Psychological Test and Assessment Modeling, 2025.67:45-66
DOI: https://doi.org/10.2440/001-0018

Treating Rapid Responses as Errant Improves Reliability of Estimates for Test Performance for NAEP

Daniel B. Wright and Sarah M. Wolff

Department of Educational Psychology, Leadership, & Higher Education, UNLV

Abstract:
Response times are often available for educational assessments and psychometricians have proposed methods for using them when estimating test performance. Several approaches have been explored to see if estimates can be improved. Previous research has shown that a simple mechanism, based on the idea that some people who guess do so rapidly, can improve the reliability of estimates for personalized formative assessments and college-admission data. The method involves Treating All Rapid Responses as Errant (TARRE) responses. Here we examine if this approach can improve reliability estimates for the National Assessment of Educational Progress (NAEP) eighth grade math assessments in the US, data for which were recently re-leased. Treating rapid responses as errant can improve reliability estimates for multiple choice questions (MCQs), but did not for non-MCQ formats. Male test takers were affected more than female test takers and those in the low and high proficiency categories were affected more than those in the middle proficiency categories. Using this procedure, or any procedure that takes into account response times, outside of a research context has further considerations. The scoring method can affect student behavior and their learning, and these consequences are discussed.

Keywords: Rapid guessing; NAEP; response times; TARRE; assessment

Correspondence:

D. B. Wright. Email: daniel.wright@unlv.edu, or Box 453001, 4505 S. Maryland Pkwy., Las Vegas, NV, USA, 89154-3001. This document was produced using knitr (Xie, 2015), which combines LaTeX with R and runs both to produce the final submitted document. The document was translated to Word for page production. The knitr document is available from the authors.

The duration to complete a task, here called response time, offers a glimpse into the cognitive processes involved in completing that task (Luce, 1986). Cognitive scientists have modeled the relationship between response time and response accuracy for simple tasks– usually those taking just a second or two–where participants are assumed to be thoughtfully working just on that task (Ratcliff, 1978). Typical items on an academic assessment are more complex than those used in cognitive science studies and students taking academic assessments have different motivations than laboratory participants. With increased use of computerized testing, the time it takes for students to respond and other process data (e.g., number of clicks, cursor/mouse movement) are often recorded. Statisticians and psychometricians have described ways in which item response times might be used: test security (Sinharay, 2021; Steger et al., 2021), identifying low-motivation responding (Wise, 2017), test speededness (Feinberg et al., 2021), estimating test performance or ability (Silm et al., 2020), *etc*. The interest here is estimating test performance, which is related to, but not that same as, estimating test takers' ability/knowledge. Other factors (e.g., motivation, sleep the previous night) can affect performance.

The cognitive science approach typically involves modeling the processes required for completing specific tasks that are simple and are hypothesized to require the same cognitive processes. An example is presenting people with dozens of trials with two grey squares and asking the people to indicate, as quickly as possible, which square is brighter (Ratcliff et al., 2018). The models (e.g., Ratcliff's diffusion model and other sequential sampling models) perform well when there is a single simple cognitive process necessary to answer the questions correctly (Ratcliff et al., 2015; Voss et al., 2019). Because different sets of multiple processes are involved in answering different academic questions, many of the models that have performed well for the response time and accuracy relationship in laboratory tasks perform poorly when applied to academic assessments (Wright, 2016).

A more common approach in educational assessment is to assume that there are latent variables related to both ability and speed, for both people and items, and that these may be correlated. One of the most popular of these approaches is van der Linden's hierarchical model (2007, 2011). Several other psychometric models have been introduced to try to account for the relationship between time and accuracy, and to provide estimates of students' performance (Bolsinova & Molenaar, 2018; Bolsinova & Tijmstra, 2019; De Boeck & Jeon, 2019).

One issue with these models is while they can work well in research contexts, they would be difficult to explain to test takers and test administrators. For assessments to be perceived as fair it is useful that the scoring procedure is as transparent as possible. In the US, the three main academic societies (American Education Research Association, American Psychological Association, and the National Council for Measurement in Education) created a document to describe best practice for measurement in the educational and social science. In it they stress that it is important "to balance technical concerns and transparency" (2014, p. 206).

Treating All Rapid Responses as Errant (TARRE) and the National Assessment of Educational Progress (NAEP)

Wright (2016) proposed a simple method that uses response times to estimate overall test performance. The method is to Treat All Rapid Responses as Errant. The first test of this was with ACT mathematics items. He discussed how long it should take for people to answer these items with those on the ACT mathematics item development team. They felt it was unlikely that anyone would be able to provide a thoughtful response to any of their questions in less than ten seconds. ACT mathematics questions are not items like "What is 4 × 3?" They require integrating and manipulating multiple bits of information. Therefore, even if someone rapidly answered one of the multiple choice questions correctly, they would be unlikely to have used appropriate cognitive processes. In lay terms, they likely made a lucky guess. This randomness decreases the reliability of the test performance estimates. Rapid responding can occur when guessing is encouraged because of time limitations, where guessing has a high likelihood of producing some correct responses, as it does with multiple choice and True/False questions, and where most test takers are motivated to score high. The ACT is a timed assessment, has multiple choice questions, and most of the students are motivated. This situation would encourage some guessing. Treating all responses faster than ten seconds produced more reliable estimates of test performance.

This procedure was subsequently used with several formative assessments that were part of a personalized learning system. Treating rapid responses as errant increased the predictive value for the estimates with these assessments (Wright, 2019b). This was shown to be true across different subject areas: history, languages, science, English, and mathematics.

The National Assessment of Educational Progress (NAEP) results are one of the main sources for comparing how well students in different US states are performing, but the results do not impact individual students' grades. Therefore, there is concern with this assessment that many students may not be expending appropriate cognitive effort. Lee & Jia (2014) examined rapid guessing with NAEP and found, overall, the amount of rapid guessing that was identified was small. One goal of this paper is to evaluate whether treating rapid responses as errant can still improve the reliability of test performance estimates when the amount of guessing appears small. An additional goal is to examine other potential boundary conditions for the efficacy of this approach. NAEP using multiple choice questions (MCQs) but also uses non-MCQs where the chances of correctly guessing are low. We predict this approach will not improve the reliability of estimates for non-MCQs.

We are estimating some underlying construct that drives test performance, rather than something more specific like the students' abilities. Wise (2017) has used rapid response times to attempt to measure ability for assessments used to evaluate teachers (and also schools, states, *etc.*). He assumes that many test takers make rapid responses for single items (or a few items) because of low motivation for those items and that the

probability of them answering correctly based on their true *ability* would usually be higher than chance.

The approach here is to estimate test performance, trying to lessen both the impact and the use of the strategy of rapid guessing. There are a large number of non-ability factors that influence test performance (e.g., feeling unwell on the test day, test-room distractions). Some of these, like low motivation, have behavioral signatures (i.e., low motivation → rapid responding). There are, however, other factors that can also lead to some of the same behavioral signatures (e.g., rapid guessing when time runs out, guessing on items that from a glance are on a topic not studied). Further, assuming that aspects of the scoring algorithm are known to test takers, it is important to consider any behavioral consequences caused by a change in the algorithm used. In the discussion it is argued that the proposed scoring algorithm could increase thoughtful responding, and thereby increase learning due to increasing the conditions that maximize the testing effect (Karpicke & Roediger, 2008).

The TARRE scoring algorithm is:

Treat all responses–whether accurate or inaccurate–that are made faster than some threshold value as errant responses and then proceed with these new data to estimate test performance.

Suppose a student answers questions 4 and 5 on a science test faster than the threshold. Suppose their answer for question 4 was correct, but their answer for number 5 was incorrect. The TARRE procedure would treat both of these as incorrect. The modified values would be used to estimate test performance. This might be made with IRT or summing the responses or whichever method one would have used with the original responses. Examples using an R function tarre for different methods to combine response accuracy and times are provided in Wright (2019b).

Choosing a Threshold

Several approaches have been described for identifying rapid responding (Kong et al., 2007; Lee & Chen, 2011; Rios, 2022; Rios & Deng, 2021, 2024; Soland et al., 2021). One approach is to use a single threshold for all items. The choice of what value to use can be based on what someone thinks is the typical response duration of someone not reading an item (e.g., $< 3s$). An alternative is choosing a threshold based on the amount of time it would take to thoughtfully answer the question. The ACT study (Wright, 2016) had subject matter experts (math item developers at ACT for their math assessment) suggest a single threshold ($10s$) below which they felt that *none* of the questions could likely be answered using appropriate cognitive effort.

Another approach is to try multiple thresholds and see which threshold produces the best results in terms of reliability. This approach was used in a study of personalized learning in high schools (Wright, 2019b). Using multiple thresholds was necessary as students were taking, on average, about 100 assessments with mostly different questions that

require different amounts of cognitive processing. It was found that the optimal threshold varied by subject with the math reliability counting answers made quicker than six seconds as errant producing the highest estimates of reliability and language assessments having a threshold of about four seconds. These are lower than the threshold assumed for the ACT data as many of these items are less complex than ACT items. The language assessment had some vocabulary items that could be answered fairly quickly. One conclusion from that study was that it is important to examine the items and depending on the threshold method (e.g., if there is a single threshold for all items) exclude those that could be answered quickly.

Choosing a single threshold for all items might be useful for describing the scoring method to test takers and it may be necessary for some adaptive assessments because of how items are picked. This approach has the disadvantage that some items require relatively little time to provide a thoughtful response and other items require a much longer amount of time for a thoughtful response. A counter argument to this is that guessing, without reading the item, should take similar times for all items. This suggests that for some items there will be response times too rapid for being able to "guess" yet too slow for thoughtful responding. Here we allow different thresholds for each item. We examine two approaches for this that take into account the distribution of response times for all test takers.

The Percentile Method

The first approach is to use a threshold based on the percentile for the response times. For example, if a test taker is in the quickest 5% of responses for an item, that could be considered a rapid response. The times associated with each of these variables will depend on the distribution for the sample. Since we do not know what proportion of people may be engaging in rapid responding, we examine thresholds from 0% to 10% and treat all responses faster than this percentile as errant and examine if this improves reliability. A consideration for this method is that it assumes that the same proportion are engaging in rapid responding on each item. It may be that there are certain items that more test takers rapidly guess on compared with others. This might be due to their placement in the assessment (e.g., towards the end) or characteristics of the question (e.g., the item starts with terminology unknown to many test takers).

The Fraction of the Median Method

The second approach comes from Wise & Ma (2012). They discuss using a fraction of the mean response time for that item. Because the response times are often positively skewed and therefore the mean is less robust than the median, a fraction of the median response time was used here. The median will usually be less than the mean for response times because response times are usually positively skewed. We explore what

fractions appear to work well for NAEP math data. We examine different fractions from 0 to 50%. This approach has that advantage that if no test taker was guessing for an item and none responded faster than the fraction of the median, none would be labeled as such. This has the potential problem that if a few people are using a much faster, but accurate, response strategy (e.g., using multiplication for the 4 + 4 + 4 + 4 + 4 problem) they could be classified as rapid responders. No threshold method can guarantee there will be no mis-classifications. Conceptually the fraction of the median approach seems to have the fewest conceptual issues of those considered.

There are other possibilities for establishing thresholds. One option is that thresholds could be established by those developing the tests for individual items. Test developers can use think-aloud protocols during development and list the necessary steps for solving the problem as intended. They could ask people to expend appropriate effort, work quickly (and perhaps be experts), do the task, and then times faster than these people could be classified as non-thoughtful (rapid) responding. This would require much work by the testing companies and, as far as we know, is not currently done.

Second, there are arguments that an individual test taker's times should be taken into account. Times could be identified as rapid outliers by taking into account characteristics of both the test taker and the item. A simple model would be the following multilevel/mixed model $ln(resptime_{ij}) = \beta_0 + u_j + v_i + e_{ij}$, where u_j is for item variability and v_i is for test taker variability, and then e_{ij} less than some threshold could be considered too rapid (Goldstein, 2011; Wright & London, 2009). However, if a person responds very slowly for most items and then as quickly as most people for one item, this could appear as rapid responding even though the person may have expended the same cognitive effort as most of the other people on this item.

There are considerations for all methods of choosing appropriate thresholds. Simulations have been done to examine which item and person characteristics affect the impact of using this procedure (Rios & Deng, 2024; Wright, 2019b).

We examine which demographic groups tend to be more negatively affected, though one of our main conclusions is that implementing any new scoring mechanism may change behavior so it will be important to consider these and how different groups adapt.

Evaluating TARRE

There are several approaches that can be taken to evaluate the TARRE approach. Here, we apply TARRE separately to the MCQ items and to the non-MCQ items, and measure Cronbach's α (Cronbach, 1951) for the entire set of items when applying (and not applying) TARRE to these subsets. There are issues with how people interpret α and issues comparing α values across different sets of items (e.g., McNeish, 2018; Revelle & Condon, 2019; Thompson, 2003), but here as the sets of items are the same this can be used for comparing using the TARRE procedure with not using it. The higher the α the more reliable the assessment. Other measures could also be used, but given the

popularity of α it was chosen. This is done for both of the methods used for creating thresholds, and a range of values for each of these methods.

Aims of Study

The purpose of this study is three-fold. First, the approach is used for a very different type of assessment from those used in earlier studies (ACT and several formative assessments). Finding the limits of when an approach can improve ability estimates is important for knowing when it can be implemented and when it should not be implemented. Second, the format of the individual items vary. Some are multiple choice, others are not. The rationale for the TARRE approach and simulations show it should only be of value when there is a substantial probability of accurate guessing (Wright, 2019b). Here the procedure is applied to the MCQ and non-MCQ items separately. The prediction is that TARRE should only improve estimates for the MCQs. Finally, for any change in scoring policy some people's scores will rise compared with others and some people's scores will drop. The method examined here will affect people who respond rapidly. It is important to examine if there are particular groups that are more affected than others. If there are differences among groups it is important to stress that this does not mean the TARRE approach is unfair in an absolute sense for those groups, only that its fairness differs from the traditional approach. It may be that the traditional approach is unfair to the groups that have their relative scores raised.

Methods

The NAEP data

NAEP (the National Assessment of Educational Progress), often referred to as the Nation's Report Card, is an assessment mandated by the US Congress that consists of a carefully sampled group of US students. This allows NCES (the National Center for Educational Statistics) to monitor progress in individual states and districts as well as make comparisons across these groups. Details about NAEP can be found at https://nces.ed.gov/nationsreportcard/about/ (all websites as accessed 28 April, 2025). NCES released process data from its 2017 Math assessment for 8th grade and more details at www.nationsreportcard.gov/process_data/ (Bergner & von Davier, 2019). The data examined here are from a thirty-minute session of 2,800 students on questions where they were allowed to use a calculator. For some parts of NAEP's math assessment they are not allowed to use a calculator. Note that all sample sizes are rounded to the nearest ten in accordance with the NCES guidelines. Nine percent (n = 260) had some sort of accommodation, including different time restrictions. Because these change the test ad- ministration conditions, these test takers are omitted from our analyses leaving n = 2, 540.

There were 19 questions in the session analyzed. Five of them were masked, but the response format could be observed from the information released to us. The wording for the other 14 questions comes with the information available from NCES when applying for access. We do not report them here. Sample questions are available at: https://www.nationsreportcard.gov/math_2017/sample-questions/?grade=8. These include multiple choice questions, tick all appropriate questions, fill-in the number ques- tions, and some technology enhanced questions where students drag multiple responses to multiple answer boxes.

Referring to the questions by number, the response formats are shown in Table 1. Two of the items allowed partial credit. In order to standardize our analyses these questions were dichotomized to be wrong or right. The first of these required filling in five numbers, but each was based on applying the same rule. Less than 10% received partial credit. We counted as correct only those that got full credit. The second partial credit question required test takers to write three numbers. The first two numbers required reading information off a plot and the third required using this information. We gave credit to people who received partial (25% of the sample) or full credit (27% of the sample) because this meant they successfully completed the first task. The overall proportion accurate for each item is shown in parentheses. Guessing completely at random would result in a 20% probability of correctly answering the five alternative MCQs. Note that some of the proportions correct are low. For example, Q8 (a five option MCQ) has proportion of .21, which is not significantly different from .20 ($\chi 2(1) = 2.63$, $p = .212$). The drag answers question required test takers to drag four different numbers to four boxes. There are 4! = 24 possible ways to do this. Thus, if guessing completely at random the probability of being correct is about 4%. The fill-in question would require the test taker to know the types of answers that are appropriate (e.g., numbers, words). Even if this was known, the probability of accurately guessing would be low. Here, the MCQs are treated as having a substantial probability of accurately guessing while the others are not.

The log/process data of the key strokes from NCES include a time at the start of each question and when the test taker submits their response. The difference between these times–the response duration–is our measure of response time. The test takers may have been doing non-test taking behaviours during this period (e.g., sneezing). Because of the nature of these data and the rules for accessing information from the US government, no information is available about which school/classroom the individual test takers are in.

Table 1
Question formats for the 19 items and the proportion correct for each in parentheses.

Format	Item #s
Five alternative MCQ	Q1 (.55), Q2 (.67), Q3 (.50), Q5 (.37), Q6 (.36), Q8 (.21), Q10 (.73), Q11 (.37), Q14 (.32), Q15 (.59), Q16 (.34), Q17 (.50), Q18 (.31), Q19 (.56)
Drag multiple answers to boxes	Q4 (.85)
Fill-in multiple values	Q7 (.72), Q12 (.55)
Fill-in single value	Q9 (.22), Q13 (.23)

SOURCE: U.S. Department of Education, National Center for Education Statistics, 2017 NAEP Grade 8 Math Response Process Data RUF, License number: 20090018.

Results

The results are divided into three sections. First, we report some descriptive statistics for the response times. Second, we explore if TARRE can be used to increase the reliability of the test performance estimates when applied to the MCQs and to the non-MCQs. We predict it will improve the estimates when applied to the MCQs, but not when applied to the non-MCQs because the probability of correctly guessing answers for these is negligible. We examine this for thresholds calculated in two ways: a percentile for each item and a fraction of the median for that item. Different percentiles and different fractions are tested to see which of these produces the most reliable estimates for test performance. We calculate Cronbach's α for the scales with and without TARRE, with the prediction that Cronbach's α will increase if using TARRE for some thresholds, but only when applied to the MCQs. The third section examines if using TARRE in this way affects some groups differently. The R environment is used for all analyses reported (R Core Team, 2023).

Descriptive statistics for Response Times

Figure 1 shows the distribution for the response times and the natural logarithm of the response times. Across all items, the response time variable is positively skewed

(skewness = 3.09). Logging the data reduces this skewness (skewness = 0.40). The QQ plot in the bottom row of Figure 1 still shows these are not perfectly normally distributed. Other transformations (e.g., Wright, 2024) could be used, but given the widespread use of natural logarithms with response times it will be used here.

Descriptive statistics for the response times are shown for each question are shown in Table 2. The distributions are positively skewed. The statistics for the natural logarithms of the response times are also shown. There is much variability in how quickly people answered the questions. With only 19 items and all asked in the same order it is difficult to judge whether any patterns in the responding are due to item characteristic or their order of presentation. The correlation between when an item was asked and the median response time was non-significant: $r = -.37, df = 17, p = .120$. With the small number of items it is important to be cautious interpreting this non-significant result, particularly without knowing how NAEP decide the order to ask questions.

We examined if people whose response patterns with respect to accuracy stood out for having high or low response times. There are several ways to measure the fit for response accuracy. Artner (2016) recommends using H_t (Sijtsma, 1986). This is available in the R package **PerFit** Tendeiro et al. (2016). There are a handful of people with high and low response times with low fits, but overall the association was small; the correlation between a person's H_t and the mean of their logged response times was: $r = .05$.

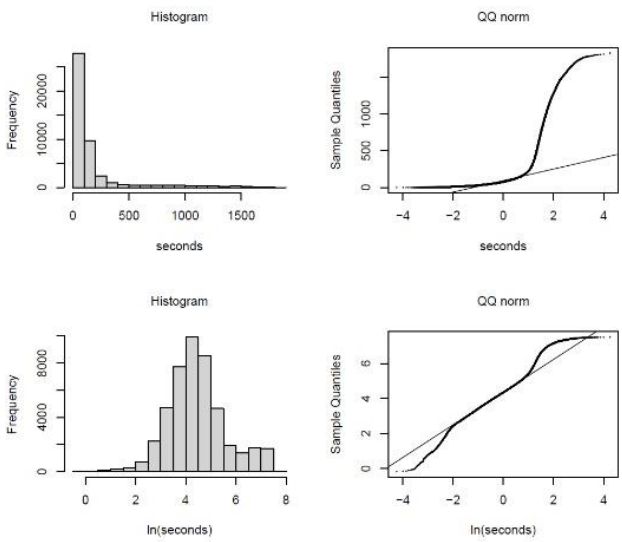

Figure 1
Histograms of response times and the natural logarithms of response times and their QQ-normal plots.
SOURCE: U.S. Department of Education, National Center for Education Statistics, 2017 NAEP Grade 8 Math Response Process Data. License number: 20090018.

Can TARRE increase the reliablity of the test performance estimates?

The TARRE method was applied to data using the two methods for classifying rapid responding for the MCQ and non-MCQ items separately. These methods are the fraction of the median and the percentile methods, and for each different values were examined (i.e., different fractions and percentiles, respectively). Cronbach's αs were calculated for these and shown in the left and right panels of Figure 2.

The horizontal lines show the value of α if TARRE is not used (0.82). The red lines show the α values when applying TARRE to the MCQ items and the blue lines show the α values when applying TARRE to the non-MCQ items. The left panel shows the results for the fraction of the median method for choosing the threshold for classifying the response as a rapid one. TARRE increases the Cronbach's α for MCQs (the red line), but not for non-MCQs (the blue line). The increase occurs for the fraction of the median method between about .1 and .3. The highest α corresponds to a fraction of 0.21 of the median. The maximum α was 0.82. The blue line remains flat as few responses were less than .2 of the median. It then drops immediately. No threshold values for the non-MCQs increased the α.

Table 2
Descriptive statistics for the response times for the 19 questions.

	Raw times				Logged times			
Item	Mean	Median	sd	skew	Mean	Median	sd	skew
Q1	206.69	27.00	462.63	2.39	3.79	3.30	1.42	1.69
Q2	266.16	78.22	459.54	2.20	4.67	4.36	1.18	1.11
Q3	260.26	90.38	426.06	2.26	4.75	4.50	1.13	0.96
Q4	281.64	128.63	391.92	2.25	5.04	4.86	0.99	0.82
Q5	263.21	127.54	364.34	2.28	4.98	4.85	1.01	0.52
Q6	171.94	56.35	324.78	2.65	4.28	4.03	1.09	1.34
Q7	244.69	143.67	286.39	2.54	5.12	4.97	0.77	1.04
Q8	232.49	107.15	313.51	2.07	4.83	4.67	1.06	0.43
Q9	193.21	81.52	281.31	2.22	4.57	4.40	1.10	0.62
Q10	171.66	85.27	233.07	2.49	4.59	4.45	0.99	0.36
Q11	145.69	74.58	204.88	2.72	4.43	4.31	0.99	0.18
Q12	171.85	100.03	197.94	2.38	4.73	4.61	0.85	0.46
Q13	188.98	119.62	192.93	2.04	4.81	4.78	0.96	-0.35
Q14	123.23	73.67	143.33	2.74	4.40	4.30	0.89	-0.05
Q15	89.22	47.32	126.46	3.22	3.95	3.86	0.98	0.21
Q16	67.56	26.40	117.67	3.35	3.48	3.27	1.06	0.85
Q17	70.87	31.02	124.97	4.84	3.61	3.43	1.01	0.68
Q18	92.68	56.47	129.94	5.01	4.02	4.03	1.01	-0.36
Q19	112.82	62.17	179.85	5.30	4.15	4.13	1.06	-0.14

SOURCE: U.S. Department of Education, National Center for Education Statistics, 2017 NAEP Grade 8 Math Response Process Data RUF, License number: 20090018.

The percentile method produced the same substantive finding. The highest increases in α when applying TARRE to the MCQs was for percentiles of about 1% to 5%, with the value producing the highest α being .03. The maximum α was 0.82. Applying TARRE to the non-MCQs lowered α.

Both methods for choosing a threshold had values that increased α when applied to MCQs, but the increase is small. Importantly, the α values dropped when the threshold increased. Further, there were no values for the threshold that increased α when TARRE was applied to the non-MCQs.

How different groups are affected by TARRE

It is important to assess if TARRE affects groups differently. Finding a difference does not mean TARRE is unfair to certain groups; it may be that not using TARRE is unfair to certain groups. If implemented it will also be necessary to assess fairness after test takers are trained to take into account this method. We compare these for gender, ethnicity, parents' education level, and the students' proficiency level. It is important to stress that overall the effects of TARRE are small, and the differences among groups are also small.

If using the number of correct answers as a measure of test performance, this can only decline or stay the same using TARRE. However, if the test performance construct is estimated from models other than the number correct that standardize estimates in some way, estimates for test takers who had no rapid responses will tend to increase slightly. The θ from the 2PL IRT model is used here with a threshold of 0.215 fraction of the median. The shifts were all small, a slight decrease for males and a slight increase for females. The percentage of males decreasing was 20.38% and the percentage increasing was 79.62%. The corresponding numbers for females are: 17.94% and 82.06%. These, with percentages for other comparisons, are shown in Table 3.

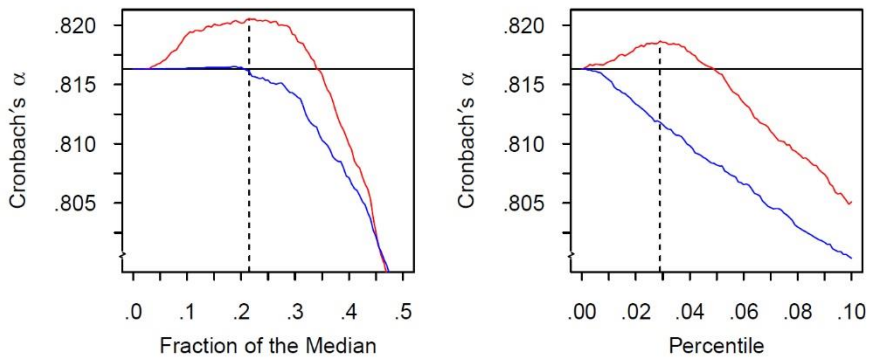

Figure 2

Finding the optimal threshold for the proportion of the median and the percentile threshold method using Cronbach's α. The red lines are when TARRE is applied to the MCQs and the blue lines are when TARRE is applied to the non-MCQs.

SOURCE: U.S. Department of Education, National Center for Education Statistics. 2017 NAEP Grade 8 Math Response Process Data. License number: 20090018.

Table 3

The percentage of people whose score decreases and whose score increases due to using TARRE for different groups.

SOURCE: U.S. Department of Education, National Center for Education Statistics, 2017 NAEP Grade 8 Math Response Process Data. License number: 20090018.

Variable	Value	Decrease %	Increase %
Gender	Males	20.38	79.62
Gender	Females	17.94	82.06
Ethnicity	White	23.34	76.66
Ethnicity	African American	9.98	90.02
Ethnicity	Hispanic	13.36	86.64
Ethnicity	Asian	34.23	65.77
Ethnicity	American Indian	13.64	86.36
Ethnicity	Nat. Hawaiian	4.55	95.45
Ethnicity	Mixed - Not Hispanic	21.52	78.48
Parent Education	Some HS	14.09	85.91

Parent Education	Grad HS	11.96	88.04
Parent Education	Post HS	14.90	85.10
Parent Education	College	24.65	75.35
Proficiency	Below	8.27	91.73
Proficiency	Basic	5.89	94.11
Proficiency	Proficient	31.65	68.35
Proficiency	Advanced	69.51	30.49

TARRE had a more negative effect on males than on females. The mean difference scores for each of the categories are shown in Table 4. A common effect size for comparing means is ω^2 (Hays, 1981). This has the advantage that it can be used both when there are two groups and when there are more than two groups. The omega_squared function from **effectsize** is used (Ben-Shachar et al., 2020). Common verbal labels for effect sizes for η^2 apply also for ω^2: small of $\omega^2 = .01$, medium of $\omega^2 = .06$, and large of $\omega^2 = .14$ (Cohen, 1988). Caution is urged using these verbal labels because the meaning of any effect size magnitude is context dependent (Lipsey et al., 2012), but these labels allow comparisons across different contexts. From this perspective none of the differences in Table 4 reaches small criterion. Our focus will be on comparing pairs of categories.

Table 4

Group differences in the effect of treating rapid responses as errant. High scores represent groups that had higher test performance estimates using TARRE than not using TARRE. We print to three digits here, but be aware these are small differences.

Variable								ω^2
Gender	male	female						
	-0.005	0.004						0.004
	1210	1330						
Ethnicity	White	African American	Hispanic	Asian	Amer Indian Alaskan	Native Hawaiian Pacific Island	mixed not Hispanic	
	-0.001	0.003	0.000	0.002	-0.014	0.002	0.005	0.000
	1330	430	520	110	40	20	80	
Parent Ed.	some HS	Grad HS	Post HS	College				
	-0.008	0.002	-0.001	0.000				0.000
	150	330	350	1160				
Proficiency	below	basic	proficient	advanced				
	-0.010	0.006	0.003	-0.006				0.001
	680	980	630	250				

NOTE: There were many missing values for parents' education: 12% or 300 of 2540. Sample sizes rounded to nearest 10.
SOURCE: U.S. Department of Education, National Center for Education Statistics, 2017 NAEP Grade 8 Math Response Process Data RUF, License number: 20090018.

The difference variable (θ_{TARRE} - $\theta_{woTARRE}$) was negatively skewed: -7.41. The people who were affected most dropped by much more than those whose scores went up slightly. The traditional intervals for skewness are less reliable than bootstrap intervals (Wright & Herrington, 2011), so bootstrap intervals are used here: 95% BCa CI = (-8.82, -6.39). Statistics that assume normally distributed residuals, like *Student*'s *t*-test, would be inappropriate. However, interest is in shifts in means, so rank-based methods, like the Wilcoxon, are also inappropriate. Therefore bootstrapping is used to compare differences between pairs of conditions (Davison & Hinkley, 1997; Efron & Tibshirani, 1993; Wright et al., 2011). Several bootstrap functions are available in R packages for bootstrap versions of the two-group *t*-test. Bias corrected and adjusted (BCa) intervals were used (Efron & Tibshirani, 1993). The boot.two.bca function from **wBoot** (Weiss, 2016) will be used here. Its default of 9,999 replications is used here.

When there are multiple groups, like with ethnicity or parents' education, compar- isons are made between each pair of values. Both the unadjusted *p*-values and the Holm adjusted *p*-values within those comparisons (i.e., comparing the four proficiency levels, only the six pairwise comparisons are used) are recorded. We discuss differences for both of these because of concerns about Type 1 and Type 2 errors. Those significant only with the unadjusted *p*-values should be viewed cautiously.

There were no significant differences, even without adjustment, for ethnicity or parents' education. For gender, the effect of males being negatively affected more than females had $p = .001$. The most interesting group differences were for proficiency. The significant pairwise differences were:

	Unadjusted *p*	Adjusted *p*
Below proficiency versus basic proficiency	< .001	< .001
Below proficiency versus proficient	< .001	= .001
Basic proficiency versus advanced proficiency	< .001	< .001
Proficient versus advanced proficiency	< .001	< .001

Those group most negatively affected are at the low and high ends of the scale. These differences are examined separately for males and females. The differences among conditions were larger for males than females. For males the same pairwise adjusted p values remained statistically significant, while for females only the difference between the basic proficiency and advanced proficiency ($p_{adj} = .0124$) and proficient and advanced proficiency ($p_{adj} = .011$) were significant.

There are important differences between the low and high proficiency group. The low group mean decrease was due to several people losing a large amount from TARRE. The high proficiency group had a large amount of people lose only a small amount because of one of their correct responses being labelled a rapid response. The standard deviation for the difference in test performance estimates between TARRE and not TARRE were: 0.11 for the low proficiency group, 0.05 for the basic proficiency

group, 0.02 for the proficient group, and 0.05 for the high proficiency group. What this means for implementing TARRE is discussed below.

Discussion

The main conclusion is that Treating All Rapid Responses as Errant (TARRE) improves the reliability of test performance estimates when applied to MCQs, but decreases the accuracy when applied to non-MCQs. The effects were predicted because rapid non-thoughtful responding is unlikely to result in an accurate response when the items require the test taker to fill in a number or require the test taker to drag multiple numbers into boxes because the probability of accurately answering when guessing completely at random is low. The improvement for MCQs is small, but similar to the effects found in other assessments (Wright, 2016, 2019b). People often choose between two approaches when one is only slightly more accurate (e.g., the difference between a *t*-test and *z*-test for large samples is minute). However, it is important to consider other consequences that changing the algorithm would have.

Testing organizations need to consider whether the size of this effect, or other effects of this size (e.g., most involving responses times, most changes in estimation methods), is worth changing the scoring method for. An important consideration is whether the primary positive consequence of adopting this approach, which might be test takers thinking more about difficult items, is sufficient to make up for the costs (both financial and otherwise) of changing scoring methods. There could also be negative consequences if test takers focus on delaying their responses until they are certain that they will not be flagged as rapid. Distractions, including just having a clock present during an assessment, can increase the amount of time pressure time takers feel (e.g., Wolff et al., 2024).

Adapting the approach of Wise & Ma (2012) we used the fraction of the median to decide the thresholds for identifying rapid responses. We tested fractions from 0 to .5 and found similar levels of improvement between .1 and .3. Because academic tests vary both in terms of item complexity and test taker motivation, it is likely that for most applications analysts would want to examine different fractions and find which ones improve the accuracy of the estimates the most. It may be that the test companies would be able to find a value for some of their products (e.g., the ACT or SAT) and establish a fraction that works well for the test so that they can report this in the materials. As discussed above, the *Standards* adopted by these companies stress the importance of transparency (American Educational Research Association et al., 2014).

For research purposes TARRE could be applied to MCQs to get slightly more accurate estimates of test performance for any assessment where item times are available without affecting the scores given to test takers. However, applying TARRE to create scores that are reported to the test takers or other stakeholders requires at least three more considerations. These apply to any new scoring mechanism.

1. Does the procedure affect groups of people differently and does it penalize any groups unfairly?
2. Can the procedure be explained fairly easily to test takers and other stakeholders?
3. How will implementing this procedure affect test taker behavior and their learning? We address each of these in turn.

Are groups affected differently?

We identified two group differences. First, males are more negatively affected than females. Males often respond more rapidly on tests and often perform less well than females (Wright, 2019a). This suggests that interventions that cause test takers to think about responses before guessing would most affect males. While we do not investigate why here, possible explanations for the gender differences in rapid responding have been suggested in relationship to individual differences, personality, motivation, and achievement identity (DeMars et al., 2013).

The other group difference that was observed, and these differences were most prominent among males, was that test takers who were classified in the lowest and highest proficiency categories were more affected than those in the two middle proficiency categories. The distribution of the decrease was different for these groups, suggesting that there are different reasons for these decrements. While some of those at the low end appear to be rapidly guessing for multiple items, those at the high end could be doing a combination of using rapid strategic guessing for selective answers and being able to provide thoughtful a response extremely quickly.

If implementing TARRE on a high stakes exam, care should be taken not to penalize thoughtful rapid responding. If the advanced proficient rapid responses are strategic, implementing TARRE could encourage these people to think about these items for a few seconds more, which in itself would not be bad, but requiring them to use a clock management strategy, which the tests do not aim to measure, would be an unfortunate consequence. It would be important to research with specific high stakes tests how often this occurs and perhaps identify those individuals. This is perhaps the biggest hurdle for implementing any procedure that takes into account response times in high stakes tests.

Can the procedure be explained easily?

Some cognitive science research of response times uses complex models. For example, the sequential sampling approaches (e.g., Ratcliff, 1978) draw on work by Einstein and others used to describe particle motion (i.e., diffusion models Einstein, 1905, see Rigden, 2005, for a summary in English, that also describes the impact of this paper). Details about how this might be used within a scoring method would be difficult to

explain to most test takers. Being able explain the algorithm fits within Knuth's idea of *literate programming*. He argues that "computer programs that are truly beautiful, useful, and profitable must be readable by people" (Knuth, 1992, p. *ix*).

Consequences of implementing TARRE

The focus in this section is on implementing TARRE, but similar issues apply to any implementation that encourages test takers to spend some time on each item, even if they think that there is only a low likelihood that they could answer the item correctly. There is a negative and a positive aspect of these approaches. The negative aspect is that test takers may use cognitive effort to decide whether they have spent enough time on the item rather than additional effort on the item itself. Having a visual cue like a clock showing time spent or time until they can move to the next question could mean test takers focus on the clock. Instructions indicating "if you read the item and the alternatives this will be enough time spent" could encourage test takers to think about how they should approach the question, not their time management. This would require checking that rapidly reading the item and alternatives could not be done so rapidly to tag the response as rapid guessing. The advantage is that thinking about even difficult questions can help you to learn about the topic (Karpicke & Roediger, 2008).

Summary

Rapid guessing is problematic for student learning and for the assessment of their learning. Several methods have been put forward to address this. Here we show a simple method, treating all rapid responses as errant (TARRE), improves the reliability of estimates of test performance. Previous research had shown this approach works with high stakes tests like the ACT and with formative assessments that are part of the personalized learning system. We extend these findings showing that it works with NAEP, an assessment that does not produce a score that directly affects the test taker. Importantly, it was shown to work for multiple choice items, but not for other items.

Particularly for high stakes tests, if there is a change in scoring methods the test preparation programs may change their advice to increase their customers' scores. There are several approaches that can be explained to test takers and other stake holders relatively easily. Each presents some issues that would need to be overcome. The important consideration is how any changes could affect test takers behaviors and the test takers' learning.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational & psychological testing (2014 edition)*. Washington, DC: American Educational Research Association.

Artner, R. (2016). A simulation study of person-fit in the Rasch model. *Psychological test and assessment modeling*, *58* (3), 531-563.

Ben-Shachar, M. S., Makowski, D., & Lüdecke, D. (2020). Compute and interpret indices of effect size [Computer software manual]. Retrieved from https://github.com/ easystats/effectsize (R package)

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*, 706-732. doi: 10.3102/ 1076998618784700

Bolsinova, M., & Molenaar, D. (2018). Modeling nonlinear conditional dependence between response time and accuracy. *Frontiers in Psychology*, *9*. doi: 10.3389/fpsyg.2018.01525

Bolsinova, M., & Tijmstra, J. (2019). Modeling differences between response times of correct and incorrect responses. *Psychometrika*, *84*, 1018-1046. doi: 10.1007/s11336-019-09682-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334. doi: 10.1007/BF02310555

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge, UK: Cambridge University Press.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.00102

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69-82.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap (monographs on statistics and applied probability)*. Boca Raton, FL: Chapman and Hall/CRC.

Einstein, A. (1905). Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, *322* (8), 549-560. doi: 10.1002/andp.19053220806

Feinberg, R., Jurich, D., & Wise, S. L. (2021). Reconceptualizing rapid responses as a speededness indicator in high-stakes assessments. *Applied Measurement in Education*, *34*, 312-326. doi: 10.1080/08957347.2021.1987904

Goldstein, H. (2011). *Multilevel statistical models (4th ed.)*. Chichester, UK: Wiley.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York, NY: Holt-Saunders International editions.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966-968. doi: 10.1126/science.1152408

Knuth, D. E. (1992). *Literate programming*. Center for the Study of Language and Information, Stanford University.

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. doi: 10.1177/0013164406294779

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

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*, 8. doi: 10.1186/s40536-014-0008-1

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., … Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-3000)*. Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195070019.001.0001

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412-433. doi: 10.1037/met0000144

Ratcliff, R. (1978). Theory of memory retrieval. *Psychological Review*, *85*, 59-108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R., Smith, P. L., & McKoon, G. (2015). Modeling regularities in response time and accuracy data with the diffusion model. *Current Directions in Psychology Science*, *24*, 458-470. doi: 10.1177/0963721415596228

Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. *Cognitive Psychology*, *103*, 1-22. doi: 10.1016/j.cogpsych.2018.02.002

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological assessment*, *31*, 13951411. doi: 10.1037/pas0000754

Rigden, J. S. (2005). *Einstein 1905: The standard of greatness*. Cambridge, MA: Harvard University Press.

Rios, J. A. (2022). When should individual ability estimates be reported if rapid guessing is present? *Applied Measurement in Education*, *35*, 222-236. doi: 10.1080/08957347.2022.2103138

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses meta-analysis. *Large-scale Assessments in Education*, *9*, 18. doi: 10.1186/ s40536-021-00110-8

Rios, J. A., & Deng, J. (2024). A comparison of response time threshold scoring procedures in mitigating bias from rapid guessing behavior. *Educational and Psychological Measurement*, *84*, 387-420. doi: 10.1177/00131644231168398

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden: Nieuwsbrief voor Toegepaste Statistiek en Operationele Research*, *7*(22), 131-145.

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test- taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. doi: 10.1016/j.edurev.2020.100335

Sinharay, S. (2021). Latent-variable approaches utilizing both item scores and response times to detect test fraud. *Open Education Studies*, *3*(1), 1–16. doi: 10.1515/edu-2020-0137

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-scale Assessments in Education*, *9*, 8. doi: 10.1186/s40536-021-00100-w

Steger, D., Schroeders, U., & Wilhelm, O. (2021). Caught in the act: Predicting cheat- ing in unproctored knowledge assessment. *Assessment*, *28*, 1004-1017. doi: 10.1177/1073191120914970

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). **PerFit**: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1-27. doi: 10.18637/jss.v074.i05

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thomp- son (Ed.), *Score reliability* (p. 3-30). Thousand Oaks, CA: Sage.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287-308.

van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, *53*, 334-358.

Voss, A., Lerche, V., Mertens, U., & Voss, J. (2019). Sequential sampling models with variable boundaries and non-normal noise: A comparison of six models. *Psychonomic Bulletin & Review*, *26*, 813-832. doi: 10.3758/s13423-018-1560-4

Weiss, N. A. (2016). **wBoot**: Bootstrap methods [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=wBoot (R package version 1.0.3)

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and im- plications. *Educational Measurement: Issues and Practice*, *36*, 52-61. doi: 10.1111/ emip.12165

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a cat item pool: The normative threshold method. Vancouver, Canada: National Council on Measurement in Education.

Wolff, S. M., Hatcher, W. J., & Wright, D. B. (2024). Task-irrelevant visual distractions and mindful self-regulated learning in a low-stakes computer-based assessment. *Frontiers of Education*, *9*. doi: 10.3389/feduc.2024.1360848

Wright, D. B. (2016). Treating All Rapid Responses as Errors (TARRE) improves estimates of ability (slightly). *Psychological Test and Assessment Modeling*, *58*, 15-31.

Wright, D. B. (2019a). Speed gaps: Exploring differences in response latencies among groups. *Educational Measurement: Issues and Practice*, *38* (4), 87-98. doi: 10.1111/ emip.12286

Wright, D. B. (2019b). Treating rapid responses as incorrect for non-timed formative tests. *Open Education Studies*, *1*, 56-72. doi: 10.1515/edu-2019-0004

Wright, D. B. (2024). Normrank correlations for testing associations and for use in latent variable models. *Open Education Studies*, *6* (20240003), 1-18. doi: 10.1515/edu-2024-0003

Wright, D. B., & Herrington, J. A. (2011). Problematic standard errors and confidence intervals for skewness and kurtosis. *Behavior Research Methods*, *43*, 8-17. doi: 10.3758/ s13428-010-0044-x

Wright, D. B., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, *62*, 439-456. doi: 10.1348/ 000711008X327632

Wright, D. B., London, K., & Field, A. P. (2011). Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology*, *2*, 252-270. doi: 10.5127/jep.013611

Xie, Y. (2015). *Dynamic documents with R and* ***knitr*** (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.