AFRICAN JOURNAL OF THEORY AND PRACTICE OF EDUCATIONAL ASSESSMENT (AJTPEA)

VOL 8(1), 2019: 152 – 162 Copyright © Educational Assessment and Research Network in Africa (EARNiA). ISSN 2545-5834
http://www.earnia.org; info@earnia.org; earnia13@gmail.com

COMPARISON OF TRADITIONAL AND RESPONSE TIME ITEM RESPONSE THEORY MODELS IN THE ESTIMATION OF EXAMINEES' ABILITY

Omonike R. Lawal & Michael Akinsola Metibemu

Abstract

Theoretically, Response Time (RT) Item Response Theory (IRT) models have been adjudged to be superior to their counterpart traditional IRT models in the estimation of examinees' ability, although empirical evidences that justify this claim are scarce. This study examined the comparability of the ability of examinees estimated using traditional and RT IRT models. The study is a non-experimental research of correlational type. Multistage sampling procedure was adopted in the selection of 874 students who participated in the study. The instrument used for data collection was a Computer-Based Mathematic Achievement Test (r = 0.89). The responses of the examinees to the test items and the time taken to respond to the test items were obtained and analysed using traditional 3-parameter logistic and RT lognormal IRT modelsas well as thepaired sample t-test. Result showed that the ability estimate of examinees under traditional and RT IRT models were quite similar. However, the relative standing of the examinees ability score under RT and traditional IRT models were different from one another. While RT IRT model was not superior to traditional IRT model in the estimation of examinees ability estimate when average ability estimates of examinee is considered, the relative standing of the examinees ability were different under the two models. Recommendations were made based on the finding.

Keywords: Response time item response theory model, traditional item response theory model, ability estimate, relative standing of ability estimates

Introduction

The world today is experiencing multidimensional changes in its entire sphere. The developed, developing and the underdeveloped nations are all technologically advancing towards making positive and innovative transformations that would aid societal growth, so as to meet up with the pressing needs of the present day world of work. The transformations are all encompassing, including education. By implication, the quality of education any individual might have acquired has potentials for contributing to the growth of society. The quest for quality education that could substantially impact a society for the expected transformations has necessitated the demand for objectively measuring students' learning outcomes through testing, which is one of the fundamental issues in education. This makes measurement and assessment in the field of education important. Measurement and assessment focuses on the design and selection of tests with minimum error so as to obtain valid and reliable assessment results. Ojerinde (2015) viewed assessment as the core of education. In fact, test scores generated from assessments are used to gauge learners' academic strengths and weaknesses. In this regard, practitioners in educational systems are interested in measuring and assessing learners' learning outcomes with utmost precision and accuracy.

In view of this, several models were developed for the measurement of learning outcomes. Among these are the classical test theory (CTT) and the item response theory (IRT) models. The CTT approach; generally known as the traditional method, centers on observed scores. The observed scores here is assumed to be made up of examinees' true and error scores. However, CTT framework is said to be limited because of its circular dependency on number of items on a test. More importantly, its inability to remove error scores from the true score of examinees. The criticism attracted by CTT lead to the development of IRT models (Hambleton and Jones, 1993; Fan, 1998). IRT framework seems to be more detailed, and it possesses invariance property in assessing students' test performance (Ogunmakin & Shogbesan, 2018; Ojerinde, Popoola, Onyeneho & Akintunde 2013; Courvile, 2004). The IRT approach rests on the principle that respondent's performance on a given item is determined by his ability and the characteristics of the item (Fulcher and Davidson, 2007).

Several IRT models have been developed to measure different assessment scale data such as binary, nominal, ordinal or continuous scale according to their unique characteristics. These models range from unidimensional dichotomous (**UIRT;** 1-, 2-, 3- and 4-parameter logistic (PL)) models, to the unidimensional polytomous types (Nominal response model (NRM); Partial Credit Model (PCM); Generalized Partial Credit Model (GPCM) and Graded Response Models (GRM)). Also, the multidimensional dichotomous models (**MIRT;** M1PL, M2PL and M3PL) and their polytomous counterparts (Multidimensional Nominal Response Model (MNRM); Multidimensional Partial Credit Model (MPCM); Multidimensional Generalized Partial Credit Model (MGPCM) and Multidimensional Graded Response Models (MGRM). These models have been used in diverse ways to estimate students' abilities (Fakayode, 2017; Metibemu, 2016; Peterson, 2014; Adegoke, 2013; 2014; Loken & Rulison, 2010;

Lanza, Foster, Taylor & Burns, 2005; Embretson & Reise, 2000; Muraki, 1992). All these models are called response accuracy models because the elicited responses extracted from examinees when answering items of a scale are used to estimate their abilities and the parameters of the supposed model used in calibration. Some of these models are presented in the given equations. Equations 1 and 2 thereby present the 3- and 4-PL unidimensional dichotomous models while equation 3 expresses the M2PL model. Equation 4 gives the polytomous generalized partial credit response model.

$$P_i(\theta_s) = \Pr(X_{is} = 1|\theta_s, a_i, b_i\ , c_i) = c_i + (1 - c_i)\frac{1}{1+e^{-1.7a_i(\theta s - b_i)}} \text{.........eqn1}$$

$$Pi(\theta_s) = \Pr\left(X_{is} = 1\middle|\theta_s, a_i, b_i, c_{i,} d_{i,}\right) = c_i + \left(d_{i,} - c_i\right)\frac{1}{1+e^{-1.7a_i(\theta s - b_i)}} \text{.........eqn2}$$

$$P\left(U_{ij} = 1|\theta_j, a_i, d_j\right) = \frac{e^{a_i\,\theta'_j + \mathrm{d}_i}}{1+e^{a_i\,\theta'_j + \mathrm{d}_i}} \text{...........................eqn3}$$

$$P\left(Y_{ij} = k|\theta_i\right) = \frac{exp\{\sum_{u=1}^{k} D\alpha_j(\theta_i - \beta_j + K_{ju})\}}{\sum_{v=1}^{k_j} \exp\{\sum_{u=1}^{k} D\alpha_j(\theta_i - \beta_j + K_{ju})\}} \text{..................................eqn4}$$

Meanwhile, the choice of any model in the calibration of the its parameters (item characteristics and examinee's ability), lies on a specific model that best fits the response data of an assessment scale, which however, necessitated model-fit analysis.

Other models that were later formulated in the IRT framework are the Response Time (RT) models, which was as a result of the shift from Paper-Pencil Test (PPT) to Computer-Based Test (CBT) in an attempt to improve the quality of assessment. CBT as a modern mode of assessment is gradually becoming popular within and outside the classrooms as well as in high and low-stakes assessments. van der Linden (2009) as well as Schnipe and Scrams (1999) affirmed that a signi? cant advancement in psychometrics was created through the administration of CBT that has enabled more sophisticated approaches to measuring some variables that were considered difficult to assess in the pre-CBT educational setting era. One of such variables is Response Time (RT), which is referred to as the time an examinee spends in supplying correct response to an item in a test and such time can only be automatically recorded in a computer-based testing.

Fox (2018) was of the view that when RT is involved in any of the IRT model, there is likelihood of a better estimate of the model's parameters. There is a presumed clue that measuring response time could improve the quality of ability estimation as well. van der Linden, Entink and Fox (2010) and van der Linden (2009) asserted that test theorists have always been fascinated with the relationship that ensued between responses to test items and the time spent by examinees to give correct response. It is then believed that computerization of educational tests has acted as a motivating factor to the current interest in RT modeling. Therefore, concerted efforts by researchers within educational body proposed and used different IRT response time models. These models incorporated (a) response timesas predictors into IRT response models; (b) response parameters as predictors for modeling response times; and (c) hierarchically modelling of response and response time models simultaneously.

Daniel (2016) expressed that integrating RT into a model that is meant for estimating ability could in a way impact the estimation of learners' response strategies and improve their ability. Various RT models that have been used in diverse studies are evident in literature. Some of which include Thiessen's Lognormal model (Thiessen, 1983); Wang and Hanson model (Wang & Hanson, 2005; 2006); Graviria model (Graviria, 2005); and van der Linden Hierarchical model (van der Linden, 2007). Thiessen's Lognormal and Wang and Hanson models are mathematically formulated in equation 5 and 6 as:

$$logT_{ij} = \mu + \tau_j + \beta_i - \gamma\left[a_i\left(\theta_j - b_i\right)\right] + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathrm{N}(0,\sigma^2) \text{ ....eqn5}$$

Where, *logTij* is the log response time of the person *I* who answers to question *j,μ* depicts overall average, *jμ* depicts overall average, τ_j is the slowness characteristics of person *i*, β_j stands for slowness characteristic of question *j*.γ shows the value attached to the log response time as the coefficient of regression for 2PL IRT model,[a_i (θ_j – b_i)] is the IRT 2PL model and ε_{ij} specifies the chance inadequacy.

$$P\left(x_{ij} = 1 \middle| \theta_i, \tau_i, a_j, b_j, c_j, \beta_j, rt_{ij}\right) = c_j + \frac{1-c_j}{1+e^{-1.7a_j\left[\theta_i - \left(\beta_j\tau_i/rt_{ij}\right) - b_j\right]}} \text{ ....eqn6}$$

The parameters ?, a_j ,b_j and c_j remain the same as in the 3PL dichotomous model. However, rt_{ij} is the time respondent i respond to item *j*, β_j gives item slowness parameter while τ_i presents respondent slowness parameter.

Empirically, few researchers have interrogated the usage of response time models in the analysis of test data. In a study by Ratcliff, Thapar, Gomez and McKoon (2004), patterns of response alongside the time taken were recorded and observed. The response data gotten from a lexical-decision task of youthful and established grown-ups respondents was subjected to a response time diffusion model. The study showed that the lengthier response times observed by more established grown-ups was as a result of the longer non-decision time and bigger limit detachment, although the finding does not explain a general mental handling slow-down process. Suh (2016) employs Bayesian estimation approach on a CBT standardized English verbal test to 978 examinees. Investigation of the responsetime was carried out on three (3) different response time models (Thissen's Lognormal, Wang and Hanson's RT and van der Linden's hierarchical framework) that have the traditional IRT models incorporated in them. Examinee's ability, item characteristics and related parameters were estimated. Findings indicated that much improvement was not shown when ability and item parameter estimates of the typical IRT and response time models were compared. However, the hierarchical framework gave the best model goodness of fit with the lowest deviance information criterion estimate.

Although the superiority of RT model over the traditional item response models in the estimation of examinees test's performance has been documented theoretically, the veracity of the claim has not been widely accepted in the testing community. This is because little is known about the practical advantage of IRT response time models over the traditional IRT model. Therefore, further research involving the use of empirical data is needed to affirm whether inculcating response time will estimate ability differently. Thus, this study;

1. calibrated a Computer-Based Mathematics Achievement Test (CBMAT) with time of response to each test item recorded and
2. assessed the comparability of the ability estimates of examinees under RT and traditional IRT models.

Methods

This study employed a causal comparative research design of non-experimental type. The entire senior secondary school II mathematics student in Lagos State was the target population. Simple random sampling was used to select education district 1 from the 6 education districts into which secondary schools in Lagos State are stratified. Education district 1 comprised 3 Local Government Areas (LGAs). Purposive sampling was used to pick 8 schools that have functional computer laboratories in each of the three LGAs. The sample size used was 874 students. The students were randomly chosen from each of the science, commercial and art classes in each school. Computer-Based Mathematics Achievement Test (CBMAT) was the instrument for data collection for the study which was made up of 40 multiple choice test items with four response options that were dichotomously scored as 1 for correct response and 0 for incorrect response. IRT empirical reliability value of 0.89 was generated after the instrument has been validated with the aid of the test-blueprint mechanism. Before the calibration of the test, the assumption of unidimensionality and item local independence were assessed and the model that fitted the data under traditional and RT IRT models were determined. Data were analyzed using 3-Parameter Logistic (3PL), Lognormal Response Time IRT (LNIRT) models and paired sample t-test. The traditional IRT and RT-IRT estimation was done using multidimensional Item Response Theory (MIRT) and LNIRT packages of the R language and environment for Statistical Computing version 3.5.3, respectively.

Results

The data generated from the CBMAT was calibrated with the unidimensional 3PL and 3PL based LNIRT models that fitted the data under the two classes of IRT models. Thereafter, the estimated ability parameters under the two models were extracted and subjected to paired-sampled t-test statistic. The result is presented in Table 1.

Table 1: Paired-sampled t -test of examinees ability estimates in CBMAT under 3PL and 3PL based LNIRT models

	Mean	N	Std. Deviation	mean dif	t	df	Sig. (2-tailed)
TRAD IRT	0.0003	874	0.862	0.00118	-0.073	873	0.942
LNIRT	0.0015	874	0.412				

Table 1 shows the ability parameter of the CBTMAT estimated under 3-PL and 3-PL LNIRT models. The table shows that the examinees' parameter estimate in the CBMAT estimated under LNIRT model was higher ($\overline{x}$ = 0.0015,*SD* = 0.412) than the ability estimate of the examinees under the traditional 3PL response model ($\overline{x}$ = 0.0003,*SD* = 0.862). However, the difference observed in the mean (mean difference = 0.001) of the ability estimates of the examinees under the two IRT models was not significant (t (873) = 0.073, $p > 0.05$). The result showed that there was a little but insignificant difference in the ability estimate of examinees under 3-PL and 3-PL LNIRT models. The implication of the result is that the average ability estimate of examinees does not change significantly with the introduction of time of response in the modeling of the ability estimate.

Comparison of examinees' ability relative standing

The estimated ability estimates of the examinees under the traditional 3PL and 3PL-LNIRT models were ranked from the highest to the lowest. Thereafter, the relative standing of the examinees on the test were indicated under the two models respectively. The result is presented as follows.

Table 2: Abridged estimated examinees ability and the ranking ability estimates

S/N	TRAD IRT	LNIRT	Rank TRAD IRT	Rank LNIRT
1	2.60360	1.57922	1	1
2	2.57757	1.49498	2	2
3	2.45534	1.47682	3	3
4	2.45524	1.42028	4	4
5	2.35440	1.37806	9	5
6	2.40534	1.36249	7	6
7	1.95678	1.29994	19	7
8	1.86759	1.23230	23	8
9	2.43194	1.21702	5	9
10	2.43194	1.21155	6	10
	+	+	+	+
864	-1.40903	-0.65303	857	864
865	-1.04004	-0.65722	793	865
866	-0.80385	-0.66207	715	866
867	-1.25052	-0.67135	837	867
868	-1.27758	-0.67952	839	868
869	-1.43795	-0.72318	858	869
870	-1.29894	-0.72372	843	870
871	-1.61624	-0.72891	870	871
872	-1.56761	-0.76458	868	872
873	-1.74482	-0.80974	873	873
874	-1.72119	-1.10813	872	874

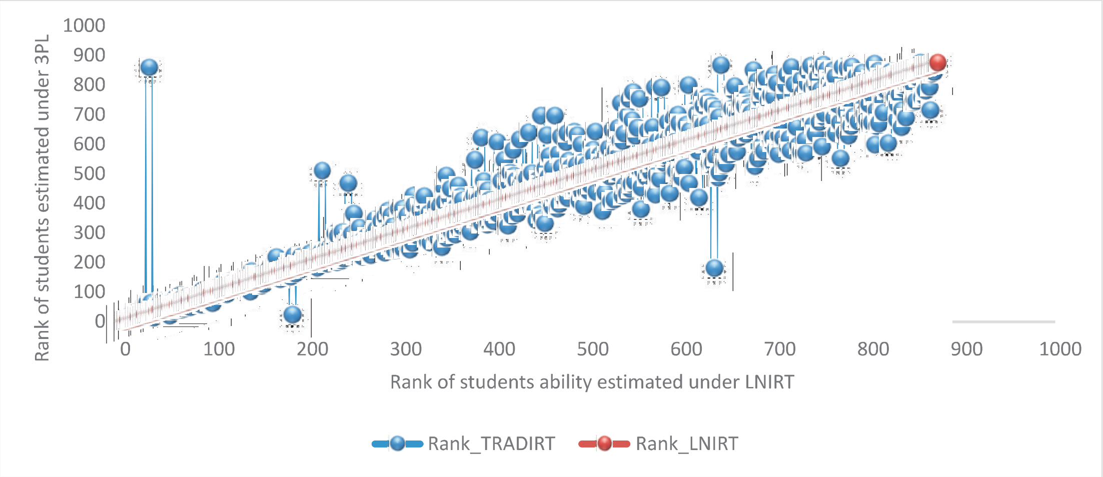

Figure 1: Ranking of examinees ability estimates estimated under 3PL and LNIRT models

Table 2 shows the abridged ability estimates of the examinees under the traditional 3PL and LNIRT and the respective ranking of the examinees based on the IRT models. While Figure 1 shows the distribution of the rank of the examinees' ability estimated under 3PL and LNIRT. The table and figure show that apart from the first four examinees that that maintained their rank under the model, all other examinees had different positions under the two IRT models. The result showed that the ranking of examinees ability under the traditional 3PL and LNIRT models. The implication of the result is that the ranking of examinees ability under the traditional IRT and LNIRT model are different from one another.

Discussion

The findingsof this study for research question one when calibration was made with two different models (3PL and LNIRT) in the same framework (IRT) revealed somewhat different estimates for each of the model's ability parameters. Mean ability estimate for LNIRT response time model estimate yielded a higher mean than estimate for 3PL response model. However, the difference observed in the ability estimates of the examinees under the two IRT models was insignificant. This finding lays credence to study of Suh (2016), where examinees ability estimated under 4PL RT model, hierarchical framework and Thissen's models showed high correlations. It however, negates the finding of Cizek and Wollack (2016) whose study indicated that test takers ability varied significantly across the pretested groups that they considered. For findings on the second research question, it was established that in the ranking of examinees ability, the traditional and LNIRT IRT models worked differently.

Conclusion and Recommendations

The campaign for innovations in the way assessment is conducted has brought development and formulations of different models in IRT framework. This has become a pointer to having more approaches to objectively measure students' learning outcomes so as to represent the true ability of examinees all stakeholders are all clamoring for (performances). In this study, the researchers employed the popularly known 3PL response accuracy model and the LNIRT response time model (fairly new in this clime as a result the bloom in CBT usage) on computer-based mathematics achievement test to make comparison on the two models' ability parameter estimates. It is believed that the usage of LNIRT helps to make more valid inferences about ability in an educational test data, because of the increasing attention CBT is offering. This study concluded on the note that, on the average, LNIRT response time model and traditional IRT model produced similar examinees' ability. However, when it come to the ranking of examinees' ability, the two model produced different results Hence, we recommended that when it comes the estimation of the average ability estimate of examinees, either of the traditional or LNIRT can be employed in the process of test calibration. But when the focus is on the assessment of relative standing of examinees on a test, the LNIRT response time model should be the preferred model.

References

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International journal of educational science,* 2(2), 107-113.

Adegoke, B. A. (2013).Comparison of item statistics of physics achievement test using classical test and item response theory frameworks: *Journal of Education and Practice,* 4,(2)

Adegoke, B. A. (2014). The role of item analysis in detecting and improving faulty physics objective test items: *Journal of Education and Practice.* 5(21)

Alordiah, C. O. (2015). A progressive step in educational measurement: An Application of the rasch model on mathematics achievement test. *Nigerian journal of educational Research and Evaluation,* 14 (3)

Ariyo, A. O. & Lemut, T. I. (2015).Ensuring quality in the test development process through innovations in item calibration: A comparison of classical test theory and item response theory eras in Jamb, Nigeria. 3rd AEAA Conference, Accra, Ghana, 28th August–2nd September.

Cizek, G. J. & Wollack, J. A. (2016).*Exploring cheating on tests: The context, the concern and the challenge.* In G. Cizek and J. A. Wollack (Eds.), Handbook of Detecting Cheating on Tests. New York, NY: Routeledge.

Courville, T. G. (2004). An empirical comparison of item response theory and classical test item/person statistics. Unpublished Doctoral Thesis, Texas A and M University.

Embretson, S. E. & Reise, S. P. (2000). Item Response Theory for Psychologists, Mahwah, NJ: Lawrence Erlbaum Associates.

Entink, R. H. K. Fox, J. P. & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika,* 74 (1), 21-48.

Fakayode, O. T. (2018). Relative Effectiveness of CTT and IRT in equating WAEC Mathematics test scores for June and November 2015. A Ph.D thesis. International Centre for Educational Evaluation (ICEE), Institute of Education.University of Ibadan.

Fox, J. P. & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement,* 54(2), 243–262. ISSN 1745-3984.

Fox, J. P. & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Journal of Multivariate Behavioral Research,* 51(4), 540-553.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment:* An advanced resource book. London and New York: Routeledge.

Gaviria, J. L. 2005. Increase in precision when estimating parameters in computer assisted testing using response time. *Quality and Quantity,* 39, 45-69.

Lanza, S. T. Foster, M. Taylor, T. K. & Burns, L. (2005). Assessing the impact of measurement specificity in a behaviour problems checklist: An IRT analysis. Technical Report 05-75. University Park, PA: The Pennsylvania State University, the methodology centre.

Loken, E. & Rulison, K. L. (2010). Estimation of a 4-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology,* 63(3), 509-525.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Metibemu, M. A. (2016). Comparison of classical test theory and item response theory frameworks in the development and equation of physics achievement tests in Ondo State, Nigeria. A Ph.D thesis, Institute of Education, University of Ibadan.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *AppliedPsychological Measurement,* 16, 159-176.

Nenty, H. J. (2004). From classical test theory (CTT) to item response theory (IRT): An introduction to a desirable transition. In O. A. Afemikhe and J. G. Adewale (Eds.), *Issues in educational measurement and evaluation in Nigeria (in honour of 'Wole Falayajo)* (Chapter 33, pp.371 – 384).Yaoundé, Cameroon: Educational Assessment and Research Network in Africa.

Nenty, H. J. (1998). Introduction to item response theory. *Global Journal of Pure and Applied Sciences,* 4(1), 93-100.

Ogunsakin, I. B. & Shogbesan, Y. O. (2018). Item response theory (IRT): A modern statistical theory for solving measurement problem in 21st century. *International Journal of Scientific Research in Education,* 11(3), 627-635.

Ojerinde, D. (2015). Innovations in assessment: JAMB experience. *Nigerian Journal of Educational Research and Evaluation*, 14(3), 1-9.

Ojerinde, D. (2013). Implementing and sustaining ICT-based assessment and evaluation in the Nigerian education system. National conference on ICT in education, National University Commission (NUC) Auditorium, Maitama, Abuja, 19th -20th November.

Ojerinde, D. Popoola, K. Onyeneho, P. & Akintunde, A. (2013). Item response function: A systematic tool for enhancing test item quality. 39th IAEA conference, Tel Aviv, Isreal, 20th - 25th October.

Ojerinde, D. (2016). The preface of vital issues in the introduction of computer-based testing in large-scale assessment. A compilation of papers presented at Local and internationalconferences. Joint Admission and Matriculation Board (JAMB). ISBN: 978-978-953-759-4.

Okpala, P. N. Onocha, C. O. & Oyedeji, O. A. (1993). *Measurement in Education*. Jattu Uzairue: Edo State, Stirling-Horden Publishers (Nig) Ltd.

Ratcliff, R. Thapar. A. Gomez, P. & Mckoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, 19, 278-289.

Rulison, K. L. & Loken, E. (2009). I've fallen and I can't get up: Can high ability students recover from early mistakes in computer adaptive testing? *Applied Psychological Measurement,* 33, 83– 101.

Schnipe, D. L. & Scrams, D. J. (1999). Response-time feedback on computer administered tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Suh, H. (2016). A study of Bayesian estimation and comparison of response time models in item response theory. Unpublish Ph.d thesis. Department of Psychology and Research in Education. University of Kansa, USA.

van der Linden, W. J. Entink, R. H. K. & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement,* 34(5), 327–347 sagepub.com/journals Permissions.

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

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

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement,* 46(3), 247–272.

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