Research Article

Daniel B. Wright*

Treating Rapid Responses as Incorrect for Non-Timed Formative Tests

https://doi.org/10.1515/edu-2019-0004
Received Nov 04, 2018; accepted Jun 21, 2019

Abstract: When students respond rapidly to an item during an assessment, it suggests that they may have guessed. Guessing adds error to ability estimates. Treating rapid responses as incorrect answers increases the accuracy of ability estimates for timed high-stakes summative tests like the ACT. There are fewer reasons to guess rapidly in non-timed formative tests, like those used as part of many personalized learning systems. Data from approximately 75 thousand formative assessments, from 777 students at two northern California charter high schools, were analyzed. The accuracy of ability estimates is only slightly improved by treating responses made in less than five seconds as incorrect responses. Simulations show that the advantage is related to: whether guesses are made rapidly, the amount of time required for thoughtful responses, the number of response alternatives, and the preponderance of guessing. An R function is presented to implement this procedure. Consequences of using this procedure are discussed.

Keywords: response times, ability, personalized learning

How quickly a student solves an academic task provides information about the student's response strategy. Rapid responding suggests little cognitive effort has been used (Wise, 2017). Luce (1986) provides a detailed account of research in cognitive science laboratories up until the mid-1980s. De Boeck and Jeon (2019), Kyllonen and Zu (2016) and Ratcliff, Smith, and McKoon (2015) provide more recent reviews of this literature.

Academic assessments are very different from these laboratory tasks. A goal of many modern rigorous assessments is to tap deep knowledge, requiring students to use deep levels of processing (Craik & Lockhart, 1972). Reaching a response may require that the students go through several thoughtful steps. High-stakes summative tests like the SAT and ACT allow students about one minute to answer each question, on average. In high-stakes timed tests like these, students are encouraged to use response strategies to increase their scores. For the SAT and ACT this includes guessing for students who run out of time:

> Your score on the test will be based only on the number of questions that you answer correctly; there is no penalty for guessing. Try to answer every question within the time allowed for each test. (ACT Inc, 2018, p. 4)

Wise and colleagues (e.g., Wise & DeMars, 2006; Wise & Kong, 2005) describe measures for saying a student has not spent enough time to be judged to have expended sufficient cognitive effort. They argue that this can occur in tests where there is little motivation for students to perform well. These students can score well below their true ability. This can have ramifications for teachers and schools because while some assessments are low-stakes for students, they can be high-stakes for teachers and schools.

It is likely that if somebody expends little cognitive effort throughout a test that the person's overall score will be low. For this paper it is assumed that students are not rewarded for lack of effort and therefore this low mark is appropriate. But what happens if somebody rapidly guesses on just a couple of items? Some of these guesses will likely be correct and some will likely be incorrect. If the responses are true guesses then these chance events add error to the ability estimates. Wright (2016) showed for high-stakes ACT Math data that treating all rapid responses as errors (TARRE) increased the accuracy of the ability estimates. He used ten seconds for the threshold for saying that test-takers had or had not expended sufficient cognitive effort on that item. The goal of the current research is to test if TARRE also improves ability estimates for non-timed formative assessments and to explore this across several areas. The TARRE procedure is summarized as follows:

1. Define threshold (e.g., quicker than 5 seconds, quicker than 95% of responses for that item).
2. If a response time is less than the threshold, treat as an incorrect response even if it was answered correctly.
3. Aggregate responses as normal.

***Corresponding author: Daniel B. Wright,** Dunn Family Endowed Chair and Professor of Educational Assessment at the University of Nevada, Las Vegas, E-mail: daniel.wright at unlv.edu or dbrookswr at gmail.com

An R function that implements the algorithm is presented in the Appendix B with examples.

Why Study Formative Assessments

Students take many formative assessments and in some school systems these outnumber summative assessments. During the past few decades it has become common for students to use computer systems to help them to learn part of their curriculum (for critical historical discussions, see Cuban, 2001; Ferster, 2014; Wright, 2018). This allows students some control over the content and the pace of their learning. This is often called self-regulated or personalized learning (e.g., Arney, 2015; Bjork, Dunlosky, & Kornell, 2013; Horn & Staker, 2015; Murphy, Redding, & Twyman, 2016; Panadero, 2017). Students typically study information in a module and then take a brief assessment to estimate how well they know the module's content. A key aspect of personalized learning is that students should only progress to the next module if they have mastered the current module. Therefore estimating knowledge well is important, though the consequences of these scores are not as high-stakes for the individual test-taker as for most summative tests.

These personalized learning assessments are different in many ways from tests like the SAT and ACT. The SAT and ACT are summative tests that are influential for college admissions and scholarships; are fixed-time tests; are composed of items that have been field tested with hundreds of thousands of students; take in total several hours; and are taken under tightly controlled settings. Each of these aspects is different from the typical circumstances when a student is doing a personalized learning assessment.

There are several reasons why a student might rapidly respond in a personalized learning context. First, the question may tap superficial and rapidly accessible knowledge. If the item is: "What is the Spanish word for *tree*?" and the first option is *árbol*, a student with knowledge of Spanish might rapidly answer this without even looking at the other alternatives. This type of question should be fairly rare on most tests as item developers are encouraged to write questions that tap more in-depth knowledge in order to assess whether the student has mastered the material (Herman & Linn, 2014). The second reason is similar to what happens in timed high-stakes tests when the student runs out of time and guesses because there is no penalty. In non-timed formative tests students will have less incentive to do this, but they may want to rush off to a class or lunch or whatever else the student believes is more important than this assessment. Third, some students may quickly glance at an item and based on this initial impression believe that they will not be able to figure out the answer, and so they guess. Ideally formative tests should be designed to entice students to think about even difficult items beyond this initial impression. Thoughtful incorrect responses can be valuable for learning (Metcalfe, 2017).

1 Study 1: Empirical Test of TARRE with Formative Assessments

1.1 Methods

Data were gathered for one year from two Northern California Bay Area Charter high schools. The personalized learning system used at these schools is called Summit Learning and is described in detail at `https://www.summitlearning.org/`. The assessments are divided into six subject areas: History (26% of the assessments), World Languages (13%), Science (22%), English (18%), Mathematics (21%), and Expeditions ($\ll$ 1%). The Expedition assessments are not examined here. Table 1 shows some of the characteristics of the sample. Demographic differences in response times, called *speed gaps*, are discussed in Wright (in press).

Students have several hours each day available to spend on this system. This means that each student takes many of these assessments for each of the topic areas and therefore there are a large number of assessments. There are data for 74,804 of these assessments. The personalized learning system reports a score out of ten to the student. Providing the score is at least eight, the student decides whether to progress to further modules. This threshold was chosen by the software developers; the consequences of this choice are discussed later in this paper. If the score is less than eight correct, the student is not given this option and must re-do the module. Only twenty nine percent of assessments were taken just once, so for most assessments the student has already been assessed on the module.

The items for each assessment are randomly chosen from evolving test banks that vary in size by module. The items were created by the same educators who developed the content in the modules, not professional item developers. The item bank is large enough so that most items will not have been previously seen by students who take the assessment multiple times, but some may. All items are included in these analyses as all are used in the system when estimating mastering. The items are four-alternative multiple choice questions (i.e., the test-taker chooses among four alternative answers for each item). This format, as

Tab. 1: Percentages for some of the demographics characteristics of the 777 students.

School	A = 54%	B = 46%		
Gender	Females = 44%	Males = 56%		
Grade	9th = 25%	10th = 27%	11th = 25%	12th = 24%
Ethnicity	Asian = 26%	Black = 2%	Hispanic = 37%	
	White = 17%	Two or more races = 17%		
American Indian, Alaska or Hawaii Native, or other Pacific Islander < 1%				

opposed to free response, is used to ensure immediate and accurate feedback for students. The response time is recorded by the amount of time spent in total on each item, combining times if the student comes back to an item. The responses are recorded in integer values. A very small number (0.08%) of times in the data set registered as negative, 0, or 1 second time. After consultation with the engineers at Summit Learning it was determined that these were recording errors and these times are treated as missing. This is a very small percentage, but having any errors detected suggests that there may be others. It is expected, if so, that the percentage would be small. The software was written by engineers from a well-known technology company (Facebook) and it is being used by more than 380 schools in the US.

1.2 Analytic Plan

The assessments considered here have ten items and the items are randomly chosen from pools of items for each module. There are several ways to estimate ability, but for transparency purposes the online personalized learning system (Summit Learning) uses the number of correct responses (other methods are shown in Appendix B). Here the sum of the number of correct responses will be compared with and without treating all responses less than different thresholds as incorrect responses. For example, if the threshold is 7 seconds, all the responses faster than 7 seconds will be treated as incorrect responses and therefore not contribute to the sum of correct responses.

Following Wright (2016), leave-out-one-item cross-validation will be used to compare TARRE and the traditional methods. This involves using the sum of nine of the ten items to predict the probability of the tenth one being answered correctly. The following logistic regression is used:

$$logit\left(\widehat{Pr_i}\right) = ln\left(\frac{\widehat{Pr_i}}{1-\widehat{Pr_i}}\right) = \beta_0 + \beta_1 sum_i \qquad (1)$$

where the probabilities are assumed to follow a binomial distribution. The procedure is as follows:

1. For each of the 10 positions in which an item can appear:
   (a) Traditional:
      i. Sum the number of correct responses of the remaining nine items.
      ii. Use this sum to predict the response on the chosen item using a logistic regression (alternatives exits, for example, using a probit regression or a more flexible curve).
      iii. Record the fit of the model.
   (b) TARRE:
      i. For the remaining nine items, change any values with an associated response time that is less than the threshold to an incorrect response.
   ii., iii., iv. Repeat i., ii., iii. from above.
2. Compare the fit of the matched pairs of models

There are several ways to compare the fit of logistic regressions. Many of these take into account the number of people in the sample and the model complexity. One set of approaches[1] are information criteria (IC):

$$IC = -2 \cdot log\text{-}likelihood + k \cdot npar \quad , \qquad (2)$$

where *log-likelihood* is the log-likelihood, *npar* is the number of parameters in the model, and k is a penalty invoked for different complexities. Popular values for this penalty are $k = 2$ for An (or Akaike's) Information Criterion (AIC) and $k = ln(n)$ for the Bayesian Information Criterion (BIC). Discussion of the differences between these is in Hastie, Tibshirani, and Friedman (2009). Because the models being compared here use the same number of students and each just uses a single predictor variable, different values of k, and therefore different ICs based on eqn. 2, will lead to the same conclusions as will using the log-likelihood value itself. The log-likelihood statistic from each logistic regression is used here. Higher values for this mean the model fits better. If the log-likelihood of the model using TARRE

1 The equations are often presented differently, for example multiplying these by the sample size. The equation shown here is based on the R function `AIC`.

minus the log-likelihood for the traditional procedure is positive, this shows a TARRE advantage. If the difference is negative it is a TARRE disadvantage.

1.3 Results

1.3.1 Univariate Time and Accuracy

Approximately 1% of the response times (7,856 of 747,451 times) were longer than ten minutes. Figure 1 shows that the log of the response times is approximately normally distributed, though with differences at the extremes. Some of the differences at the low end may relate to the discrete measurement of the data and rapid guessing. Differences at the high end may be related to students doing an unrelated task (e.g., going to the bathroom). For the purposes of this paper response time is dichotomized at the threshold so very long response times are not treated differently from those lasting only a minute.

The proportion accurate for each subject area, overall, were:

History:	130,717 correct of 191,990, or 68%,
World Language:	64,912 correct of 98,710, or 66%,
Science:	108,602 correct of 163,560, or 66%,
English:	88,071 correct of 134,150, or 66%,
Mathematics:	98,861 correct of 157,020, or 63%.

Thus, the mathematics assessments have the lowest accuracy.

Each individual student's assessment is made from ten different items that are randomly chosen from item banks that continue to evolve. Because these are only ten items long the aggregate measures are less reliable than if they were longer. The 95% confidence interval for getting 8 of 10 correct using Wilson's method, as recommended in (Agresti, 2002, p. 16), is 4.90 to 9.43, so quite wide. Cronbach's α, calculated across all assessments using item position to demarcate the different items, was .56. This would be considered low for many purposes and is in part because of having only ten items. Using the Spearman-Brown prophecy formula, if using twenty of these items the predicted α would be .72, and with fifty items it would be .87. Having about ten items is common for formative tests, but is less than most summative tests.

1.3.2 Relation between Accuracy and Response Time

Figure 2 shows the proportion correct for each of the five subject areas for each amount of time. The proportion accurate increases from extremely rapid responses to a plateau for most subject areas at about 7–8 seconds. The accuracy for mathematics items continues to increase until about 15 seconds. The accuracy for the mathematics items is lower than for the other subjects for each of the responses less than about 25s.

1.3.3 Comparing TARRE and the Traditional Methods

Analyses were done separately for each of the five subjects areas (History, World Languages, Science, English, and Mathematics). Consistency across these five sets of comparisons is used to judge whether TARRE consistently provides more accurate estimates of ability for these formative assessments. Following the procedures in the Analytic Plan, the TARRE advantages were calculated for each subject area for each possible threshold from 3 seconds to 10 seconds.[2] The mean log-likelihood value for each of these sets of ten was found and the mean difference between the TARRE method and the traditional method is shown in Table 2. Positive values, highlighted in yellow, show when the TARRE estimates had on average higher log-likelihood values and therefore were more accurate estimates of ability (a TARRE advantage). To assess the consistency of this advantage within each set of ten, paired t-tests were conducted. All those with t values above 2.262 (the t associated with an upper-tail in the probability distribution of approximately .025 for $df = 9$) are in *italics*. If a single cut-off were required for all tests, having it at 4 or 5 seconds seems appropriate to maximize accuracy. This choice is discussed further in the General Discussion.

1.3.4 How Many Test-Takers does this affect?

Table 3 shows the number of tests that have scores of eight or more (i.e., "passing") depending on the cut-off used. The change in percentage is fairly small, but if these students only reach this level, which is assumed to denote mastery, by guessing, having them re-do the module may be in their best interest. If five seconds is used as the threshold, about 1% more students would be required to re-do the module. If students learn that they are not rewarded for rapid guesses it is hoped that they will take more time and make more thoughtful responses. Even if

2 Analyses were conducted with thresholds up to 30 seconds, and also for treating long responses as incorrect responses. For all of these additional evaluations, the traditional procedure had a better fit than the TARRE procedure. These analyses are available from the author.

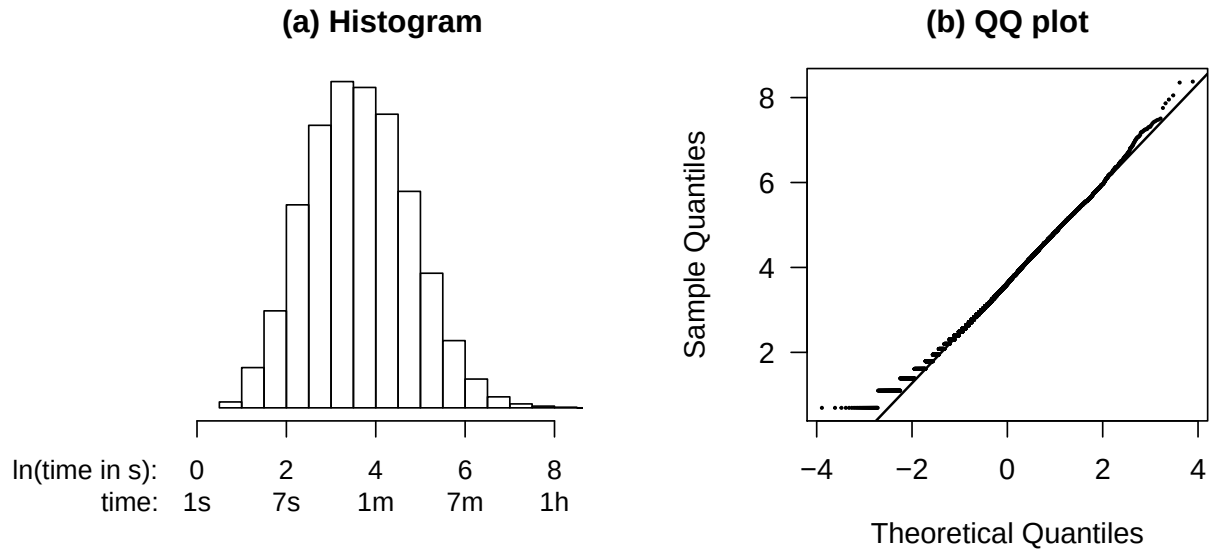

Fig. 1: Distribution of the logged response time. Panel (a) shows the histogram and panel (b) the quantile-quantile plot with the normal distribution line superimposed (10,000 of the points are shown).

Tab. 2: The mean difference in log-likelihood R^2 for the TARRE approach minus the mean for the traditional approach for the different thresholds. Highlighted cells show a TARRE advantage and those which are also statistically significant at $p < .025$ are shown in italics.

	History	Languages	Science	English	Math
< 3s	*4.19*	1.07	*2.59*	*5.02*	*4.19*
< 4s	*10.27*	*3.30*	*8.41*	*13.20*	*10.62*
< 5s	*7.03*	-3.31	*7.70*	*10.22*	*11.31*
< 6s	-11.57	-18.78	-2.26	-1.18	*8.30*
< 7s	-36.60	-44.73	-20.75	-17.05	1.05
< 8s	-67.59	-72.75	-44.09	-34.40	-6.26
< 9s	-97.77	-102.89	-66.06	-53.70	-14.54
< 10s	-128.08	-130.86	-88.20	-73.00	-28.10

Tab. 3: The number of tests with scores less than eight or greater than/equal to eight, and this percentage, for the different thresholds for treating responses as incorrect.

	# < 8	# ≥ 8	% ≥ 8
None	46,843	27,961	37.38
< 3s	46,881	27,923	37.33
< 4s	47,148	27,656	36.97
< 5s	47,863	26,941	36.02
< 6s	48,987	25,817	34.51
< 7s	50,501	24,303	32.49
< 8s	52,059	22,745	30.41
< 9s	53,634	21,170	28.30
< 10s	55,127	19,677	26.30

the student is incorrect, thinking about the problem can help them to learn (Metcalfe, 2017). Ideally any changes would stop students just glancing at items and believing that they cannot figure out the answer, and entice them to spend some time thinking about the item if only for a few extra seconds.

2 Study 2: Simulation of TARRE

Simulation methods are used to explore when the TARRE advantage is likely to occur. Simulation methods allow insight into how the data might have arisen and to test alternative what-if questions (Feinberg & Rubright, 2016; Gentle, 2009).

The first step with a statistical simulation is to decide which probability distributions to use. The assumption here is that the response time distribution is a mixture of at least two distributions. One is a log-normal distribution (alternatives to this, like the Weibull distribution, could also be used, see Palmer, Horowitz, Torralba, and Wolfe, 2011) for the thoughtful responses. A second distribution is for rapid guesses. With the ACT data (Wang & Hanson, 2005; Wright, 2016) exploratory analyses revealed a clear bi-modal distribution with about 7% of the sample being rapid responses around five seconds, fewer at around 6–9 seconds and then the main log-normal distribution. The distribution for the data here (Figure 1b) has more low (and high) times than predicted by the log-normal distri-

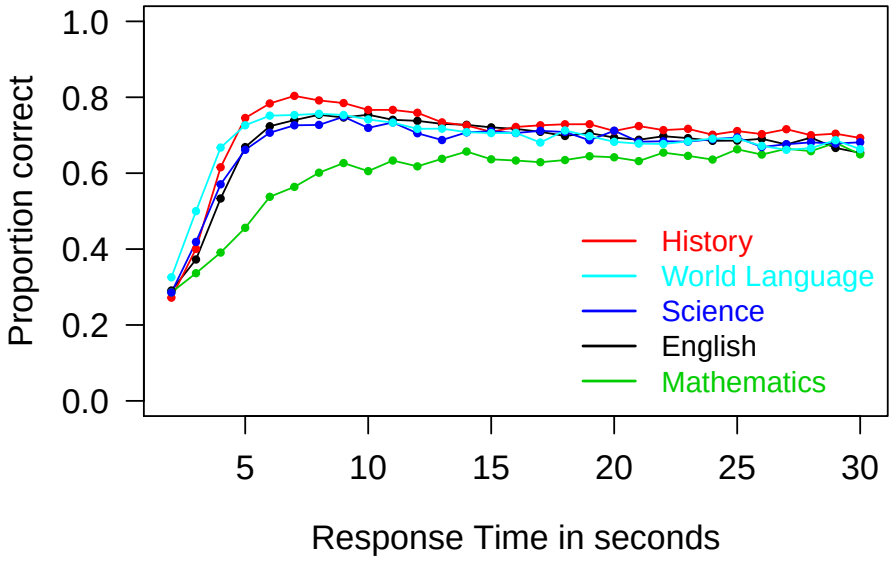

Fig. 2: The relationship between the proportion correct for each subject area with the response time up to 30s.

bution, but because of the discrete nature of these data it is difficult to see if the distribution is bi-modal. Further, because the rapid responses constitute only a small proportion in this data set, analytic procedures to detect mixtures of distributions (e.g., those procedures in **mixtools**, Benaglia, Chauveau, Hunter, and Young, 2009) did not identify a separate distribution for these rapid responses. A distribution was created to be similar to those in Figure 1 by combining a log-normal distribution and a uniform distribution. The excess of slow times compared with the log-normal is not included here because there was not a drop in accuracy for these so they do not appear to be guesses. Further, very slow times are not treated different from those just above the threshold for the current analyses.

2.1 Methods

Four variables are examined in this simulation. First, as noted earlier, a suggested strategy in high-stakes fixed-time tests that do not penalize incorrect answers is for students to guess rapidly, but for a non-timed test students have less incentive to guess rapidly. Two uniform distributions used to simulate guessing behavior are $RT \sim U[1, 30]$ for the non-timed version and $RT \sim U[1, 10]$ for the fixed-timed version. Second, the items on tests can require different amounts of thought. Those that require deep knowledge (rigor) may average twice the amount of time that items only requiring superficial, easily accessible, knowledge. Two log-normal distributions will be examined. Their means are 30 and 60 seconds, which correspond to ln(30) = 3.40 and ln(60) = 4.09 for the logged distributions. The standard deviations will be 1 and ln(60)/ln(30) = 1.20, so that the ratio of the mean to the standard deviation remains constant for the logged distributions. Occasionally (about 3 in every 10,000 trials) these will result in a value less than one second or a negative log. The absolute values of these logs were taken so that both the thoughtful responding distribution and the guessing distribution have the same theoretical minimum (1 second). Third, the number of alternatives presented to students can vary. The values of 2 (which corresponds with a TRUE/FALSE item), 4 (as typical with many multiple choice items including those considered in Study 1), and 8 (this is an arbitrary larger number so that the likelihood of guessing accurately is low, but not zero) are used. Finally, how often students guess is varied from 0% to 50% in increments of 5%. Thus, there were $2 \times 2 \times 3 \times 11 = 132$ different conditions. There were 5,000 replications for each of these, so 660,000 trials in total.

A number of variables are not examined in this simulation. For example, item variability is not taken into ac-

count nor is that student ability will likely be related to guessing likelihood. Also, 10 item assessment with 1000 students are fixed for this simulation. Using around ten items is common with formative assessments, but is less than used with most summative assessments. Only the very basic method of treating responses that are less than ten seconds as incorrect and a similarly basic method of summing correct responses to estimate accuracy are used. More involved methods are shown in Appendix B.

The expectation is that the TARRE advantage will be most pronounced when there is a high likelihood of accurate rapid guesses and that non-guesses will be slower so not often treated as incorrect. This corresponds with being encouraged to guess rapidly (the $U[1, 10]$ condition), having few thoughtful responses near the threshold so that these are not removed (the log-normal for thoughtful responding with the higher mean), having few response alternatives (the advantage should be largest when there are only two alternatives), and having lots of guessing (here 50% guessing is the maximum). Because this is a simulation the true ability of each student is known. Ability is based on a random uniformly distributed variable from .01 to .99. Item difficulty is the square root of a random variable drawn from a uniform .01 to .99 distribution. The TARRE advantage will be measured by correlating student true ability with the estimates from the TARRE procedure and the traditional procedure. The difference between these (un-transformed) values is stored and the mean of them reported.

The code for this simulation is in the Appendix A and at `https://github.com/dbrookswr/tarre` (the file SimulationStudy2.r).

2.2 Results and Discussion

Figure 3 shows the relationship of the correlation between true ability (known since this is a simulation) with ability estimates using the TARRE procedure and true ability with estimates from the traditional procedure. Most of the data points are near the diagonal, showing that the two procedures usually yield similar levels of accuracy. However, when they differ the TARRE procedure usually performs better as evident from the data points above and to left of the red diagonal line. Of the 4.44% trials where there is more than a .1 difference between the two correlations, all have the TARRE correlation higher than the traditional correlation. Thus adopting this procedure has little risk of producing substantially worse estimates for the conditions examined.

The main effects for the TARRE advantage were as predicted and are shown in Figure 4. The units of the effects are differences in Pearson correlations. When the mean time of the thoughtful responses was short (30 seconds) the advantage was only 0.0029, but when it was long (60 seconds) the advantage was 0.0122. This was the smallest main effect for any of the four variables manipulated. If the differences between these distributions are made larger, the corresponding differences in the TARRE advantage become larger (e.g., 10 seconds versus 60 seconds). When the rapid guesses ranged from $U[1, 10]$ the mean TARRE advantage was 0.0193 and there was a TARRE disadvantage (-0.0042) when it was $U[1, 30]$. The number of alternatives had the predicted effect, with the TARRE advantage largest when there were only two alternatives (0.0284), next when four alternatives (0.0010), and an an overall TARRE disadvantage when there were 8 alternatives (-0.0068). When the preponderance of guessing was 0% there was a TARRE disadvantage (-0.0143). This stayed negative until guessing was above 25%, and reached a TARRE advantage of 0.0487 when the guessing preponderance was 50%. When all these main effects were included in the model to predict the difference, the mean for the thoughtful responding distribution had a p-value of .03. The ps associated with the other three variables were all $p \ll .001$.

3 General Discussion

Rapid responses on formative tests, like those examined in Study 1, have near chance levels of accuracy. The length of time taken means students are not expending much cognitive effort on these questions and the accuracy rates suggest that many of these responses are guesses. What can and should be done?

It is worth differentiating summative and formative assessments. The following are three priorities for each of these types of assessment.

Summative
1. Accurately measure ability.
2. Require rigor/assess mastery.
3. Thoughtful responding.

Formative
1. Thoughtful responding.
2. Require rigor/assess mastery.
3. Accurately measure ability.

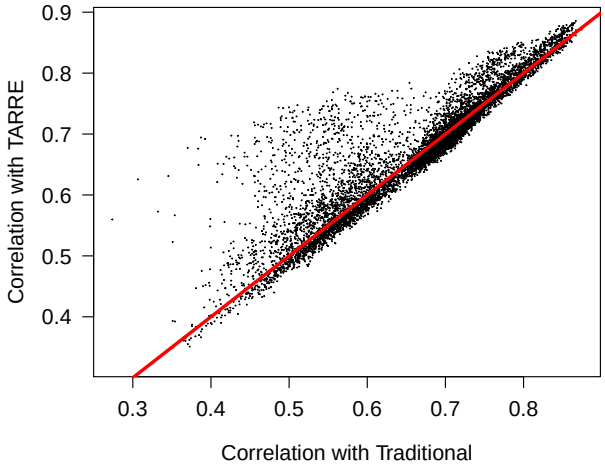

Fig. 3: Scatter plot of the correlations between true ability and the estimates from the TARRE procedure with the correlations between true ability and the estimates for the traditional procedure. Ten thousand points shown. The red line shows the diagonal where the correlations are the same.

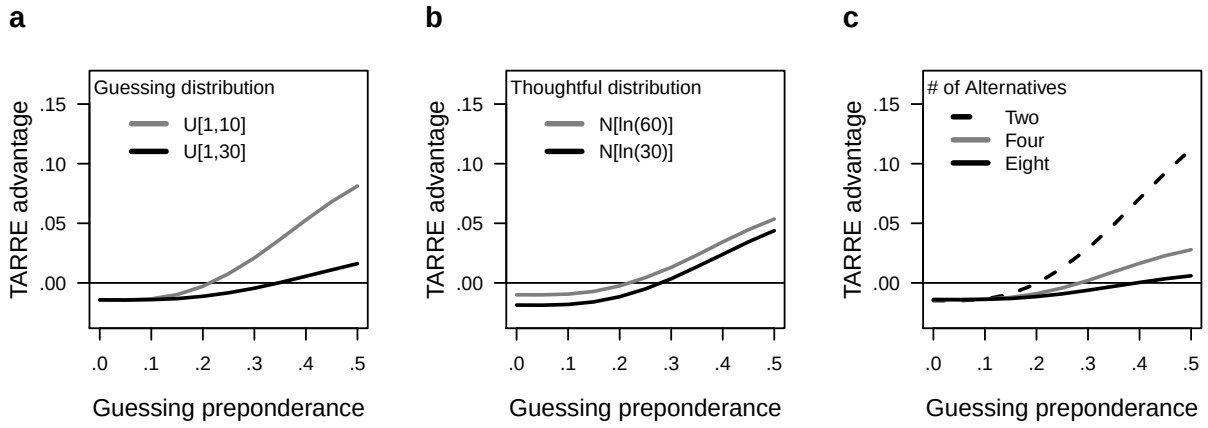

Fig. 4: The differences between the correlation between true ability and the estimates of the TARRE procedure and the estimates of the traditional procedure, with the amount of guessing. The main effects for the guessing distribution, the thoughtful responding distributions, and the number of alternative are shown in panels a, b, and c.

Notice the priorities are they same; all three of these are valuable for both types of assessment. But the priorities are in opposite orders for the two types of assessment.

For summative tests accurately measuring ability–or whichever construct is being measured–is the primary goal. Because TARRE improves ability estimates of timed summative tests (Wright, 2016), the main reasons not to implement this for these tests is if it is too complicated to explain to students and, for very high-stakes tests, any negative consequences resulting from how test preparation companies would adapt to the change. It is important to recognize that there are several methods to make scores more accurate (e.g., adding more items, field testing items in order to remove weak ones), and it is best consider these together.

For formative tests the priorities are in the opposite order. The goal of these assessments is to help students to learn. Thoughtful responding during assessments is one of the key aspects of the testing effect (Roediger & Karpicke, 2006). This is particularly important for schools that use a lot of personalize learning and formative assessments. Accurate ability estimation is worth pursuing, but any changes should also improve student learning. Suppose that the administrators were to change their policy

so that all rapid responses were treated as incorrect. Students would likely be told this. This might negatively affect students wanting to answer the *tree-árbol* example rapidly, but arguably item developers may wish to tap less superficial knowledge anyway. The second group of fast responders–those wanting to finish the assessment and get to recess–would likely accept not getting credit for accurate guesses since they have already prioritized recess over the assessment. The third group are those who glance at the item and believe that they will not be able to answer the question and therefore guess. Using TARRE should encourage this group to spend more time on these questions and hopefully to think more about them.

Table 2 suggests thresholds from 4–7 seconds would improve estimates for the data in Study 1. The optimal threshold will vary by assessment and perhaps by sample. The choice of threshold should be consistent across test type to avoid confusing test-takers (assuming that they are informed about the procedure). It is important to consider how students change their behavior and evaluate how this affects ability estimation.

It is worth considering how these formative assessment scores are used. One goal of many personalized learning systems is that students should demonstrate *mastery* at each level before progressing to the next level. For the present discussion, assume that the true state of *mastery* is binary; somebody either has mastered the content or has not. No assessment can perfectly demarcate these states; there will be errors where the assessment says someone has mastered the material when the person has not (called a false alarm) and errors where the assessment says the person has not mastered the material, when in fact the person has. These are shown in Table 4.

For ten-item tests, like used in the data here, the percentage of different types of errors will depend on how high the threshold is for declaring if a student has passed. If the threshold is low, say getting only 5 of 10 correct, the percentage of times incorrectly declaring someone has mastered the concept (a false alarm) will be higher than if the threshold is 9 of 10 correct. But with this high threshold, many students, who have mastered the concepts, will fail (misses). The costs of these different types of errors should be considered (Swets, Dawes, & Monahan, 2000). False alarms in an educational context mean students advance to part of the curriculum that they are not prepared for. If the false alarm rate is high the system should include ways to check for fundamental knowledge that may be missing (i.e., to catch these errors subsequently) and require these students to learn this material (e.g., revert to these previous "passed" levels). If the percentage of misses is high (equivalently if the hit rate is low), students may get discouraged with the system and be spending too much time on simple tasks without being challenged with more rigorous material. The relative costs of these will vary with context.

Tab. 4: Different types of outcomes for an assessment. Hits and correct rejections are correct outcomes, false alarms and misses are errors.

		True State	
		Not Mastered	Mastered
Test says	Not passed	Correct Rejection	Miss
	Passed	False Alarm	Hit

The hits, misses, correct rejections, and false alarms can be combined to create several measures of the diagnostic value of the assessment for that sample. People often discuss the hit rate (HR) and the false alarm rate (FAR). HR is the number of hits divided by the total number of students who have mastered the material, i.e., HR = hits/(hits + correct rejections). FAR is the number of false alarms divided by total number of students who have not mastered the material, i.e., FAR = false alarms/(false alarms + misses). In medical contexts the HR is often called *sensitivity* and one minus the FAR called the *specificity*. This area is called signal detection theory and a common statistic is called *d'*. It is the *z* score associated with HR minus the *z* score associated with the FAR: $d' = z_{HR} - z_{FAR}$. d' values above zero mean ratings are more accurate than guessing and negative d' values means ratings are less accurate than chance. The hope is that the d' values are high, at least above 1.

A popular visual plot is the receiver operating characteristic (ROC) plot. ROCs show the hit rate with the false alarm rate. Because researchers often have only a small number of HR and FAR pairs, some assumptions about the decision process are made and ROCs plotted in an idealized form. Figure 5 is an example. Two ROCs for two d' statistics are shown. These show what the HR and FAR would be depending on the threshold that the assessment developers choose to say someone has passed. Suppose that they say the highest acceptable FAR is 50%. They can use this to decide the threshold and this would show them the expected HR. For the ROC with $d' = 1.5$ this is .93 and for the $d' = 1.0$ this is .84. Conceptually the choice of the passing threshold, for the tests considered in Study 1: 8 out of 10, involves moving along a single ROC. With this conceptualization, the choice between TARRE and the traditional method would involve moving from one ROC to another. The TARRE method is more diagnostic (a higher d').

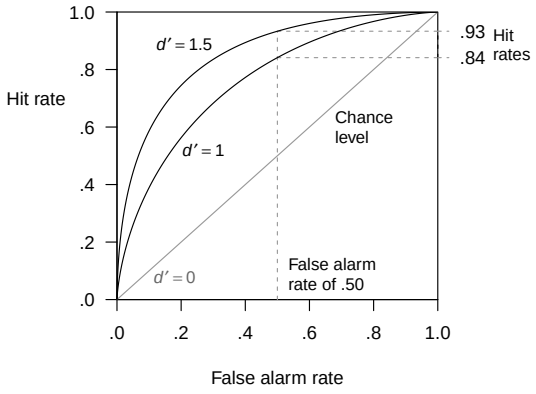

Fig. 5: The receiver operating characteristics (ROCs) for two tests. One has $d' = 1.5$ (the one that is better able to assess student mastery) and one has $d' = 1.0$ (assesses mastery better than chance, $d' = 0$, but not as well as the other test). If chosing the threshold so that the false alarm rate is .50, the ROC with $d' = 1.5$ has a higher hit rate (.93) that does the ROC with $d' = 1.0$ (.84).

If keeping the same threshold (and assuming these idealized ROC forms are correct) this would slightly raise the HR and slightly lower the FAR.

3.1 An Alternative

Consider a more radical proposal: allow students to replace a question they believe that they do not know the answer of with three (other numbers could also be used) new questions, but the students have to answer all of these replacement questions correctly to get credit for the original one. This would require students to think about their metacognitive state regarding the question, one of the key 21st century skills (Griffin & Care, 2015) and could prevent guessing. They would also have to think about the chances of answering the replacement questions correctly. This type of thinking, trying to imagine oneself as a thinking machine and predicting outcomes, is central to the skills Minsky (2019) argues are critical for student growth. Clark (2016) argues that we are, in essence, prediction machines. The number of replacement items (e.g., two or three), how to choose replacement questions (e.g., they should be from the same content area, but this could be difficult if the question bank is not large), and whether students would be limited in how often they could use this option would need to be decided by test administrators and after further testing.

Another group of people interested in ability estimates from assessments are researchers conducting studies on how ability is related to other variables. For them, measurement accuracy is of the utmost importance. Any improvements in ability estimates are welcomed, but if the effects are small the value of the improved estimates might not be worth over-complicating the results sections of articles. At present many articles still use the sum of correct responses even though estimates from IRT (Mair, 2018) could be more accurate. Authors, editors, and reviewers might not feel the psychometric advantages of these more complex procedures are worth explaining to their audience for non-methodological papers. If researchers wish to explain TARRE to their audiences, they could use the leave-out-one item cross-validation procedure shown with Study 1 to test if the ability estimates are improved for their data set, and then use TARRE if appropriate. Further research on TARRE should be conducted and if the findings continue to show improved estimates in different situations then this method should be widely recommended in research contexts.

All research has limitations. It is important to stress that Study 1 was conducted with students from two high schools that are not typical. They use personalized learning systems more than is typical and the students are from a part of the United States (silicon valley) where computer use–at almost all ages–is very common. Further, these formative assessments were all produced by one organization, and they are continuing to develop the test batteries and how they are administered. Study 2 also has limitations. While some of the key factors were manipulated, it is important that future studies examine others, the covariation among these, and estimate what values for these parameters are associated with different testing conditions. The covariation between true ability and likelihood of rapid guessing is particularly important. In complex latent variable models, like van der Linden's (2011) hierarchical model, the covariation between these, and other latent constructs, is estimated. It is important to stress that the format of this assessment is very different than that used for Wright (2016), and yet for both treating rapid responses as incorrect improved the ability estimates. The R function and the use of leave-out-one cross-validation allows researchers to check for their own data to see if TARRE improves their estimates. It is encouraged that people do this and report any TARRE advantages and disadvantages for different thresholds.

Response times are a window into the cognitive processes of the student (Luce, 1986). Beyond helping to estimate ability, the response times can tell students and their teachers about how they are answering questions. If teachers know which questions students spend the most time on, this can be helpful for their lesson plans. Response times are one of the most readily available forms of *para-*

data in online tests and are available for each item from many software systems. They are a valuable resource that should be used. Finding that treating rapid responses as incorrect can improve ability estimates is one way in which this resource can be used.

Acknowledgment: Thanks to Connie Choi, Adam Carter, and others at Summit Learning for access to the data and discussion. Thanks to Kristin Smith Alvarez and Heather Kirkpatrick for discussion. This research is supported by the Chan Zuckerberg Initiative. Some aspects of this study were reported at the Developmental Methods Conference, Whitefish, MT, US, in September 2018.

References

ACT Inc. (2018). *Preparing for the ACT test*. Iowa City, IA: ACT, Inc. Retrieved from https://www.act.org/content/dam/act/unsecured/documents/Preparing-for-the-ACT.pdf

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley Interscience.

Arney, L. (2015). *Go blended! A handbook for blending technology in schools*. San Francisco, CA: Jossey-Bass.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using **lme4**. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6), 1–29. Retrieved from http://www.jstatsoft.org/v32/i06/

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. doi: 10.1146/annurev-psych-113011-143823

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York, NY: Oxford University Press.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, *11*, 671–684. doi: 10.1016/S0022-5371(72)80001-X

Cuban, L. (2001). *Oversold & underused: Computers in the classroom*. Cambridge, MA: Harvard University Press.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*, 102. doi: 10.3389/fpsyg.2019.00102

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*, 36–49.

Ferster, B. (2014). *Teaching machines: Learning from the intersection of education and technology*. Baltimore, MD: Johns Hopkins Press.

Fox, J.-P., Klotzke, K., & Entink, R. K. (2019). **LNIRT**: Lognormal response time item response theory models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=LNIRT (R package version 0.4.0)

Gentle, J. E. (2009). *Computational statistics*. New York, NY: Springer.

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

Griffin, P., & Care, E. (Eds.). (2015). *Assessment and teaching of 21st century skills: Methods and approach*. New York, NY: Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Herman, J., & Linn, R. (2014). New assessments, new rigor. *Educational Leadership*, *71*, 34–37.

Horn, M. B., & Staker, H. (2015). *Blended: Using disruptive innovation to improve schools*. San Francisco, CA: Jossey-Bass.

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(14). doi: 10.3390/jintelligence4040014

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

Mair, P. (2018). *Modern psychometrics with R*. Cham, Switzerland: Springer.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. Retrieved from 10.1146/annurev-psych-010416-044022

Minsky, M. (2019). *Inventive minds: Marvin Minsky on education*. Cambridge, MA: The MIT Press.

Murphy, M., Redding, S., & Twyman, J. S. (Eds.). (2016). *Handbook on personalized learning for states, districts, and schools*. Charlotte, NC: Information Age Publishing, Inc.

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 58–71. doi: 10.1037/a0020747

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2017.00422 doi: 10.3389/fpsyg.2017.00422

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

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210. doi: 10.1111/j.1745-6916.2006.00012.x

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26. doi: 10.1111/1529-1006.001

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

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wickham, H. (2015). *Advanced R*. Boca Raton, FL: CRC Press.

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

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*, *18*, 163–183.

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

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. (2018). A framework for research on education with technology. *Frontiers in Education*, *3*. Retrieved from `https://www.frontiersin.org/article/10.3389/feduc.2018.00021` doi: 10.3389/feduc.2018.00021

Wright, D. B. (in press). Speed gaps: Exploring differences in response latencies among groups. *Educational Measurement: Issues and Practice*.

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

Appendix A: Code for the Simulation

The following is the code used for the simulation. It is also available at https://github.com/dbrookswr/tarre (the file SimulationStudy2.r). If you adapt this code for your own research please email the author.

```
set.seed(1643)
replics <- 5000
n <- 1000
k <- 10
mu <- c(log(30),log(60))
#sd included here in paras
unifs <- c(10,30)
nafc <- c(2,4,8)
guess <- seq(.0,.5,.05)

paras <- as.data.frame(matrix(ncol=9,
  nrow=replics*length(mu)*length(nafc)*length(guess)*length(unifs)))
paras[,1] <- rep(1:(length(mu)*length(nafc)*length(guess)*length(unifs)),
                 replics)
paras[,2] <- rep(1:replics,each=nrow(paras)/replics)
paras[,3] <- rep(rep(mu,replics),each=nrow(paras)/(replics*length(mu)))
paras[,4] <- paras[,3]/log(30)
paras[,5] <- rep(rep(unifs,length(mu)*replics),
   each=nrow(paras)/(replics*length(mu)*length(unifs)))
paras[,6] <- rep(rep(nafc,replics*length(mu)*length(unifs)),
   each=nrow(paras)/(replics*length(mu)*length(nafc)*length(unifs)))
paras[,7] <- rep(rep(guess,replics*length(nafc)*length(mu)*length(unifs)),
    each=nrow(paras)/(replics*length(mu)*length(nafc)*
                        length(guess)*length(unifs)))
colnames(paras) <- c("Trial","replic","mu","sd","unifs","nafc",
                     "guess","rtarre","rtrad")

for (i in 1:nrow(paras)) {
  ability <- runif(n,.01,.99)
  itemdiff <- sqrt(runif(k,.01,.99))
  probright <- (matrix(rep((ability),k),ncol=k) +
                matrix(rep((itemdiff),n),byrow=TRUE,ncol=k))/2
  rt <- exp(matrix(rnorm(prod(dim(probright)),
              paras$mu[i],paras$sd[i]),ncol=ncol(probright)))
  rt[probright < paras$guess[i]] <-
           runif(sum(probright < paras$guess[i]),1,paras$unifs[i])
  probright[probright < paras$guess[i]] <- 1/paras$nafc[i]
  probright[probright < 1/paras$nafc[i]] <- 1/paras$nafc[i]
  correct <- matrix(rbinom(n*k,1,probright),ncol=k)

  trad <- rowSums(correct)
  correctTARRE <- correct
  correctTARRE[rt<5] <- 0
  tarre <- rowSums(correctTARRE)
  paras[i,8:9] <- c(cor(tarre,ability),cor(trad,ability))
}
```

Appendix B. Examples using `tarre` with R

The `tarre` function uses R's functional programming capabilities (Wickham, 2015). The function itself is succinct:

```
tarre <- function(x,respt,fabil,ftarre,...){
  newx <- ftarre(x,respt,...)
  abil <- fabil(newx,respt,...)
  return(abil)}
```

This is a minimal function that allows much flexibility because the user can input their own functions to define how rapid responses are treated and the scoring algorithm. The flexibility is at the cost that users will need some R skills. This is why the examples are provided here. This function and examples are located on GitHub at `https://github.com/dbrookswr/tarre`. Readers are encouraged to add more examples.

The user needs to create functions for deciding which responses to treat as incorrect (the `ftarre` slot above) and how to create the ability estimates (`fabil`). Both of these functions can use the response accuracy matrix `x` and the response time matrix `respt`, and the four examples below illustrate this. The assumption is that the columns for these will refer to the item numbers, but this is only a requirement if made so by the user's functions. Other parameters can also be passed to these functions because of the *ellipses* (the `...`) in the `tarre` function, and this is shown in Example #2 below. If there is a particular approach that is common in your organization you can make this approach the default for your version of `tarre` (if you do this on GitHub, please create a new function with a different name).

For ease, the same toy data set will be used for each example. `x` is the response accuracy matrix (1 correct, 0 incorrect) and `respt` are the response times in seconds. In this example, as response time increases the probability of being accurate increases, and then at the median the probability of being accurate plateaus. These data are designed only to illustrate the function, not to mimic any particular situation. This toy data set was created with the following:

```
set.seed(1984)
x <- matrix(nrow=100,ncol=10)
ability <- matrix(runif(100,0,.4)[row(x)],ncol=10)
nability <- apply(ability,1,mean)
LT <- abs(rnorm(1000,60,20))+2
probacc <- (LT<median(LT))*LT/105 +
  (LT>=median(LT))*median(LT)/105 + c(ability)
x <- matrix(rbinom(1000,1,probacc),ncol=10)
respt <- matrix(LT,ncol=10)
```

These examples are included to illustrate the versatility of the function. Their inclusion is not an endorsement.

Example #1

This example uses a single threshold for guessing and returns the student sums for the new response matrix. This is the method used with the high school student data analyzed in Study 1 and would be easily understood by students if it were implemented. Because both the accuracy and the response time matrices are named in `ftarre` and `fabil` without defaults, it is important when writing these that either both are included as input variables or `...` is used to show further parameters could be included. The second approach is done in the `sumall` function below. Just a single variable of the ability estimates are returned. The user can return more complex objects (e.g., multiple ability estimates per student); it is whatever is returned by the user-defined `fabil` function.

```
ltthresh <- function(x,respt,thresh=10)
  {x[respt<thresh] <- 0; return(x)}
sumall <- function(x,...) rowSums(x,na.rm=TRUE)
```

```
ex1 <- tarre(x,respt,sumall,ltthresh)
```

Example #2

The ACT data, discussed earlier, were fit using a single threshold and the ability estimates, the θs, from item response theory (IRT) models (Wright, 2016). Here the two-parameter logistic (2PL) model is estimated using the **mirt** package (Chalmers, 2012). The `mirt` function requires that the data are in a data frame. The defaults (other than minimizing output) are used for this function. The `fscores` function estimates a single latent variable; the `1` in the `mirt` function call means only a single latent variable is estimated per student. The **mirt** package can estimate many different IRT models and any of these can be used with the `tarre` function. A different threshold, 15 seconds, is passed to the `ltthresh` function to show how this can be done.

```
suppressPackageStartupMessages(library(mirt))
ltthresh <- function(x,respt,thresh=10)
    {x[respt<thresh] <- 0; return(x)}
irtabil <- function(x,...)
  fscores(mirt(as.data.frame(x),1,itemtype="2PL",verbose=FALSE,
               technical=list(message=FALSE)))
ex2 <- tarre(x,respt,irtabil,ltthresh,thresh=15)
```

Example #3

Because the time required for a thoughtful response will vary by item (Wise, 2017), it can be worth having different thresholds for each item. For this example the threshold is defined by whether the response is faster than 10% of the times for that item. Wise and Ma (2012) refer to this as the normative threshold method. With 100 students in this toy example and R's default method for calculating quantiles, the quickest ten responses would be treated as incorrect. This requires that that the cells of the response accuracy and response time matrices refer to the same unique trials. If students see different items, as with adaptive testing, more columns would be needed and missing values would need to be dealt with by the user defined functions. The ability estimates are calculated here using van der Linden's (2011) hierarchical model as implemented in the **LNIRT** package (Fox, Klotzke, & Entink, 2019). This package is in version 0.4 and therefore some of the syntax may change. The function `LNIRT` assumes the response times have been logged. If the user wanted only the slowest 5% to be treated as errors, `tarre(x,respt,vdlabil,botperc,bottom=.05)` could be used. There are lots of parameters that could also be passed to the `LNIRT` function.

```
suppressPackageStartupMessages(library(LNIRT))
botperc <- function(x,respt,bottom=.1){
  x[respt <  matrix(apply(respt,2,quantile,bottom),
                    ncol=10,nrow=100,byrow=TRUE)] <- 0
  return(x)}
vdlabil <- function(x,respt)
  return(LNIRT(log(respt),x)$Post.Means$Person.Ability)
ex3 <- tarre(x,respt,vdlabil,botperc)
```

Example #4

This example is more complex and is used to illustrate how more involved functions can be used with the `tarre` function. It is worth noting that the usefulness of the succinct `tarre` function is mostly conceptual here, stressing the separation of the TARRE and ability estimation processes, than for making the computations simple. Anyone using it for an ex-

ample this complex an example would need good knowledge of R. A brief review of the model used in this example is warranted.

The diffusion model (Ratcliff, 1978) is the most extensively studied model within cognitive psychology that incorporates response time and response accuracy. It has been valuable for understanding how people make simple decisions. It can also be applied to more complex decisions, like how students answer questions. There are several extensions to the basic model. Here, for this toy problem, the *EZ-diffusion* model (Wagenmakers, van der Maas, & Grasman, 2007) will be used because of its computational ease. It is available in Wagenmakers et al.'s appendix and at: `http://www.ejwagenmakers.com/2007/EZ.R`. There are no plans in the foreseeable future to remove this web page (Wagenmakers, personal communication). The function is called `get.vaTer`. Because within cognitive psychology the diffusion model has been developed for simple decisions, here the `ftarre` function will treat long response times as incorrect. This might be appropriate in an educational context if it was believed that long delays suggest students may be looking up answers on the internet. The user function is called `slow`. Slow response times will be based on the residuals, the e_{ijk}, of variance component model that includes random variables for the student ($u1_j$) and item ($u2_k$). This is often called a multilevel or cross-classified model (Goldstein, 2011; Wright & London, 2009).

$$ln(RT_{ijk}) = \beta 0 + u1_j + u2_k + e_{ijk} \tag{3}$$

This means that an item is deemed slow based on both other students' times on the item and how slow the time is for the particular student. This can be estimated using the `lmer` function from the **lme4** package (Bates, Mächler, Bolker, & Walker, 2015). Responses with e_{ijk} values above one standard deviation are treated as incorrect. This requires restructuring the data into a "long" format, as is necessary for many multilevel functions.

The input for the `get.vaTer` function requires the percentage correct (`Pc`), the mean response time for correct answers (`MRT`) and the variance of response times for correct decisions (`VRT`) for each student. `Pc` values of 0, .5, and 1 result in errors so are changed in the code below. With the typical cognitive psychology laboratory study these values do not often occur because most responses are accurate and each subject may take part in hundreds of trials so that there will usually be some errors. For the response times, the un-transformed values are used. The function calculates several parameters for each person, the first of which is the drift rate, and this is what the code below returns. The drift rates are used to estimate each student's ability.

```
source("http://www.ejwagenmakers.com/2007/EZ.R")
suppressPackageStartupMessages(library(lme4))

slow <- function(x,respt){
 RTs <- c(respt)
 studID <- rep(1:nrow(respt),ncol(respt))
 itemID <- rep(1:ncol(respt),each=nrow(respt))
 rs <- matrix(resid(lmer(log(RTs)~1 + (1|studID) + (1|itemID))),ncol=10)
 x[rs > 1*sd(c(rs))] <- 0
 return(x)}

EZabil <- function(x,respt){
  Pc <- apply(x,1,mean,na.rm=TRUE)
  Pc[Pc==0] <- .05
  Pc[Pc==1] <- .95
  Pc[Pc==.5] <- rbinom(sum(Pc==.5),1,.5)/10 + .45
  RTright <- respt
  RTright[x==0] <- NA
  MRT <- apply(RTright,1,mean,na.rm=TRUE)
  VRT <- apply(RTright,1,var,na.rm=TRUE)
  EZabil <- vector(length=length(Pc))
  for (i in 1:length(Pc))
```

```
    EZabil[i] <- get.vaTer(Pc[i],MRT[i],VRT[i])[[1]]
  return(EZabil)
  }
ex4 <- tarre(x,respt,EZabil,slow)
```

Table 5 shows the correlations among the estimated abilities from these different procedures and the true ability, known because the data were simulated and the data were the same for all examples. Because there were only ten items, none of the methods provide very accurate estimates.

	Ex. 1	Ex. 2	Ex. 3	Ex. 4
Ex. 2	0.69			
Ex. 3	0.85	0.73		
Ex. 4	0.85	0.58	0.68	
True Ability	0.65	0.53	0.52	0.60

Tab. 5: Correlations among the ability estimates from the different R examples. The examples use the following methods for treating responses as incorrect responses: less than ten and fifteen seconds (#1 and #2), less than 90% (#3), and with a residual greater than one standard deviations. The ability estimates were found by summing, the 2PL IRT model, van der Linden's hierarchical model, and the EZ diffusion model. True ability is known since the data are simulated.