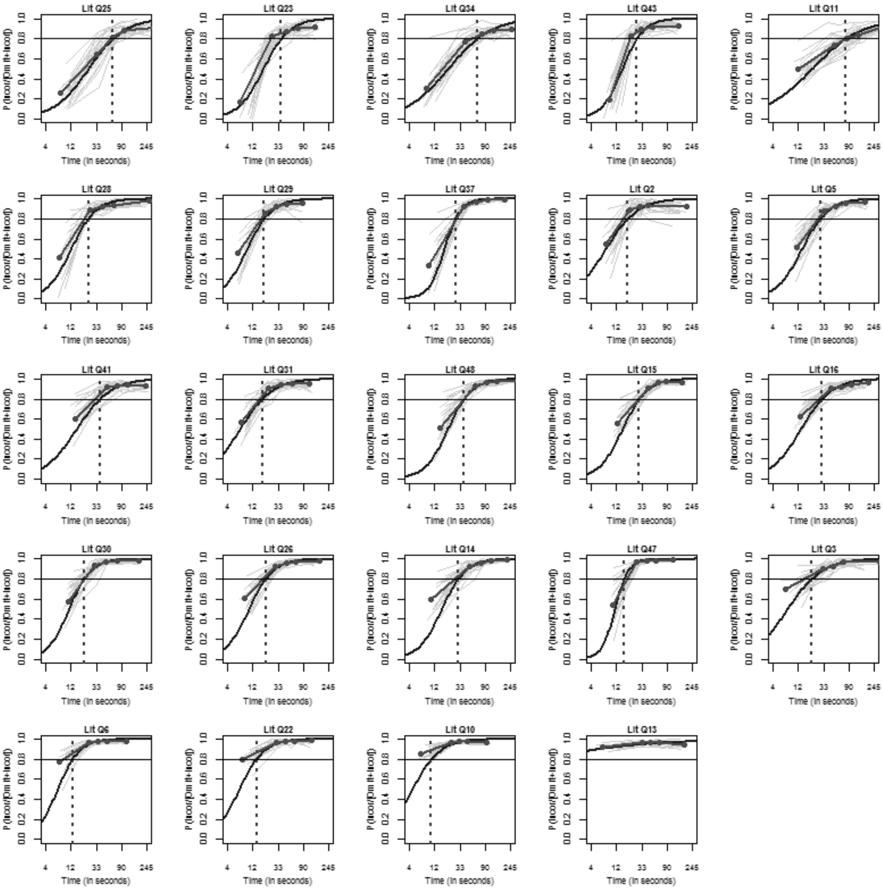


Figure A1:
Model 1 fit plots – literacy



Continuation of Figure A1: Model 1 fit plots – literacy

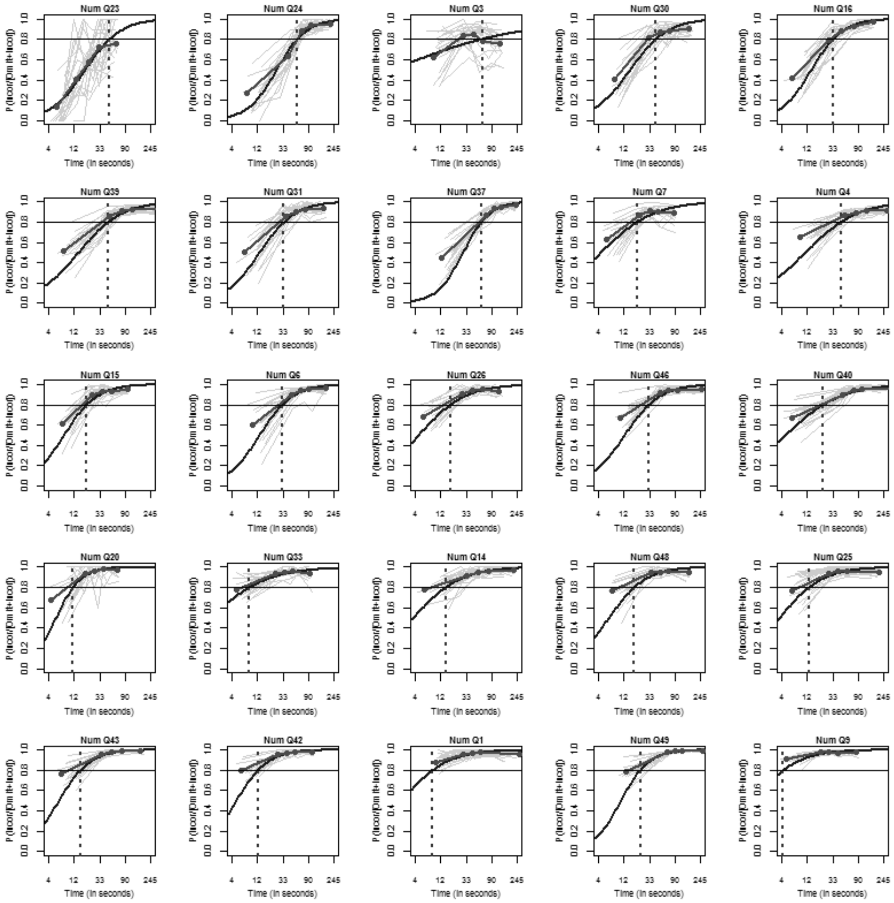
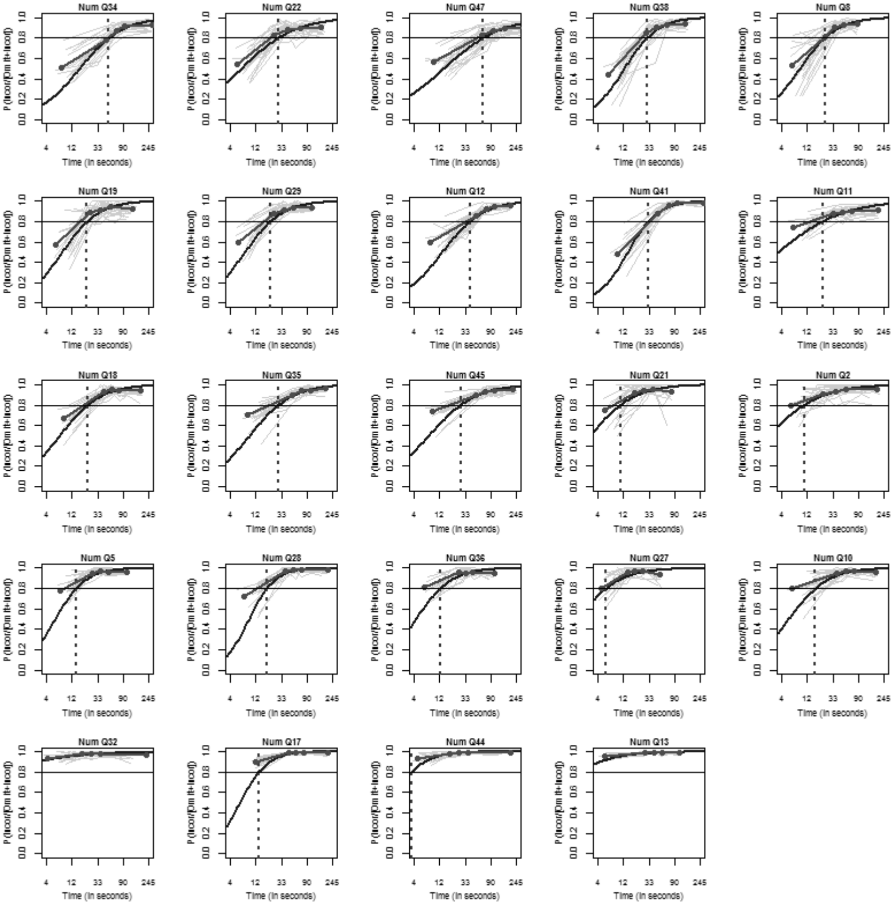


Figure A2:
Model 1 fit plots – numeracy



Continuation of Figure A2: Model 1 fit plots – numeracy