Article

Modeling Individual Differences in Numerical Reasoning Speed as a Random Effect of Response Time Limits

Applied Psychological Measurement
35(6) 433–446

DOI: 10.1177/0146621611407305
http://apm.sagepub.com

Robert Semmes[1], Mark L. Davison[1], and Catherine Close[1]

Abstract

If numerical reasoning items are administered under time limits, will two dimensions be required to account for the responses, a numerical ability dimension and a speed dimension? A total of 182 college students answered 74 numerical reasoning items. Every item was taken with and without time limits by half the students. Three psychometric models were fit to the data—one including no time-limit effect, one including a fixed time-limit effect, and one including a random effect of time limits. The latter model best fit the data, suggesting that a speed dimension, the random effect of time limits, is needed to account for time-limited responses. The estimated reliability of the Speed scores was .39. Despite this low reliability, Speed scores were correlated with American College Testing (ACT) math scores and response times. Speed scores added significantly to the ACT math score variance accounted for by the numerical reasoning dimension in the model. A within-person log-odds ratio interpretation of the Speed score is proposed. Possible methods of improving Speed score reliability and methods for studying the speed dimension are discussed.

Keywords

hierarchical linear model, linear logistic model, multidimensional item-response theory, numerical reasoning, quantitative reasoning, Rasch model, random effects, response time limits, test time limits, speededness

Suppose a person was asked to solve a set of cognitive test items that were all of the same type (e.g., mathematical reasoning items) but varied with respect to their difficulty. If the person was given a stringent time limit within which to answer items, would the person's total number-correct score measure the same ability (or abilities) as would be measured if the person was allowed to answer the items without time constraints? Hambleton and Swaminathan (1985)

[1]University of Minnesota, Minneapolis, USA

Corresponding Author:
Mark L. Davison, Department of Educational Psychology, University of Minnesota, 56 E. River Rd., Minneapolis, MN 55455
Email: mld@umn.edu

have stated what seems to be the consensus opinion—the speededness induced by time limits changes the dimension(s) measured: ''When speed affects test performance, then at least two traits are impacting on test performance: speed of performance, and the trait being measured by the test content'' (p. 30). Their conclusion seems largely based on studies of time-limited tests. According to Nunnally (1978), such studies largely appeared before 1960 (e.g., Davidson & Carroll, 1945; Lord, 1956) culminating in a 1960 special issue of *Educational and Psychological Measurement*. A notable later study was Peterson's (1993) review of test time-limit effects on the General Aptitude Test Battery. In Peterson's opinion, subsequent advances in statistical methodologies, particularly random coefficient modeling, warrant a reconceptualization of the speed dimension and suggest better methods for the study of that dimension.

In this study, the authors investigated the effect of item-response time limits capitalizing on techniques not readily available or not in wide use at the time of the earlier research: computer administration of test items and hypothesis testing methods based on random coefficient models. The purpose was to study the existence of a speed dimension underlying timed item performance, the reliability with which such a speed dimension could be measured, and the validity of the speed dimension.

The evidence of the earlier studies, based on a total test time limit rather than item time limits, is actually somewhat mixed and open to alternative interpretations. From the very beginning, researchers investigating test time limits studied several tests covering different content domains. For instance, Davidson and Carroll (1945) studied 13 content domains. They found evidence for speed factors, but the factors were defined by a combination of response times and responses administered under total test time limits, not just by the responses administered under total test time limits. In what is probably the most widely cited study of test time limits, Lord (1956) focused primarily on the existence of a second, speed factor in three domains: arithmetic reasoning, spatial visualization, and vocabulary knowledge. He concluded that the data provided evidence for a speed factor in two of the domains, spatial visualization and vocabulary but not arithmetic reasoning. Jöreskog (1971) reexamined some of Lord's data. His best fitting model included arithmetic reasoning, spatial visualization, and verbal knowledge. Jöreskog's results support the existence of a speed factor in all three domains. However, Lohman (1979) questioned whether Lord's data provided evidence for a speed factor in any domain. He analyzed the correlations themselves rather than factors derived from those correlations. Analyzing each content domain separately, he found that the average correlation between two tests administered in the same speed condition (both tests unspeeded, both tests moderately speeded, or both tests highly speeded) was no higher than the average correlation of two tests administered in different speed conditions.

Furthermore, when tests administered with time limits do generate a second factor, it may not be a speed factor. A second dimension may reflect individual differences in test strategy. Attali (2005), Boughton (2009), and Wise and DeMars (2009) discussed a response strategy for speeded multiple-choice tests in which examinees proceed at their own (unspeeded) pace for items early in the test and then either omit or randomly guess for the remaining items. When examinees adopt this strategy, only items toward the end of the test are truly speeded. If there is an additional factor underlying responses to time-limited tests, it may reflect individual differences in pacing, guessing, and omission strategies, not speed of reasoning.

In the present study, all items were computer administered. A separate time limit was imposed on every item. All items required a short, open-ended response. Because of the item time limits, every item was speeded, not just those near the end of the test. There could be no effect of guessing on multiple-choice items because there were no multiple-choice items. All items in a form were presented to every examinee; there could be no effect of individual differences in unreached items because there were no unreached items. Item content was crossed with the

Table 1. Summary of Participants' Demographic Characteristics and ACT Math Scores

Variable	Sample VW	Sample WV
Sample size	91	91
Percentage of women	60.2	37.2
Median age	20	20
Lowest age	18	18
Highest age	43	47
Percentage of native English speakers	83.9	83.0
Median ACT Math score (% sample with scores)	27 (82.8)	26 (69.1)
Lowest ACT Math score	17	15
Highest ACT Math score	34	35

Note: ACT Math scores are reported on a scale ranging from 0 to 36. During the 2003-2004 school year, an ACT Math score of 26 was at the 85th percentile among college applicants who took the ACT. Too few students released SAT Quantitative scores for meaningful summaries to be reported.

time-limit/no time-limit factor; every item was taken by half the examinees with a time limit and by the other half of the examinees without a time limit. In other words, every item was administered in both timed and untimed test administrations so that the effects of item content were crossed with, not nested within, the time-limit factor. The hypothesis of a second, speed factor was formally tested using a random coefficient model.

In this study, it was hypothesized that there would be an effect of time limits on item performance, such that the probability of correctly answering an item would decrease as a result of the time limits. Furthermore, it was hypothesized that the effect of time limits would be a random effect varying across persons. This random effect would constitute the second dimension introduced into the response process by time limits. Items in the mathematics/arithmetic domain were studied, a domain in which Lord (1956) was unable to find evidence for a second, speed dimension. The results reported here are based on 74 items divided across two 37 item forms, Form V and Form W, forms that were administered in both the self-paced (untimed) and experimenter-paced (timed) conditions in a counterbalanced design such that item content and speed condition would be fully crossed; that is, every item was administered under both self-paced and experimenter-paced conditions.

Method

Participants

Participants were 182 college students at a Midwestern university enrolled in psychology department courses and who earned extra credit points for their participation. They underwent both self-paced testing (untimed) and experimenter-paced testing (timed). There were two parallel test forms, V and W, and students were randomly assigned to Sample VW or WV. In the sample designations (Sample VW and Sample WV), the first letter indicates the form taken first under self-paced conditions, and the second letter designates the form taken second under experimenter-paced conditions. Table 1 summarizes the primary demographic characteristics and ACT Math scores of the two samples. Despite random assignment, Samples VW and WV differed in their percentages of women: 60% in Sample VW and 37% in Sample WV. In other respects, the two samples were similar.

Measures

The items selected for the final test forms included Differential Aptitude Test (DAT) and published quantitative items drawn from the SAT and Graduate Record Examination (GRE). To eliminate guessing and the need for a pseudoguessing parameter in these psychometric models, all items were presented in a constructed-response format. As compared with their original multiple-choice formats, the constructed-response versions of the items were presumably more difficult.

A total of 74 numerical reasoning items were chosen for analysis in the study. Of them, 37 were assembled into a form called Form V; the remaining 37 were assembled into a form called Form W. All items included in the analysis had been administered in both the self-paced and experimenter-paced conditions, so that item content and speed conditions would be fully crossed; that is, every item was administered under both self-paced and experimenter-paced conditions. Forms V and W were created to be similar in difficulty and content. The self-paced raw-score means of Forms V and W were not significantly different, 25.61 and 24.82, respectively ($p > .05$), but Levine's test led to rejection of the null hypothesis that the variances were equal, 6.96 and 8.71, respectively ($p < .01$).

Setting item time limits. For the timed administration, a time limit was established for each item using data from a pilot administration. For each item, the authors identified examinees in the pilot administration who had correctly answered that item given unlimited time. They then computed the median response time among the examinees answering correctly. After rounding to the nearest 5 s, the rounded median response time became the item's time limit for experimenter-paced testing. Given the way the time limits were set, some people could answer the item correctly without decreasing the time that they would have spent on the item given unlimited time. However, given an announced time limit, there may still be an effect of time limits on such people in that they may answer the item more quickly or may feel more anxiety than they would without a time limit. The time limits ranged from 25 s to 135 s.

Procedure

Students participated in both self-paced (Session 1, untimed) and experimenter-paced (Session 2, timed) testing. A participant's self-paced session always preceded his or her experimenter-paced session, and Session 2 always occurred 1 to 16 days after the participant's Session 1.

For the experimenter-paced Session 2, the test administrator informed participants that each item would be administered with a time limit, the item would disappear from the screen when the time limit expired, and that if no answer were given within 5 s after the time limit expired, the item would be scored as incorrectly answered. Before actual experimenter-paced testing began in Session 2, students answered eight practice items to acclimate them to the experimenter-paced performance conditions. In the actual experiment-paced testing, each item's presentation was preceded by a screen stating the time limit for answering the given item. That time limit also appeared on the item's presentation screen, though no countdown clock appeared on that screen. Participants did not control how much time elapsed between items. As soon as a participant's answer to a given item was recorded, the time-limit screen for the next item appeared.

Analysis Models

Using hierarchical linear modeling (HLM), three explanatory item-response models were fitted to the data (De Boeck & Wilson, 2004). Using the Bayesian information criterion (BIC), the Akaike (1974) information criterion (AIC), and the deviance fit statistics, the three hypothesized

models were compared. Recent psychometric research (e.g., Kang & Cohen, 2007; Kang, Cohen, & Sung, 2009) suggest that, of the various information statistics, the BIC more often identifies the best model among several competing models. Then reliability statistics are reported for the Level and Speed dimensions in the authors' best fitting model. The reliability index equation (Raudenbush, Bryk, Cheong, & Congdon, 2004) as computed in the HLM6 software is as follows:

$$\lambda = \frac{\tau}{\tau + \frac{\sigma^2}{n}}. \quad (1)$$

where τ is the between-person variation estimate, either $\sigma^2(\theta_j)$ or $\sigma^2(\psi_j)$; σ^2 is the Level 1 error variance; and n is the number of items. When the link is the logit function, the Level 1 error variance is commonly taken to be $\sigma^2 = \pi^2/3 = 3.29$ (e.g., Cho & Rabe-Hesketh, 2011). In Equation 1, λ is De Boeck's (2008) ICC(k), the reliability of the sum for all items, not the more commonly discussed (e.g., Cho & Rabe-Hesketh, 2011) ICC(1), the intraclass correlation of a single response for which $n = 1$ in Equation 1. Finally, the validity of the Speed and Level dimensions was explored by regressing ACT scores and response time data onto the Level and Speed scores.

The theory guiding the formulation of the authors' three models is called the Speed–Level hypothesis because it contains two ability dimensions, a Speed dimension and a Level dimension. It can be expressed as follows:

$$\pi_{ij} = \frac{e^{(\theta_j + \psi_j \times \delta_i - \beta_i)}}{1 + e^{(\theta_j + \psi_j \times \delta_i - \beta_i)}}. \quad (2)$$

where π_{ij} is the probability that person j will correctly answer item i. θ_j is the location of person j on a Level dimension underlying self-paced item performance; it reflects the person's maximal mathematical reasoning ability in the sense of his or her performance when given unlimited time. ψ_j is the location of person j on a Speed dimension, a second dimension underlying performance on experimenter-paced items. δ_i is a dichotomous indicator variable that dummy codes the administration condition; that is, $\delta_i = 1$ if the item is administered under experimenter-paced conditions, and $\delta_i = 0$ if the item is administered under self-paced conditions. Given that $\delta_i = 1$ if the item is administered under experimenter-paced conditions and $\delta_i = 0$ if the item is administered under self-paced conditions, the model implies that self-paced responses are a function of the Level dimension only, whereas the experimenter-paced responses are a function of both the Level and Speed dimensions. β_i is the difficulty of item i.

In the model of Equation 2, δ_i functions like a dichotomous discrimination parameter for item i along the Speed dimension, but a discrimination parameter that is given by the administration condition, self- versus experimenter-paced, rather than estimated from the data. For most examinees, ψ_j is expected to be negative, leading to the prediction that experimenter-paced items are less likely to be answered correctly by examinee j than are self-paced items of equal difficulty β_i. However, ψ_j can be positive in which case, person j is more likely to correctly answer experimenter-paced items than self-paced items. If $\psi_j = 0$, person j is equally likely to correctly answer experimenter-paced and self-paced items of equal difficulty. Both θ_j and ψ_j are random effects that vary over persons. Equation 2 can be viewed as a nonlinear mixed-effects model in which θ_j is a random intercept term varying over persons, ψ_j is a random effect of the experimenter-paced administration condition that varies over persons, and β_i is a fixed effect of item i.

Equation 2 contains no guessing parameter because the items were administered in a constructed-response format, and so the probability of correctly guessing an answer is approximately zero. Implicitly, Equation 2 contains an assumption that all items have equal discrimination parameters along the Level dimension, and all experimenter-paced items have equal discrimination parameters along the Speed dimension. These assumptions were adopted despite the fact that they are overly

simplistic. Given the sample size and number of items of this study, the authors probably could not accurately estimate item-difficulty parameters and item-discrimination parameters along two dimensions; a model with equal discrimination parameters can often provide an acceptable fit to the data.

The first hypothesis, the Level-only model, is the special case of Equation 2 in which there is no random effect of response time limits; that is, $\psi_j = 0$ for all j. In this special case, both self-paced and experimenter-paced items would satisfy the standard Rasch model:

$$\pi_{ij} = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}}. \tag{3}$$

Hypothesis 2 is called the Fixed Speed–Level hypothesis. It corresponds to the special case in which the effect of time limits is a fixed effect that does not vary over persons (or items): that is, $\psi_j = \psi$ for all j and $\sigma^2(\psi_j) = 0$. It posits that there is only a single dimension of ability θ_j that underlies responses to both self- and experimenter-paced responses but that there is a fixed effect of time limits on experimenter-paced items, ψ.

$$\pi_{ij} = \frac{e^{(\theta_j + \psi \times \delta_i - \beta_i)}}{1 + e^{(\theta_j + \psi \times \delta_i - \beta_i)}}. \tag{4}$$

The third model is called the Random Speed–Level hypothesis. It is the full model of Equation 2 in which there is a single dimension underlying self-paced items, θ_j, and two dimensions underlying experimenter-paced responses, θ_j and ψ_j. Like Model 2, it posits that experimenter-paced responses require an experimenter-paced effect, but unlike Model 2, it posits that the size of the effect varies across persons.

Model Fitting

To fit each of the hypothesized models, the authors used HLM 6 (Raudenbush et al., 2004). For purposes of HLM, the analysis must be specified as a two-level design. Items were viewed as the Level 1 sampling units nested within people, the Level 2 sampling units. As there were no predictors in the Level 2 model, only the Level 1 equation, Equation 5, is shown. To fit these models, the authors recast the full model of Equation 2 into a logit form. Let

$$\eta_{ij} = \ln\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right).$$

Then the logit form of Equation 2 is

$$\eta_{ij} = \theta_j + \psi_j \times \delta_i - \beta_i. \tag{5}$$

The fixed effects of the items were then dummy coded to create an equation of the form

$$\eta_{ij} = \theta_j + \psi_j \times \delta_i - \sum_{i'=1}^{i'=(I-1)} \beta_{i'} x_{i'}, \tag{6}$$

where $x_{i'} = 1$ if $i = i'$ and $x_{i'} = 0$ otherwise and I is the number of items. β_i in Equation 5 has been replaced by the extended form $\sum_{i'=1}^{i'=(I-1)} \beta_{i'} x_{i'}$ in Equation 6. As dummy coding I categories, one for each item, requires only $I - 1$ predictor variables, $I - 1$ is the upper limit of the sum in

Equation 6 (Kamata, 2001). Equation 6 forms the basis for a nonlinear regression in which the criterion variable is the dichotomous response by person j to item j, the nonlinear link function is logistic, and the predictors are δ_i and the item dummy codes $x_{i'}$. In this model, θ_j is the random coefficient intercept, ψ_j is the random coefficient slope on the predictor δ_i, and $\beta_{i'}$ is a fixed effect slope on the item dummy code $x_{i'}$. In the first step of the analyses for each model, restricted maximum likelihood estimation (REML) with Laplace iteration (Raudenbush et al., 2004) was used to estimate the fixed effect parameters in the model: the mean of the random coefficient intercepts, $\mu(\theta_j)$, the variance of the random coefficient intercepts, $\sigma^2(\theta_j)$, the mean of the random coefficient slopes, $\mu(\psi_j)$, the variance of the random coefficient slopes, $\sigma^2(\psi_j)$, the covariance of the random coefficient intercepts and slopes, $\sigma(\theta_j, \psi_j)$, and the $(I-1)$ fixed item effects β_i. The above procedure yields estimates of item difficulties β_i for which the origin has been set in an unorthodox manner. That is, the origin is set so that the item difficulty for the last item is zero: $\beta_I = 0$. However, this should not affect any of the study's conclusions. As the difficulty of item I is fixed to zero, only $(I-1)$ item-difficulty parameters are estimated.

Model 1 was fit in exactly the same manner as described above except that the predictor δ_i was dropped from the equation. Model 1 posits that there is no effect of item time limits. Fitting Model 1 yielded REML estimates of the mean random coefficient intercept, $\mu(\theta_j)$, the variance of the random coefficient intercepts, $\sigma^2(\theta_j)$, and the fixed item effects β_i; by implication $\mu(\psi_j) = 0$ and $\sigma^2(\psi_j) = 0$ in Model 1.

Model 2 was fitted as described above for the full model except that ψ_j was specified as a fixed effect. Model 2 posits that there is a nonzero effect of time limits, but that the effect is constant across all persons. Fitting Model 2 yielded REML estimates of the mean random coefficient intercept, $\mu(\theta_j)$, the variance of the random coefficient intercept, $\sigma^2(\theta_j)$, the fixed effect coefficient, ψ, and the fixed item effects, β_i; by implication, $\mu(\psi_j) = \psi$ and $\sigma^2(\psi_j) = 0$ in this model.

In Model 1, there is no effect of time limits. In Model 2, the effect of the time limits is specified as a fixed effect ψ common to all persons, an effect that can be considered a shift in the ability scale. In Model 3, the effect of time limits is specified as a person-specific shift ψ_j with mean $\mu(\psi_j)$ and variance $\sigma^2(\psi_j)$ that is a second dimension of individual differences.

Results

Table 2 shows the parameters and the fit measures for the three hypothesized models: Level-only (Model 1), fixed Speed–Level (Model 2), and random Speed–Level (Model 3). For Models 1 and 2, the deviance statistics were 39472.52 and 38843.20, and the difference in degrees of freedom is 1 because Model 2 contains only one additional parameter corresponding to the administration effect, $\psi = \mu(\psi_j)$. The difference in these two deviance statistics was highly significant, $\chi^2(1) = 629.32$, $p < .0001$, leading to rejection of the hypothesis that the Level-only and the Fixed Speed–Level models fit equally well. For Models 2 and 3, the difference in the deviance statistics of 38843.20 and 38813.92 was also significant, $\chi^2(2) = 29.28$, $p < .0001$. Because Model 3 has two additional parameters corresponding to the variance and covariance of $\psi_j, \sigma^2(\psi_j)$ and $\sigma(\psi_j, \theta_j)$, the chi-square difference statistic has two degrees of freedom. Model 3 also had the lowest BIC and the lowest AIC. Therefore, Model 3 seems to be the best model to explain performance on the test items as indicated by the deviance, the AIC, and BIC statistics.

As shown in Table 2, HLM 6 also provided estimates of the correlation between the two random effects θ_j and $\psi_j, \rho(\theta_j, \psi_j) = -0.17$. Although there is some small tendency for those with higher Level scores to have lower Speed scores, the Level and Speed dimensions were only very modestly correlated.

Table 2. Parameters and Fit Measures From Model Fitting

	Hypothesis 1: Level-only model	Hypothesis 2: Fixed Speed–Level model	Hypothesis 3: Random Speed–Level model
Fit measures			
Deviance	39472.52	38843.20	38813.92
Number of estimated parameters	75	76	78
BIC	39862.82	39238.70	39219.83
AIC	39622.52	38995.20	38969.92
Parameters			
$\mu(\theta_j)$	–1.20*	–0.74*	–0.76*
$\mu(\psi_j)$		–1.04*	–1.04*
$\sigma^2(\theta_j)$	1.08	1.21*	1.24*
$\sigma^2(\psi_j)$			0.21*
$\rho(\theta_j, \psi_j)$			–0.17
Reliability θ_j	0.93	0.94	0.87
Reliability ψ_j			0.38

Note: BIC = Bayesian information criterion; AIC = Akaike information criterion. θ_j is the score of person *j* on the mathematical reasoning dimension, and ψ_j is the random time-limit effect for person *j*.
* indicates significance at the .01 level (two-tailed).

The analysis also produced intraclass correlation-based reliability estimates for the Level and Speed scores. These estimates were computed according to Equation 1 and are shown at the bottom of Table 2, $\lambda = .87$ for Level and $\lambda = .38$ for Speed. As compared with Level, the Speed reliability was much lower and certainly not suitable for high-stakes application.

Table 3 shows the means and standard deviations of item difficulties in the three hypothesized models. The correlations of the difficulties estimated under the three models were 1.000 to three decimal places indicating that the relative difficulties of items were stable across models. However, for Model 1, the difficulty parameters were, on average, higher than for Models 2 and 3. In Models 2 and 3, an item's difficulty is its difficulty in the self-paced condition, whereas in Model 1, it is (roughly speaking) an ''average'' of the item's difficulty in the self- and experimenter-paced conditions.

Correlations of Speed Scores With Criterion Variables

For these analyses, Level and Speed scores were estimated for each examinee. To evaluate the validity of the Speed scores, the authors correlated the Speed scores (and the Level scores) with several criterion variables. In self-paced and experimenter-based testing, response times were recorded for each item. For each person, six response time variables were created: median response time for all self-paced items, median response time for all correctly answered self-paced items, median response time for all incorrectly answered self-paced items, median response time for all experimenter-paced items, median response time for all correctly answered experimenter-paced items, and median response time for all incorrectly answered experimenter-paced items. For a subset of examinees, ACT mathematics scores were available. A positive correlation between Speed scores and ACT scores was predicted, but it was not expected to be high; the ACT mathematics test does have a time limit, but it is lenient.

In Table 4, the row labeled ''Median response time (self-paced)'' refers to an examinee's median answer time across all items in the self-paced session. The row labeled ''Median correct response time (self-paced)'' refers to the examinee's median response time across all correctly

Table 3. Descriptive Statistics for the Difficulties of the Three Models

	Hypothesis 1: Level-only model	Hypothesis 2: Fixed Speed–Level model	Hypothesis 3: Random Speed–Level model
Fit Measures			
M	–1.68	–1.76	–1.78
SD	0.94	0.99	1.00
Skewness	–0.52	–0.50	–0.50
Kurtosis	–0.01	–0.07	–0.05

Table 4. Correlations of Level and Speed Scores With Various Variables

	Level	Speed
Median response time (self-paced)	–.13	–.50**
Median correct response time (self-paced)	–.04	–.52**
Median incorrect response time (self-paced)	.16*	–.34**
Median response time (experimenter-paced)	–.16*	–.45**
Median correct response time	.05	–.14
Median incorrect response time	.21**	–.19*
ACT	.72**	.21*
Level		–.18*

Note: ACT =

* indicates correlation is significant at the .05 level (two-tailed).

** indicates correlation is significant at the .01 level (two-tailed).

answered items in the self-paced session, and the row labeled ''Median incorrect response time (self-paced)'' refers to the examinee's median response time across all self-paced incorrectly answered items. The correlations of Speed with median response time, median correct response time, and median incorrect response time in the self-paced sessions were –.50, –.52, and –.34, respectively ($p < .01$). All were negative indicating that those with higher Speed scores tended to take less time in the self-paced sessions. Speed was more highly correlated with the three response time variables than was Level, which was not significantly correlated with median response time and median correct response time but positively correlated with median incorrect response time ($p < .05$).

Speed scores were not quite as consistently correlated with response times in the experimenter-paced sessions. Speed was significantly and negatively correlated with experimenter-paced median response time ($r = –.45$, $p < .01$) and experimenter-paced median incorrect response time ($r = –.19$, $p < .05$) but not significantly correlated with the median correct response time in the experimenter-paced session. However, the latter result might not be unexpected because response times in the experimenter-paced condition are determined more by the imposed time limits than by individual differences in processing speed or preferred pace. In the experimenter-paced condition, time limits produced a rather severe restriction of range on response times. Standard deviations of response times in the experimenter-paced administration tended to be less than half those of the self-paced conditions.

Speed scores had a modest but significant ($p < .05$) correlation with ACT performance, $r = .21$. As the ACT is speeded, but not highly so, only a modest association with Speed scores was

expected. To investigate the extent to which Level and Speed could explain variation in an external criterion variable, a multiple regression analysis was carried out with ACT math scores as the criterion variable. Level alone explained 52.1% of the variation in ACT math scores. With Level already in the model, Speed explained an additional 11.4% of the variation in ACT math scores ($p < .01$).

In summary, results provided evidence that a second, Speed dimension, was needed to account for response accuracy when item time limits were imposed. However, in this data set, individual differences in Speed were measured with only very modest reliability. Evidence for the validity of the Speed dimension is provided by its correlations with response times, particularly response times in the self-paced conditions, its significant correlation with ACT mathematics scores, and its unique contribution over and above Level scores to the prediction of ACT math scores.

Discussion

A major conclusion arising from this research is that the second dimension underlying experimenter-paced responses can be modeled as a random effect of administration condition: self- versus experimenter paced. To best isolate the effect of administration conditions, one needs to use a crossed design in which administration condition is crossed with item content so effects of content are orthogonal to effects of administration condition. Although this study involves item time limits, the effects of test time limits can also be modeled as a random effect using any one of the several designs.

One design is a fully crossed extension of Lord's (1956) design that involves administering one or more short tests with and without time limits. In this design, there are several short tests—say Tests A_1, A_2, A_3, B_1, B_2, and B_3. Separate time limits are imposed on each time-limited test. One group, called Group AB, takes Tests (A_1, A_2, A_3) with total test time limits and (B_1, B_2, and B_3) without time limits. That is, Group AB takes Test A_1 with its time limit, then A_2 with its time limit, then A_3 with its time limit, but then B_1 followed by B_2 followed by B_3 each without time limits. A second group, BA, takes B_1, B_2, and B_3 with time limits but A_1, A_2, and A_3 without time limits. A random coefficient can be used to model the effect of the time limits, and the design is fully crossed in that each test appears in both the timed and untimed administration conditions. The design is similar to that of Lord in that it uses several short tests each with its own time limit, but unlike Lord's design, every test appears in both timed and untimed conditions. In Lord's design, tests were nested within administration conditions so that effects of administration condition were confounded with test content differences.

A variation on the design above involves manipulating the position of an item. For instance, some testing programs include time warnings; that is, ''You have 10 min left.'' An item administered after the warning might be considered more speeded than the same item administered before the warning and hence to depend on a second factor, a random effect of the warning. In the model of Equation 2, it would be possible to set $\delta_i = 0$ or $\delta_i = 1$ for the item administered before and after the warning, respectively.

When there are time limits on the test, examinees can adopt various strategies. The first is a time management strategy in which the examinee self-imposes a time limit on each item to ensure that all items are reached. This strategy may introduce a second dimension similar to the one observed in this study. A second strategy is to proceed as if there were no time limits and to leave the last few items blank if not reached within the time limit. This strategy may introduce a different dimension of individual differences associated with individual differences in omitted responses. For multiple-choice items, a third strategy is to proceed as if there were no time limits for most items and then rapidly guess at the remaining items (Attali, 2005; Wise

& DeMars, 2009). Test time limits may have differing effects on performance depending on the response strategy adopted by the examinees. This would imply that the time-limit effect would be random, varying across individuals, and that the multiple causes of this variation may be difficult to unravel.

In this study, time limits detracted from average examinee performance in that examinees correctly answered fewer items with imposed time limits. However, this does not mean that imposing time limits necessarily detracts from the validity of the test. Scores on the Speed dimension were significantly correlated with scores on the ACT math test, and scores on the Speed dimension added significantly to the variance accounted for by the Level dimension. Despite the low reliability of the Speed dimension, the data provided some evidence for its criterion-related validity and its incremental validity. If the explicit goal is to measure an examinee's maximal performance level, then time limits would interfere with the goal of the measurement. However, if the measurement is intended to predict performance in situations where time or time pressure is a factor, then the imposition of time limits may be appropriate, at least if individual differences in time-limited performance on tests generalizes to individual differences in performance in more real-life situations.

The Speed dimension observed in this study may be of interest in its own right, at least if more reliable measures of the dimension can be constructed. The reliability may be enhanced by increasing the number of time-limited items, increasing or decreasing the time limits, and/or administering time-limited items using a computer adaptive strategy. At each step of the testing, an adaptive test could select an item (with a time limit) whose administration would maximize the Fisher information along the Speed dimension. Such an item-selection algorithm could minimize the number of items needed to reach a target standard error of measurement for each person's Speed score.

In recent years, there has been extensive study of models for response times in essentially self-paced conditions (e.g., Thissen, 1976/1977, 1983; Van Breukelen, 2005; van der Linden, 2006, 2009). Such models often include a Speed dimension or its inverse, a Slowness dimension. It is not clear whether the Speed dimension that underlies self-paced response time variables is the same Speed dimension that seems to underlie the experimenter-paced item responses. However, the correlation in Table 4 between the Speed dimension and the self-paced response time variables suggests that there is a connection. If the same Speed dimension underlies both experimenter-paced item responses and self-paced response times, it may be possible to simultaneously model self-paced item responses, experimenter-paced item responses, and self-paced response times in a manner which yields a more reliable estimate of speed scores.

The Speed dimension has a log-odds ratio interpretation. In this final model, the expression for a self-paced response is the usual Rasch model:

$$\pi_{ij(s)} = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}}. \tag{7}$$

The expression for an experimenter-paced response is

$$\pi_{(ij(e))} = \frac{e^{(\theta_j + \psi_j - \beta_i)}}{\left(1 + e^{(\theta_j + \psi_j - \beta_i)}\right)}. \tag{8}$$

The odds ratio is

$$\frac{\pi_{ij(e)}/(1 - \pi_{ij(e)})}{\pi_{ij(s)}/(1 - \pi_{ij(s)})} = e^{\psi_j}.$$

The log-odds ratio is equal to ψ_j. This ratio expresses the odds of correctly answering an item with time limits as compared with the odds of correctly answering the same item without time limits. If ψ_j equals 0, then the odds ratio is 1.00 for person j; the odds of answering an item with and without time limits are the same. If ψ_j is negative, as its estimate was for most examinees, then the odds of answering an item with time limits are less than the odds without time limits. If ψ_j is positive, as its estimate was for a few examines, the odds are greater for the experimenter-paced than for the self-paced administration of an item.

ψ_j expresses the person's ability under time limits, not in an absolute sense, but relative to performance with no time limits. A person for whom $\psi_j = 0$, a high score in this data, tended to answer the same number of items correctly in both the experimenter-paced and self-paced conditions; that person may or may not have answered a large number of items correctly in the experimenter-paced conditions, depending on his or her performance without time limits. In this model, θ_j is a measure of maximal performance, if maximal performance is taken to mean performance with no time limits. ψ reflects the degree to which performance with time limits was affected by those limits, the degree to which the person could increase his or her speed to the point where they answered within the time limit without sacrificing the accuracy achieved without time limits. This Speed dimension is what Davison (Davsion, Kim, & Close 2009; Kim, Davison, & Frisby) has called a within-person factor in that the log odds associated with the factor score reflects a within-person contrast between the person's model predicted ability to answer items of a given difficulty with and without time limits; for example,

$$\ln\left(\frac{\pi_{ij(e)}}{1-\pi_{ij(e)}}\right) - \ln\left(\frac{\pi_{ij(s)}}{1-\pi_{ij(s)}}\right) = \psi_j.$$

Findings suggest that the Speed factor is best interpreted as a within-person factor, which reflects the ability to maintain maximal performance in circumstances requiring speeded responses. This interpretation would suggest that the speed/accuracy trade-off is an individual differences variable; individuals vary in the degree to which they sacrifice accuracy to achieve greater speed. If reported along with the Level dimension, it would provide information about the person's ability to maintain his or her maximum level of performance under less than ideal conditions.

The model in Equations 7 is a Rasch model and that in Equation 8 is an extension of the Rasch model in which all items were constrained to have discriminations of 1.00 on the Level dimension, and experimenter-paced items were constrained to have discriminations of 1.00 on the Speed factor. Both models can be extended to a three-parameter form in which there is a guessing parameter and, for items with nonzero discriminations on a dimension, the loadings are allowed to vary. In this three-parameter form, self-paced items would be hypothesized to have positive loadings on the Level dimension only; experimenter-paced items would be hypothesized to have positive discriminations on both the Speed and Level dimensions. Fitting such a three-parameter form would seemingly require a much larger sample size than that of the current study.

Three issues were investigated in this study: the existence of a Speed dimension underlying timed item performance, the reliability with which such a Speed dimension could be measured, and the validity of the Speed dimension. The existence of a Speed dimension was supported by the finding that the data were best reproduced by a model that included a speed dimension as a random effect. The Speed score estimate based on the model was only modestly reliable. However, despite the modest reliability, correlations of the Speed score with ACT math scores and response times provided some evidence for the validity of the Speed score.

Consistent with results reported by Lord (1956) and Jöreskog (1971) for time-limited tests, results suggest that performance with item time limits reflects two factors, a maximal performance Level factor and a Speed factor. However, this model suggests a new interpretation of

the Speed factor. In the context of this model, the Speed factor has a within-person odds ratio interpretation, which suggests that the Speed factor reflects the ability to maintain one's maximal performance level under speeded conditions. Because it reflects the ability to maintain one's maximal performance level under speeded conditions, Speed factor scores cannot be interpreted in isolation but must be interpreted relative to a person's maximal performance Level factor score.

Although most factors are interpreted in terms of content common to items loading on the factor (e.g., verbal, numerical, and spatial), the Speed factor must be interpreted in terms of an administration condition manipulation common to all experimenter-paced items. The results of this study suggest that individual differences factors or dimensions can be random effects associated with experimental manipulations and that mixed-effects models may be useful tools for studying such factors.

Acknowledgment

The authors' thank two anonymous reviewers and the acting editor, Allan Cohen, for their helpful suggestions in the improvement of this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a United States Army Research Institute for the Behavioral and Social Sciences contract award number 1435-04-03-CT-74083 (Mark L. Davison) and by Grant No. R305C050059 from the Institute of Education Sciences in the U.S. Department of Education. The rights of research participants were protected and applicable human research guidelines were followed.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement, 29*, 357-368.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis, 55*, 12-25.

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement, 5*, 411-427.

Davison, M. L., Kim, S.-K., & Close, C. (2009). Factor analytic modeling of within person variation in score profiles. *Multivariate Behavioral Research, 44*, 668-687.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533-559.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models*. New York, NY: Springer.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nihoff.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331-358.

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*, 499-518.

Kim, S.-K., Davison, M. L., & Frisby, C. L. (2007). Confirmatory factor analysis and profile analysis via multidimensional scaling. *Multivariate Behavioral Research, 42*, 1-32.

Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of the correlational literature* (Technical Report No. 8). Stanford, CA: Stanford University, Aptitude Research Project, School of Education. (NTIS No. AD-A108–003).

Lord, F. M. (1956). A study of speed factors in tests and academic grades. *Psychometrika, 21*, 31-50.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Peterson, N. G. (1993). *Review of issues associated with speededness of GATB tests*. Washington, DC: American Institutes for Research.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Thissen, D. M. (1976/1977). *Incorporating response latencies in latent trait estimation* (Doctoral dissertation, University of Chicago, IL). Dissertation Abstracts International, 37, 4658B. (University Microfilms No. T-26116).

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

Van Breukelen, G. J. P. (2005). Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika, 70*, 359-376.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181-204.

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

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's ''Reliability of Speeded Number-Right Multiple-Choice Tests.'' *Applied Psychological Measurement, 33*, 488-490.