Psychological Test and Assessment Modeling, Volume 58, 2016 (1), 15-30

Treating all rapid responses as errors (TARRE) improves estimates of ability (slightly)

Daniel B. Wright[1]

Abstract

Response times can be modeled along with response accuracy to estimate ability. Models that do not use response times were compared with three models that do. The predictive accuracy of the models were assessed using leave-out-one-item cross-validation where for a k item test each method is used k times with $k-1$ items to create ability estimates and these estimates are used to predict responses on the remaining item. The conceptually simplest method using response times, which treats all rapid responses as errors (TARRE), produced the most predictive values. However, the increase was less than would be achieved by having one extra item on the test. Possible effects of changing the scoring algorithm on student test taking behavior need to be explored before implementing any such a change.

Keywords: Response times; Measurement; IRT

[1] *Correspondence concerning this article should be addressed to:* Daniel B. Wright, PhD, Public Education Department, 300 Don Gasper Ave., Santa Fe, NM, 37501, USA; email: dbrookswr@gmail.com

I was always deeply uncertain about my own intellectual capacity; I thought I was unintelligent. And it was true that I was, and still am, rather slow. I need time to seize things ...
(Schwartz, 2001, p. 30)

A young Frenchman, Laurent Schwartz enjoyed mathematics. While he got good grades, it took him a long time to understand what his mathematics teachers were saying. The relationship between speed and ability is much researched within cognitive science. For within-subject comparisons there is a speed-accuracy trade-off (Luce, 1986) whereby if individuals speed up their responding to individual items, this lowers their accuracy on these items. Comparisons between different people are more difficult. While it is clear that if one person rapidly answers "20" to the prompt 4×5 and another person counts fingers and toes, that the first person has been using a more advanced arithmetic skill, the case of Schwartz highlights that speed is not necessary for mathematics greatness. The Fields Medal (often described as the Nobel Prize for mathematics) winner argued "being quick or slow isn't really relevant" (p. 31). The goal of this paper is to examine if response speed can help to assess ability.

Many standardized tests are now administered on computer, allowing response times to be recorded. While there are yet-to-be-resolved issues about these values (e.g., what happens when students go back and forth between questions?, are the values equally reliable on tablets and desktop computers?), testing organizations are now considering what should be done with these values. It is agreed that response times provide information about the cognitive processes occurring while responding (Luce, 1986), and therefore they are being used for test development and research. More controversial is whether the response times should be used to estimate ability and therefore be part of the algorithms used to generate test scores.

When students take a multiple choice standardized test, it is common to treat the data as 0s (incorrect) and 1s (correct) and to use item response models to estimate ability, usually denoted θ.[1] These are estimated; they are measured with error. Because of this, if response times also are in part related to true ability, then they could lessen this measurement error. Novick and Jackson (1974) described this as *collateral information*. For example, someone with very low ability may rapidly guess and get a response correct by chance. Incorporating response time into the estimation of the ability estimates could improve the estimation.

Using response times to estimate θ could affect students' response strategies. Given that for high-stakes tests like the ACT® and SAT®, students often take courses to

[1] Item response models have also been adapted for partial credit and other response types.

improve their scores, it is likely that if testing organizations penalized certain responses based on how quickly the responses were made, then these courses would take this into account. This could affect students' strategies and have unintended consequences. These consequences should be examined prior to implementing any change to the scoring the algorithm and are discussed further at the end of this paper.

Models examined

Item response theory

Item response theory (also called item response models and latent trait models) is often used to estimate the ability of students by assuming the probability of a correct response follows a specific function of the students' ability and characteristics of the items. Let θ_i be the ability of the ith student. Multiple θs are sometimes used if the test is measuring multiple constructs, though here only uni-dimensional person models will be considered. Common uni-dimensional models are referred to as the 1PL–4PL models for the number of item parameters in the logistic model (for an introductory book see Embretson & Reise, 2000; for introductory articles see Harris, 1989; Wright & Skagerberg, 2006; Zicker, 1998; for relating these to other latent variable models see Chapter 4 of Bartholomew, Knott & Moustaki, 2011). Different notations are often used for these models. Here the notation is based on Embretson and Reise (2000, p. 71). For example, the 3PL equation is:

$$P(correct_{ij}) = c_j + (1 - c_j)\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad , \tag{1}$$

where a_j is the discriminability (sometimes a scalar is included), b_j is the difficulty, and c_j is a guessing parameter, all for the jth item. For the 2PL, $c_j = 0$, and for the 1PL, a_j is the same for all items (and $c_j = 0$). The 4PL includes a ceiling parameter that is the highest probability correct for someone with the highest ability (the c_j is the same but at the lowest ability levels). The 4PL is seldom used in educational assessment since as $\theta \to \infty$ the probability of being correct should be very near 1 if the item is well-written.[2]

Several packages can estimate IRT models. Here the R package **mirt** (Chalmers, 2012) is used. This package allows many alternatives and options. Unless otherwise mentioned, the default settings for the 1PL–4PL models are used. One useful alternative for the 3PL

[2] A ceiling threshold is better suited for other applications. For example, modeling winning at games that involve chance (e.g., poker) would require both floor and ceiling boundaries. If educational assessments are based on results from so-called *serious games* then the 4PL might become more widely used if the games include chance aspects.

model is to set the floor values to a constant. Setting the floor parameter to $1/k$ for a k-alternative forced choice test is often used.

Treating all rapid responses as errors (TARRE)

If somebody randomly guesses the correct answer this gives a false impression of the person's ability. Therefore, if a test taker spends less time than is required for a thoughtful response then this arguably is a guess. The adjustment here is that all rapid responses are assigned zero (incorrect) for the IRT ability estimation (though not for cross-validation when assessing predictive accuracy). The threshold will depend on the type of questions and sample of test takers so the choice of threshold will need to be addressed for different circumstances.

Here, the choice of less than 10 seconds as a threshold for declaring a response as too quick for demonstrating proper cognitive processes was informed by both of knowledge of ACT mathematics questions and the empirical distributions. The lowest quartile was calculated for all 20 items, and the lowest was 19 seconds. Since 9.5 seconds would round down to 9 seconds, it means that items on this test were deemed lacking appropriate cognitive effort if the test taker took less than half of the lower quartile of time for the item with the lowest quartile. Wise and Kong (2005) discuss methods for detecting low-effort responding in more detail. They consider taking into account test taker and item characteristics. Thus, rapid answers to a question like 5×4 would not be treated as an error for older students who likely have memorized the *times tables*, but a rapid response to a lengthy mathematics word problem would be. There is an advantage having a single threshold, however, as this could be conveyed to test takers more easily.

An advantage of the TARRE approach is that the same analyses that would normally be done using techniques that do not incorporate response time can still be done in a straight-forward manner and therefore the standard diagnostic output is available.

Diffusion models

Diffusion models have long been used in physics to account for Brownian motion, but have recently become popular in cognitive science to account for the joint distribution of response time and accuracy in decision making. Ratcliff's (1978) diffusion model is the most widely evaluated and it has been adapted by many. It assumes information either for or against the correct response is sampled (it is an example of a sequential sampling method). A test taker with high ability would be sampling more information

leading to the correct response than to an incorrect response; the test taker *drifts* towards the correct response. Once the amount of information has reached a threshold for responding either with the correct and an incorrect response, the test taker makes this response. An important aspect of these models is that they are meant as descriptions of the cognitive processes occurring when the person responds; the parameters have psychological meaning.

Within these models the *drift* rate is the gradient towards a correct response and thus this parameter is the most similar to ability from IRT (van der Maas *et al*., 2011). Alternative models, for example the linear ballistic accelerator models, assume the test taker is sampling from a multinomial distribution for each possible response (Brown & Heathcote, 2008). Adaptions like forcing the ability estimates to be positive have also been examined (van der Maas *et al*., 2011).

Most of the cognitive science research with diffusion models has been with relatively small numbers of subjects answering a large number of items that are very similar to each other (or differ based on an experimental factor). With educational testing the number of test takers can be very large and the items purposefully differ in many ways. Many of the algorithms are prohibitively slow with the sample sizes used in standardized testing. Wagenmakers, van der Maas and Grasman (2007) produced an algorithm for a simplified model they call the *EZ* diffusion model. Research comparing it with other diffusion model adaptions has shown the *EZ* model works quite well (van Ravenswaaij & Oberauer, 2009). The drift rate, v, for this model is shown in eqn. 2 (Wagenmakers *et al*., 2007, eqn. 7). P_c is the proportion correct, VRT is the variance of the response times, and s is a measure of the amount of information accumulated at each step and here is set to .1, which is the default for Wagenmakers *et al*. function. Because of its simplicity it can be used with large samples without creating computational difficulties.

$$v = sign\left(P_c - \frac{1}{2}\right) s \left\{ \frac{\text{logit}(P_c)\left[P_c^2 \text{logit}(P_c) - P_c \text{logit}(P_c) + Pc - \frac{1}{2}\right]}{VRT} \right\}^{\frac{1}{4}} \qquad (2)$$

An R function, adapted from http://ejwagenmakers.com/2007/EZ.R (accessed 20 January 2016), is shown below. Because P_c can equal 0 or 1 in sample data (and the quantile logistic function below would yield $-\infty$ and $+\infty$, respectively), the P_c values are multiplied by .98 and .01 added to the result (their function returns these as errors, which is not practical for educational assessments). This has the effect of shrinking values towards .5. The function takes matrices of response times (`tt`) and responses (`rr`) and returns the estimated drift rate.

```
estv <- function(tt,rr){
    s <- .1
    Pc <- apply(rr,1,mean)*.98 + .01
    VRT <- apply(tt,1,var)
    L <- qlogis(Pc)
    x <- L*(L*Pc^2 - L*Pc + Pc - 0.5)/VRT
    sign(Pc-0.5)*s*x^(1/4)}
```

van der Linden's Hierarchical Model

van der Linden (2007, 2011) described modeling the response accuracy and the response latency separately, but allowing the latent variables used in the separate models to be associated. He describes this as a hierarchical model because information about the response accuracy and latency are informed by some higher level relationships. I will follow his lead (van der Linden, 2011, pp. 341–342) and refer to level 1 and level 2 parts of his model. Level 1 has two models: one for the response times and one for the responses themselves. For the response times a log-normal model is used (an alternative being a Box-Cox normal model or a log with a starting value normal model) such that:

$$log(RT_{ij}) = \beta_j - s_i + \varepsilon_{ij}, \text{ with } \varepsilon_{ij} \sim N(0, {\alpha_j}^{-2}) \quad . \tag{3}$$

The model for the probability of the response being correct is an IRT model. Here, the 2PL is used:

$$P(correct_{ij}) = \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}} \quad . \tag{4}$$

Level 2 allows associations between the latent variables. The covariance matrices, Σ_p and Σ_l, describe the associations among these parameters, and the mean vectors, μ_p and , μ_l allow differences among people and items.

The model, with 2PL (other IRT models could also be used), can also be depicted as a directed acyclic graph (i.e., a DAG) as in Figure 1.

Fox et al.'s (2007) **cirt** package estimates the parameters of van der Linden's model. It allows several options. The response part of the model can be estimated using a 2PL or a 3PL. The time part of the model can be estimated using both latent variables or having the discrimination variable, α_j, constant. The number of iterations was set to 5000 as recommended in the package's help pages.

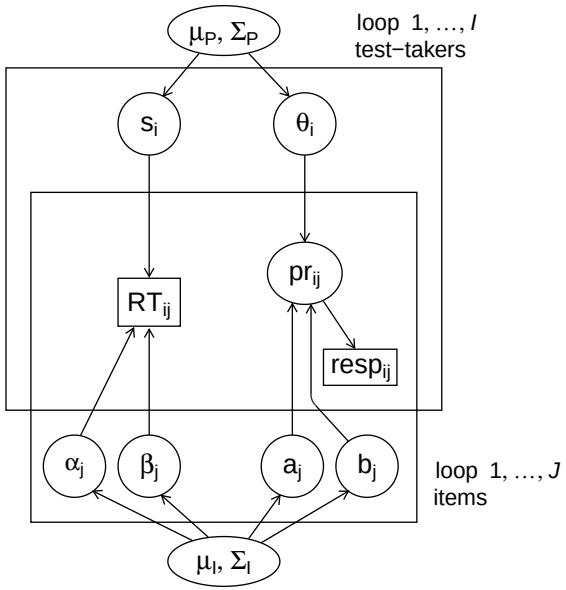

Figure 1: The van der Linden (2011) hierarchical model with 2PL for the response model and 2 item parameters for modeling response time. The response time model is shown on the left. The response accuracy model is shown on the right.

Methods for comparing models

Several other models have been proposed that incorporate response time and accuracy. Luce (1986) provides a thorough review of those up to 1986. Since then there has been a more concerted effort within education. Examples relevant to education include Thissen (1983) and Wang and Hanson (2005). Lee and Chen (2011) review this literature.

All of these models examined attempt to estimate an unknown person level latent variable that can be interpreted as the person's ability. Both TARRE and van der Linden's hierarchical models build directly upon the standard IRT models so are closely related and are psychometric models not designed to represent the underlying cognitive processes. The diffusion models have been developed and are often interpreted as models of the cognitive processes. However, the studies usually used within the cognitive science

studies involve items that are designed to require only a single cognitive process. The typical item on a standardized test requires several cognitive processes. The focus here is just to compare the predictive accuracy of the methods, rather than to assess if the models are accurate descriptions of the underlying processes.

There are two methods that have been used within education to compare models that use response times with ones that do not. The first is showing that the models yield similar estimates with real data. For example, Wang and Hanson (2005) showed that their model produced similar estimates to the 3PL with real test data (multiple-choice mathematics questions from the ACT). The second method involves simulation. Wang and Hanson (2005) simulated data and compared the $\widehat{\theta}$ from the two models with the true θ used to create the data. They found the $\widehat{\theta}$ from the 3PL was correlated .884 with θ, but the $\widehat{\theta}$ from their model was correlated .937. Similarly, van der Linden, Klein Entink and Fox (2010) used simulation methods to show that the van der Linden model improved the estimates of θ.

A difficulty with simulations is that the data need to be created in a way consistent with how they arise in real testing situations. Widiatmo and Wright (2005) created data that were either consistent with a diffusion model or with the van der Linden model. They found that the estimates of ability were more accurate than a 2PL when the data were consistent with the statistical model used to estimate θ, but were less accurate than the 2PL when they were created in a manner inconsistent with the statistical model. Therefore, it is important to examine the model fit with data consistent with real test data. The problem is that with real data the true abilities are unknown.

Here the data reported in Wang and Hanson (2005) are analyzed in a manner that allows the predictive accuracy of the models to be compared. It is important to observe how well a model from training data can predict new data (Hastie, Tibshirani & Friedman, 2009). Leave-out-one cross-validation (CV) is a popular technique where the models are fit for all but one of the *n cases*. The model is then used to predict the remaining case.[3] This is repeated for all cases and then an aggregate predictive value from all n estimates is used as a measure of the predictive value of the approach. Here the models are estimated for all but one of the *items*, the ability estimated for each test taker, and then these ability estimates are used to predict responses on the remaining item using a logistic regression.

McFadden's pseudo-R^2 is used to estimate the fit for all the items. Finding the mean of R^2 values is complicated by them being bounded by 0 and 1 (all the relationships reported in this paper were positive). Here, their square-roots are taken, Fisher's z

[3] Five-fold and ten-fold CV are also common and have some advantages over leave-out-one CV (Hastie *et al.*, 2009).

transformation applied, the mean found, then the mean z value is back-transformed into an R value, which is squared. This has the advantage that the final values are in a well-known metric.

The main research question is how treating all rapid responses as errors (TARRE) and then applying IRT affects the accuracy of ability estimates compared with IRT and with two existing methods that incorporate response times. An auxiliary question is if there is an improvement using one of the response time methods, is the improvement enough to change scoring algorithms. Finally, *leave-out-one-item* cross-validation is used to compare the accuracy of the different methods and it is argued this provides a useful way to measure these differences.

Methods and results

The data used here are those used and discussed in Wang and Hanson (2005). They are responses and response times from 20 ACT multiple-choice mathematics questions from 1161 test takers. The response times are all rounded to the nearest integer in seconds. The code for all analyses is available from the author.

Comparing IRT models

The 1PL–4PL models are compared using leave-out-one item CV. In addition, a 3PL fixing all of the guessing parameters at .25 is included. The package **mirt** (Chalmers, 2012) was used for all estimations. The defaults for the function `mirt` were used except that the number of EM cycles was increased to 10,000 so that all solutions converged. The `fscores` function by default uses the expected *a posteriori* method (Chalmers, 2012).

Table 1 shows the McFadden pseudo-R^2s for each model. The bottom row shows the means using Fisher's transformation and then back-transformed into R^2 scale. To three digits the 2PL, the 3PL with the guessing value fixed at .25, and the unconstrained 3PL provide the same level of accuracy. Table 2 shows that the only two models that fit significantly poorer than others are the 1PL and 4PL (the 4PL over-fits the training data). For simplicity, the 2PL is used in the next section for comparison.

	1PL	2PL	3.25PL	3PL	4PL
1	0.186	0.196	0.195	0.195	0.192
2	0.171	0.179	0.181	0.181	0.182
3	0.168	0.178	0.180	0.180	0.179
4	0.232	0.237	0.241	0.241	0.239
5	0.135	0.136	0.132	0.132	0.131
6	0.101	0.107	0.109	0.109	0.109
7	0.118	0.121	0.121	0.121	0.120
8	0.202	0.208	0.212	0.212	0.211
9	0.282	0.293	0.296	0.296	0.297
10	0.229	0.241	0.245	0.245	0.243
11	0.124	0.125	0.123	0.123	0.121
12	0.076	0.078	0.074	0.074	0.075
13	0.034	0.035	0.036	0.036	0.036
14	0.083	0.083	0.080	0.080	0.081
15	0.109	0.107	0.106	0.106	0.105
16	0.092	0.094	0.097	0.097	0.097
17	0.270	0.272	0.270	0.270	0.268
18	0.131	0.128	0.130	0.130	0.127
19	0.197	0.195	0.193	0.193	0.191
20	0.076	0.074	0.072	0.072	0.071
Column mean	0.146	0.149	0.149	0.149	0.148

Table 1: IRT models. McFadden's pseudo-R^2 for each method with 19 items predicting the remaining item. The mean was calculating with Fisher's transformation. 3.25PL is the 3PL model with the guessing parameter fixed at .25.

	1PL	2PL	3.25PL	3PL	4PL
1PL		0.004	0.020	0.020	0.064
2PL	0.003		0.558	0.560	0.526
3.25PL	0.004	0.000		0.412	0.009
3PL	0.004	0.000	-0.000		0.009
4PL	0.003	-0.000	-0.001	-0.001	

Table 2: The mean differences (lower triangle) and p values (upper triangle) comparing each of the IRT models. Positive values in the lower triangle mean the row method performed better than the column method. 3.25PL is the 3PL model with the guessing parameter fixed at .25.

Comparing 2PL to approaches that use response times

$\widehat{\theta}$ from the 2PL is compared with $\widehat{\theta}$ from the 2PL that treats all rapid responses (operationalized as less than 10 seconds) as errors (TARRE), the drift rate from the *EZ* diffusion model, and $\widehat{\theta}$ from van der Linden's hierarchical model. van der Linden's model is estimated twice, once allowing each item to have its own response time discrimination value (α_j from Figure 1) and once having these equal for all items. In the Tables these are referred to as vdl and vdltm1 respectively.

Table 3 shows that the model that treats all quick responses as errors has the highest predictive value of all the models. The lowest predictive value is for the drift rate variable of the *EZ* diffusion model. Table 4 shows the predictive value for the *EZ* model was statistically significantly below all other models and the model treating all responses less than 10 seconds as incorrect was statistically significantly above all other models.

The predictive accuracy of the ability estimates was improved with TARRE compared with the 2PL procedure. However, the increase was small: McFadden's pseudo-R^2 increased by less than .001. In order to interpret a shift of this magnitude it is worth examining what the shift is for including an extra item. The leave-out-one-item CV was repeated for all sets of 19 items using both the 2PL and the TARRE procedures. The mean R^2 was 0.14719 for 2PL and 0.14796 for TARRE. The TARRE value is higher, but is below the 2PL value for all 20 items, which was 0.14878. It is about halfway between these values. Therefore, a larger increase in predictive accuracy could be accomplished in many ways including adding an extra item.

Discussion

There are many issues to consider prior to recommending a fairly large change in scoring algorithms. The main consideration is whether the improvement in accuracy outweighs any potentially negative consequences of the change. In addition, it is necessary to decide if there are other ways to improve the estimation of ability that are more cost effective.

Although response time can be recorded during the computerized tests without affecting the test taker, if response times are used to estimate ability for individuals on high-stakes tests, this could affect test taking strategy. This could increase or decrease the predictive accuracy of the scores and the fairness of the tests, and therefore this will need to be investigated. If the change is to treat all rapid responses as erroneous, this would have the benefit on timed tests that there would not be an advantage (nor would there be a disadvantage, other than the time taken to guess on several items might allow a

	2PL	TARRE	EZdiff	vdl	vdltm1
1	0.1964	0.2007	0.1889	0.2022	0.1996
2	0.1786	0.1783	0.1674	0.1792	0.1790
3	0.1780	0.1777	0.1711	0.1780	0.1776
4	0.2368	0.2367	0.2189	0.2341	0.2351
5	0.1359	0.1338	0.1187	0.1350	0.1351
6	0.1074	0.1071	0.0853	0.1057	0.1062
7	0.1212	0.1225	0.1028	0.1197	0.1207
8	0.2076	0.2095	0.1944	0.2075	0.2069
9	0.2928	0.2927	0.2690	0.2919	0.2920
10	0.2407	0.2431	0.2175	0.2423	0.2412
11	0.1250	0.1259	0.1127	0.1245	0.1252
12	0.0782	0.0791	0.0704	0.0778	0.0777
13	0.0345	0.0340	0.0301	0.0332	0.0336
14	0.0829	0.0837	0.0742	0.0809	0.0816
15	0.1069	0.1078	0.1017	0.1060	0.1061
16	0.0938	0.0952	0.0927	0.0921	0.0926
17	0.2716	0.2737	0.2629	0.2701	0.2698
18	0.1277	0.1293	0.1301	0.1282	0.1278
19	0.1953	0.1953	0.1707	0.1920	0.1932
20	0.0737	0.0735	0.0763	0.0721	0.0728
Column mean	0.1488	0.1495	0.1373	0.1480	0.1481

Table 3: McFadden's pseudo-R^2 for each method with 19 items predicting the remaining item. The means use Fisher's transformation. The column heading refer to: the two item parameter IRT model (2PL), treating all rapid responses as errors and then applying the 2PL (TARRE), the EZ diffusion model (EZdiff), van der Linden's hierarchical model allowing response time discrimination to vary among items (vdl), and his model with a single discrimination value (vdltm1).

	2PL	TARRE	EZdiff	vdl	vdltm1
2PL		0.0274	0.0000	0.1589	0.0414
TARRE	0.0007		0.0000	0.0006	0.0001
EZdiff	-0.0115	-0.0122		0.0000	0.0000
vdl	-0.0006	-0.0014	0.0108		0.7133
vdltm1	-0.0006	-0.0013	0.0109	0.0001	

Table 4: The mean differences (lower triangle) and p values (upper triangle) comparing each of the models. Positive values in the lower triangle mean the row method performed better than the column method. Columns are defined as in Table 3.

thoughtful response on one item) for test takers randomly and rapidly guessing if they ran out of time at the end of a test. Based on the results presented in this paper, this would be beneficial for scoring accuracy. However, it would be problematic if students were worried about responding too quickly when they were able to answer a response thoughtfully. This could increase test anxiety and be unfair to those who are not good at assessing duration (which is not a construct that is supposed to be measured by the main standardized tests). However, test scores are used for many purposes. TARRE, with its better predictive accuracy, could be used in some circumstances without concern. For example, these methods could be used with historical data for research purposes.

If the algorithm used by testing organizations was changed to reflect the general speed-accuracy trade-off, as with the diffusion models (and others), then there would be an advantage for responding quickly (if accuracy were maintained). This would be contentious. In some cases the argument can be made that rapid responding shows better grasp of the material. For example, if two students are presented with $5 \times 4 = X$ and one rapidly responds "20" and other counts fingers and toes before saying "20", then it is likely that the first student has mastered a skill that the other has not. However, with more complex mathematical problems speed may be less informative, as the opening quotation from Laurent Schwartz suggests. In addition, if this prompted test preparation companies to encourage students to proceed too fast this could have negative consequences. Further, many classmates take these tests together at the same test centers. This could result in classmates racing each other. While students are required on timed tests to have some competence in time management, the tests are not designed to measure this.

Another class of non-thoughtful (or not the right kind of thought) response that can be rapid are when one test taker copies from another or in some other way has the answers available without the need to read the question. While mixing the order of questions and response alternatives on computer based tests means the test taker would still need to skim the question and alternatives, it is likely that responses to these could also be

quicker than 10 seconds. Of course if someone were trying to cheat and avoid being caught, the person might also be aware about any timing rules and respond at a rate appropriate for thoughtful responding.

Different models may preform better or worse with different kinds of tests. Mathematics tests are different from reading tests. It is important to continue testing different response time models. As Widiatmo and Wright (2015) showed, how well a response time/accuracy model fit depends on how closely the statistical model corresponds with how the data were generated. It is important to understand the relation between the cognitive processes used to answer a question and the response time. Further, there are variations of both sequential sampling models (of which the *EZ* diffusion model is an example) and hierarchical latent variable models (of which van der Linden's model is an example). The method used to compare the models, leave-out-one-item cross-validation, can be used to compare the predictive value of these variations with a single sample.

The focus of this paper has been on using response time as collateral information to improve the estimation of ability. There are other types of *metadata* that are increasingly available because of computer administered tests. This includes tracking mouse/cursor movements. It is also possible to measure different bio-markers (e.g., heart rate, galvanized skin response) during tests. Future research is necessary to explore if these measurements can be used to improve psychometric measurement.

Summary

The procedure that had the highest predictive value of those examined here was Treating All Rapid Responses as Errors (TARRE). In addition to predictive value, other advantages include that it is relatively simple to justify because the responses are too rapid to demonstrate thoughtful responding, that the adjustment is simple to explain, and that all the diagnostic output of IRT packages can be used. However, the consequences of using this adjustment on test taker strategies, fairness, and anxiety need to be explored before any changes are made. Shakespeare used the word *tarre* to mean *invoke*. The TARRE procedure described here, or any procedure that uses response times, should not be adopted without considering the consequences. Particularly with standardized high-stakes tests, changing the scoring algorithm without considering the consequences could, quoting the Bard of Avon, "tarre them to Controuersie".

The Nation holds it no sinne, to tarre them to Controuersie.
(Shakespeare's Hamlet, 1623, ii. ii. 354)

References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach. 3rd ed.* West Sussex, UK: Wiley.

Brown, S. & Heathcote, A. (2008).The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153–178.

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

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package **cirt**. *Journal of Statistical Software, 20* (7).

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*, 35–41.

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

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

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Novick, M. R. & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

Schwartz, L. (2001). *A mathematician grappling with his century*. Basel, Switzerland: Birtkhäuser:Verlag.

Thissen, D. (1983). Timed testing: An approach using item response testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 179-203). New York: Academic Press.

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.

van der Linden, W. J., Klein Entink, R. H, & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*, 327–343.

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

van Ravenzwaaij, D. & Oberauer, K. (2009). How to use the diffusion model: Monte-Carlo simulations evaluating parameter recovery of three methods: EZ, Fast-DM, and DMAT. *Journal of Mathematical Psychology, 53*, 463–473.

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.

Widiatmo, H. & Wright, D. B. (2005). Comparing two item response models that incorporate response times. Paper presented at NCME. Chicago, IL.

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.

Wright, D. B., & Skagerberg, E. M. (2006). A dialogue about MCQs, reliability, and item response modelling. *Psychology Teaching Review, 12*, 66–79.

Xie, Y. (2013). *Dynamic documents with* R *and* **knitr**. Chapman and Hall/CRC.

Zickar, M. J. (1998). Modeling item-level data with item response theory. *Current Directions in Psychology, 7*, 104–109.