Measuring Speed of Numerical Reasoning

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Lan Huang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF ARTS

Mark L. Davison

December 2010

© Lan Huang 2010

ACKNOWLEDGEMENTS

I would like to thank Mark Davison, Michael Rodriguez, David Weiss, and Robert delMas for their teaching and mentoring, enormous expert knowledge and good conversations. Particularly, I want to thank my adviser Mark Davison for his incredible supports and patience to walk me through this work.

I would like to thank Robert Semmes for allowing me to use his data.

I would like to thank my colleague, Catherine Close, for her great help at the early stage of this project.

I would like to thank my best friend Yu Chen in Finland, who just has finished her degree in Computer Science this summer, for her reminders and encouragements every day. With this "overseas support network" we built together, I was able to set aside all the distractions, sometimes even frustrations, and to concentrate on writing.

I would like to thank my parents, Yuefeng Gong and Jiujia Huang who have been loving and supporting me unconditionally since I was born. I must thank my parents for their effort to help me become an educated, responsible and compassionate person. What is more important, they taught me how to be a happy person. The secret is to listen, to share and to give.

Last but not least, I would like to thank Wei Zhang, who has been kind and loving since the first day we met. His love and support without any complaints has enabled me to complete this thesis.

ABSTRACT

If numerical reasoning items are administered with time limits, will two dimensions be required to account for the responses, a numerical ability dimension and a speed dimension? If we want to know how quickly a person solves a problem, how can we obtain a reliable measure of speed? This study reanalyzed the data collected by Semmes, Davison, & Close (2009) in which one hundred and eighty-one college students answered 74 numerical reasoning items. Every item was administered with and without a time limit by half of the students. Three two-dimensional models were fit to item responses under self-paced and experimenter-paced conditions and response times under self-paced administrations. The best fitting model suggested that, other than the Level dimension, a second Speed dimension was needed to account for variation in numerical reasoning performance under experimenter-paced administration. After adding response time to the model, we saw a significant increase in the reliability estimate for the Speed factor compared to prior research with the same data, but estimating speed scores using only the experimenter-paced responses (Semmes et al., 2009). The validity of the Speed dimension was supported by its unique contribution to the prediction of ACT scores after controlling for the variation accounted for by the Level dimension. An alternative method of measuring Speed is mentioned. Some previous research using response times for other purposes besides measurement of speed are also discussed.

TABLES OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

"Other things being equal, the more quickly a person produces the correct response, the greater is his intelligence."

Edward L. Thorndike et al. (1926, p.24)

Imagine that two test takers A and B are trying to solve one math problem at the same time. It turns out that both of them arrive at the correct answer, but person A finishes it much earlier than person B. Intuitively, would we evaluate them as having the same math ability? Some people may disagree. They may even think it is justifiable to give person A more credit since quickness of performance is considered to be another valuable characteristic in the real world over and above accuracy. In fact, for some occupations particularly, accurate and immediate responses are very important, e.g., the soldiers on battlefields have to respond accurately and rapidly to volumes of information. Thus, the Army needs to develop a system to identify personnel who can complete difficult tasks both accurately and quickly (Davison, Semmes, & Close, 2009).

From a theoretical perspective, test theorists are always intrigued by the relationship between Speed and Ability. This is because in real world testing, most tests are administered with time limits for convenience or for the reason that measurement of Speed is part of the goal. In the former case, when tests are not pure power tests, the unidimensionality assumption of item response theory (IRT) might be violated and thus it may distort the estimates of the trait we intend to measure. Then controlling the effects

introduced by speededness becomes a problem. In the case of measuring speed, scoring methods and the quality of such speed measures need to be investigated. In the rest of this paper, some prior psychometric studies on speededness are reviewed in a framework that Davison, Semmes, and Close (2009) have proposed; then a modeling strategy is described.

Linear Factor Analysis Framework

"Is speed on cognitive tests a unitary trait? If so, how highly correlated are speed and level on the same task? How do various criteria relate to speed?" Frederic Lord asked this series of questions in 1956. He applied maximum likelihood factor analysis to scores from seven tests of different levels of speededness in three task domains (N = 649). He identified not just one "Speed" factor, but three of them, two different speed dimensions for two particular cognitive domains, verbal and spatial, and a second-order general speed factor. Also, small positive correlations between students' academic grades and their speed factor scores were found. However, the study left some unanswered questions.

First, no speed factor for the arithmetic-reasoning tests could be found. Second, for each test, the time limit was imposed on the test as a whole, which may cause severe item interdependence. Lord noted that in the highly speeded tests, examinees could not reach all the items. Finally, he used multiple-choice items which made things more complicated due to possible rapid random guessing. Lord's study is quite influential for his experimental design- the use of both unspeeded and speeded tests- and his conclusion that the speed factor may be domain/task specific.

Item Response Theory Framework

"An implicit assumption of all commonly used item response models is that the tests to which the models are fit are not administered under speed conditions…When speed affects test performance, then at least two traits are impacting on test performance: speed of performance, and the trait measured by the test content."

Hambleton and Swaminathan (1985, p.30)

In the item response theory framework, researchers are able to study accuracy and response times (RTs) at the item level. The use of time limits on each item has solved the problem of item interdependency.

Thurstone (1937) was the first to address the relationship between responses and RTs from an IRT perspective. He considered ability as a power factor that is independent of speed. The ability of an examinee is defined as the difficulty of the item for which his probability of obtaining the correct answer is 0.50 given infinite time. Speed is usually defined as the number of easy tasks that are completed in one unit time. However, as he admitted, "…a faster working speed with difficult tasks is more socially valuable." (Thurstone, 1937). In his model, for a fixed person, a decrease in the probability of correct response is associated with an increase of the difficulty, but the probability of success increases as the given time increases. He believed that items always have a speed aspect and a power aspect. Two assumptions of Thurstone's hypothesis seem to need more reflection (van der Linden, 2009): First, he treated item responses as a random variable, but not RTs. If accuracy and RT are both indicative of the same cognitive process, both of them should be treated as random variables. Second, the probability of a

correct response is explained by a person parameter and an item parameter while RT is not explained. Under what condition can RTs be independent of person and item parameters? One of the possible conditions may be when the items are extremely easy. Otherwise, for items with non-trivial difficulty, we would expect person and item effects for RTs. Thus, both person and item parameters should be taken into consideration when modeling accuracy and RTs.

In 1983, Thissen used an IRT model for item analysis and test scoring in timed testing, a model which is a revision of the model proposed by Furneaux (1961). This model was applied to three tests: a verbal analogies test, the Progressive Matrices Test, and a test of spatial ability. Both item responses and the RTs were treated as dependent variables and they were modeled separately. For the item responses, a 2-parameter logistic model was used which was very similar to the conventional 2PL IRT model today. For the RTs, a linear model of the logarithm of RTs was proposed:

$$\log(t_{ij}) = v + s_j + u_i - bz_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

$$z_{ij} = a_i(\theta_j - b_i) \tag{2}$$

where v is the mean log(RT), s_j is the person slowness (negative speed) and u_i is the item slowness, z_{ij} is the weighted distance between the person's ability location and the item location, and b is a regression coefficient reflecting the relationship between ability and easiness with RTs. Item response has been found to be systematically related to item features, such as sentence structure complexity in reading or calculation workload in numeric reasoning. Thus, two items can have the same difficulty but different time intensity (van der Linden, 2009). Thissen's model describes a negative relationship

between RTs and ability. For a fixed item, the higher the ability level, the shorter the response time it takes to solve the problem. However, the relationship (across persons and items) between ability and slowness is still unclear. It seems that these two person parameters have different effects on different types of tasks. An example would be the spatial test data where people do mental rotations; the estimated correlation between person slowness and effective ability was .82. In the Progressive Matrices tests, given enough time, people could get most items correct. Therefore, individual differences are actually differences in person slowness. In the Verbal Analogies tests, however, "both ability and speed will absorb part of ability defined by the speed test." (Thissen, 1983).

For the RTs part, the log(RT) model is one of several models that have been used in later studies (Schinipke & Scrams, 1997; van der Linden, Scrams, & Schnipke, 1999). In an empirical study with a sample size of 1104 and the RTs of 94 items, the lognormal model showed a good fit to the RT data (van der Linden et al., 2007). The distributions of RTs tended to be positively skewed because it is naturally bounded by zero and it is always possible to spend more time on an item. So the log transformation will make the distribution more symmetric (van der Linden, 2009). However, Thissen did not compare his model to alternative models.

An early model that tries to incorporate RTs into the item response model is Roskam's model (1997). He made a strong assumption that is also assumed in some other studies (Thurstone, 1937; White, 1982; Roskam, 1999; Van Breukelen, 1989) that the probability of a correct response increases with response time, that is, given infinite time, any item can be answered correctly. The increasing rate is defined as mental speed. He

redefined the person parameter in a Rasch model as the "effective ability parameter" θ_j which equals mental speed τ_j times processing time t_{ij},

$$\theta_j = \tau_j \times t_{ij} \quad (3)$$

If they are put on an exponential scale, then τ_j is rescaled to $\ln(\tau_j)$,

$$\ln(\theta_j) = \ln(\tau_j) + \ln(t_{ij}) \quad (4)$$

His model can be written as

$$P_i(\tau_j) = \left\{1 + \exp\left[-\left(\tau_j + ln_{t_{ij}} - b_i\right)\right]\right\}^{-1} \quad (5)$$

where t_{ij} is RT and b_i is the item difficulty. This model captures the speed-accuracy tradeoff: an increase in time is associated with a better chance of a correct response. Verhelst et al. (1997) proposed a very similar model to Roskam's except that their model contains a shape parameter. Also, they assumed a generalized extreme-value distribution for the latent response variable conditional on RT and a gamma distribution for the marginal distribution of RTs.

Besides separate models of RTs and accuracy, Wang and Hanson (2005) proposed a model called the Four Parameter Logistic Response Time model (4PLRT) that incorporates response time into a conventional 3PL model that can be used to model speeded tests with non-trivial difficulty:

$$P\left(x_{ij} = 1 \middle| \theta_j, \rho_j, a_i, b_i, c_i, d_i, t_{ij}\right) = c_i + \frac{1-c_i}{1+e^{-1.7a_i[\theta_j - (\rho_j d_i / t_{ij}) - b_i]}} \quad (6)$$

where a_i, b_i, c_i, and θ_j are the usual parameters in a 3PL model. Parameter d_i is the item slowness parameter, t_{ij} is the RT of this examinee on this particular item, and ρ_j is an examinee slowness parameter. $\rho_j d_i / t_{ij}$ implies that the product of a particular person

and item slowness parameter determines the rate of increase in probability of a correct response as a function of response time. When response time t_{ij} goes to infinity, this model reduces to the conventional 3PL model. Wang and Hanson fit this model and the conventional 3PL model to simulated data exceeding 1,000 simulees. The simulation results showed that the correlation (.937) between the 4PLRT θ estimates and the true θs is higher than that between 3PL θ estimates (.884) when the generated data fit the 4PLRT model. But no further model fit indices were provided to show that it is worth the effort to use the more complex model. Moreover, this model requires a restrictive assumption that RT distributions must be independent of the person parameters. It treats RT as a fixed predictor rather than a random variable, which is determined by the test giver, but not applicable in the situation where examinees can vary in their RTs.

In 2005, Van Breukelen proposed a mixed-effects regression model for RTs and accuracy respectively. In his experiment, the mental rotation task of Shepard and Metzler (1971) was used, in which the participant has to decide whether two objects with different orientations are isomorphic or not and their responses and RTs were recorded. The ln(RT)was treated as the sum of a random person effect, two fixed item effects and an error term.

$$ln(RT_{ij}) = \gamma_{0j} + \gamma_1 diff_{ij} + \gamma_2 angle_{ij} + e_{ij} \quad (7)$$

For the response accuracy, the model assumes that the log odds of a correct response to a given item is the sum of a random person parameter and two fixed item effects, same vs. different and the angle of rotation needed.

$$ln\left(\frac{p_{ij}}{q_{ij}}\right) = \beta_{0j} + \beta_1 diff_{ij} + \beta_2 angle_{ij} \quad (8)$$

The IRT model and the RT model are linked by assuming that the probability of a correct response is an increasing function of RT for person *j* on item *i* which assumes a within-person correlation between RTs and accuracy.

Van Breukelen fit 13 different models to each of response accuracy and ln(RT) data. For log odds, both best-fit models have a random person effect, a random same vs. different status effect and a fixed effect- angle of rotation. For ln(RT), both best-fit models have a person effect, a random angle of rotation effect, and a fixed same vs. different status effect. In addition, he found no correlation between the two person effects in the log odds model and ln(RT) for each pair of best-fit models. As he acknowledged, this finding might be due to his small sample size (N=30). However, as Davison et al. (2009) observed, the results are not sufficient to support "the existence of distinct speed and level abilities" in Shepard-Metzler's mental rotation task. First, in this experiment, participants were asked to work as quickly as possible while maintaining accuracy. This instruction could not guarantee truly self-paced performance. Second, Van Breukelen did not take into account that respondents have a 50% chance of getting the correct answer- same vs. different.

In a recent article by van der Linden (2009), there is a nice summary of the basic issues of which we should be aware when modeling RTs, among which five of the six conclusions are related to our study:

1. RTs on test items should be treated as realizations of random variables.

2. RT models should have a person parameter as well as an item parameter for their "time intensity".

3. The speed-accuracy tradeoff is a within-person phenomenon. For a fixed examinee, ability "displayed" is a monotonically decreasing function of speed. This can only be checked by letting one single examinee perform at different rates and check his or her accuracy, which is obviously different from the designs we described above. "The response models with a single ability parameter can fit only when the person operates at a constant speed during the test." (Van der Linden, 2009). Thus, we do not need to model the speed-accuracy tradeoff. Models of response accuracy and RTs need fixed parameters for both Speed and Ability.

4. An item difficulty parameter is needed in the response accuracy model while a time intensity parameter is needed in the RT model, that are two different conceptions.

5. Local independence of responses to different items should be assumed. Likewise, it seems to be reasonable to assume local independence between RTs on different items.

In fact, in van der Linden's modeling framework, he has two lower-level models for each examinee and two higher-level models to explain observed correlations between accuracy and RTs across persons and items. The higher-level models are ignored here.

Van der Linden (2009) derived what he called the "fundamental equation" which is saying that for person j, Speed τ_j can be measured as amount of labor needed to solve item i (β_i), divided by $RT_{ij.}$. From this, the following equation can be derived:

$$t_{ij} = \frac{\beta_i}{\tau_j} \qquad (9)$$

As other researchers did, he took a natural logarithm transformation of RT, which gives

$$\ln t_{ij} = \beta_i - \tau_j \quad (10)$$

$(\beta_i - \tau_j)$ is fixed for every person item combination. But RT is a random variable so the equation should be:

$$E(\ln t_{ij}) = \beta_i - \tau_j \quad (11)$$

Then an extension was added to the model which assumes a normal distribution of the RTs around $E(\ln t_{ij})$. So the final equation turned into the following:

$$\ln t_{ij} = \beta_i - \tau_j + e_i, \;\; e_i \sim N(0, \alpha_i^{-2}) \quad (12)$$

and its lognormal density for the distribution of t_{ij} is:

$$f(t_{ij}, \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\{-\tfrac{1}{2}[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2\} \quad (13)$$

where α_i is the reciprocal of the standard deviation of the RTs on item *i* and it can be interpreted as its discrimination parameter.

The model was fit to a real dataset which came from a large scale computerized examination with 96 items of 1,104 test takers (van der Linden & Glas, 2010). The results suggest that, the two local independence assumptions were quite plausible except for the violation of local independence of responses for one item.

Research has shown that the correlation between Speed and Ability can be positive or negative (van der Linden, Scrams, & Schnipke, 1999; van der Linden et al., 2007; Klein Entink et al., 2009). A plausible explanation provided by van der Linden is that more able examinees have better time-management skills. When the time limit is stringent, they know how to distribute time properly and they can speed up to finish the item in time (resulting in positive correlations). However, when the time limit is lenient,

they may also choose to take their time to maximize their performance (resulting in negative correlations). The sign of the correlation probably depends on the type of test and test conditions. “Sometimes ‘hard work’ will pay off, but sometimes “take your time is the best advice.” (van der Linden & Glas, 2009).

CHAPTER TWO
OUR MODELING FRAMEWORK

This is a reanalysis of data collected by Davison et al. (2009). In their experiments, each examinee took two parallel tests- one was self-paced and the other was experimenter-paced. In the self-paced conditions, examinees were told to take their time in order to maximize their performance. Both responses and RTs were recorded for every item. In the experimenter-paced conditions, a time limit was set for each item based on pilot study data and examinees were told if no answer were given before the time limit expired, this item would be scored as incorrect. Previous analyses of the data have included only the experimenter- and self-paced responses, but not response times. The analyses have provided some support for the hypothesis that two dimensions underlie self-paced and experimenter-paced responses: a Level and a Speed dimension. Using just the experimenter- and self-paced accuracy variables, however, the Speed dimension was not reliably measured. Some evidence for the validity of the Speed dimension did emerge from these analyses despite the modest reliability (Semmes et al., 2009).

The two goals of this study are to further investigate empirically whether a speeded test with non-trivial item difficulties has two dimensions- Level and Speed- and whether one can reliably measure Speed by using a combination of experimenter-paced items and self-paced response times.

In this thesis, "Level" will be used as a synonym for a very specific kind of ability, not general intelligence (Thorndike et al., 1926). Here Level refers to the person's mathematical reasoning ability level given unlimited time.

Level-only Hypothesis and Speed-only Hypothesis

Our first two basic hypotheses state that under self-paced conditions (unlimited time), a single factor Level can account for the variance of response accuracy. Under self-paced conditions, for a fixed person, we can assume that he/she can work at the pace he/she wants. Once speed is fixed, his/her "effective" ability is constant. Letting *j* index persons and *i* index items, the non-linear factor model for the probability of a correct response on item *i* by person *j* under self-paced conditions can be expressed as:

$$\pi_{ij(s)} = \frac{\exp(\lambda_{1i}F_{1j} - \tau_{1i})}{1+\exp(\lambda_{1i}F_{1j} - \tau_{1i})} \quad (14)$$

where $\pi_{ij(s)}$ is the probability of a correct response under **s**elf-paced conditions, λ_{1i} is the factor loading of item *i* on Factor 1 (Level) and τ_{1i} is the intercept.

Similarly, a single factor Speed factor can account for the variance of RTs under self-paced conditions. A factor model of log RT on item *i* by person *j* is proposed as the following:

$$E\big(Z\mathrm{Ln}(RT_{ij})\big) = \lambda_{2i}F_{2j} + \mu_i \quad (15)$$

The left side of Equation (15) is the expectation of the standardized ln(RT) on item *i* by person *j*. λ_{2i} is the factor loading of item *i* on Factor 2 (Speed) and μ_i is the intercept. After standardizing ln(RT), the scales of RTs and item responses were more comparable, thus the Speed-Level model described in the next section could be identified.

Speed-Level Hypothesis

Unlike under self-paced conditions, when every item has an imposed time limit, some examinees may need to choose to work at a new speed in order to finish the item as quickly as possible; their maximal ability, Level, may not be fully displayed; in other

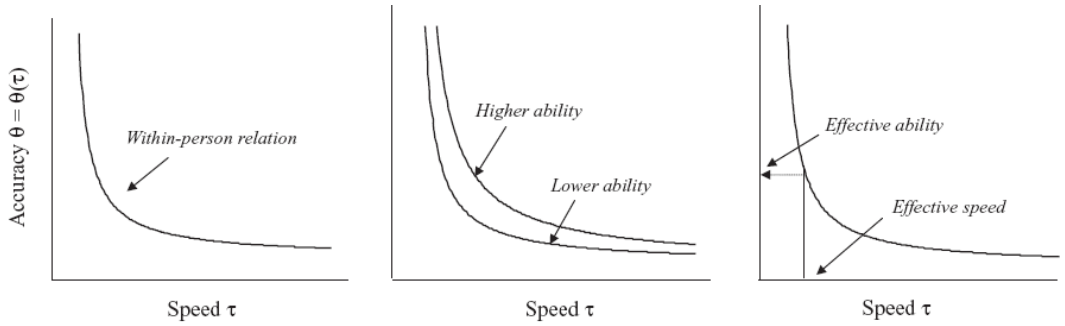

FIGURE 1. Examples of ability as a monotonically decreasing function of speed τ (left panel), ability functions for a more and a less able test taker (middle panel), and effective speed and ability of a test taker during the administration of a test (right panel).(van der Linden, 2009)

words, the person may have to make a speed-accuracy tradeoff. Thus we can consider the ability displayed under speeded conditions as "effective ability" (van der Linden, 2009). Some test takers can finish item *i* within time T_{ij} under self-paced conditions. If T_{ij} > Time Limit$_i$ and the test taker cannot speed up to a degree that he/she can finish item *i* within Time Limit$_i$, he/she may fail to produce a correct response, resulting in an estimate of his/her effective ability lower than his/her maximal ability Level. For those more able test takers who can still finish most items under time limits, the discrepancies between their ability and effective ability will not be as large as those of the less able test takers who need more time. Thus, according to our model, the accuracy in a speeded power test is the manifestation of two traits- Level and Speed. Figure 1 is an illustration of the relationship between effective level and speed within and across persons. In the middle panel, we can see that the accuracy of a lower ability test taker decreases faster than that of a higher ability test taker over the full range of speed.

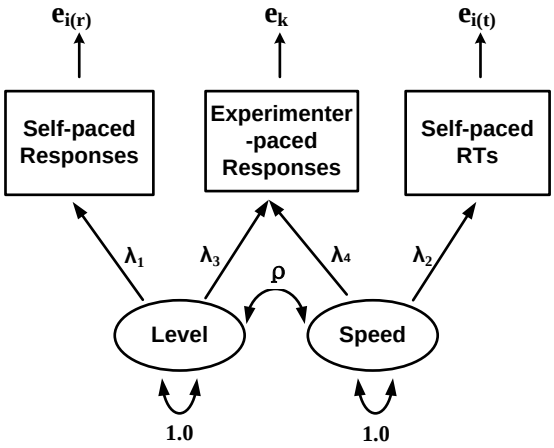

FIGURE 2. The Speed-Level hypothesis (Model 3)

Thus, according to our Speed-Level hypothesis, under the experimenter-paced condition, two factors-Speed and Level are needed to account for the variance of response accuracy:

$$\pi_{ij(E)} = \frac{\exp(\lambda_{3i}F_{1j}+\lambda_{4i}F_{2j}-\tau_{2i})}{1+\exp(\lambda_{3i}F_{1j}+\lambda_{4i}F_{2j}-\tau_{2i})} \qquad (16)$$

$\pi_{ij(E)}$ is the probability of a correct response on item *i* by person *j* under experimenter-paced conditions. λ_{3i} is the factor loading of item *i* on Factor 1 (Level). Likewise, λ_{4i} is the factor loading of item *i* on Factor 2 (Speed). τ_{2i} is the intercept. Figure 2 is an illustration of this model.

Again, our major hypothesis is that a single Level dimension can account for variation of accuracy on self-paced items; however, a second dimension, Speed, is needed to account for the individual differences in performance on experimenter-paced items. In addition, after controlling for the time intensity of items, RTs under self-paced

condition are the manifestation of a Speed dimension. If we want to obtain a reliable measure of Speed, RTs should also be treated as a response variable of the model. In the next section, the data collection process is described.

CHAPTER THREE

METHODS

This is a reanalysis of data collected by Davison and Semmes (Davison et al., 2009; Semmes et al., 2009). The method section below describes their data collection.

Participants

Participants were University of Minnesota college students enrolled in Psychology Department courses who earned extra credits for their participation. Students took two forms of tests, a first form under self-paced administration, and a second form under experimenter-paced administration. To balance the order effect that might occur, participants were divided into two subgroups. Sample VW took Form V under self-paced conditions, then Form W under experimenter-paced conditions. Sample WV took Form W under self-paced conditions, then Form V under experimenter-paced conditions.

Participants were recruited and tested from February 2006 through July 2006. They took tests under both self-paced and experimenter-paced conditions. For participating in the self-paced testing, each student received only extra credit points except those who enrolled during the summer session who were additionally awarded $5.00. When they completed the second session of the experimenter-paced testing, they received both extra credit points and $20.00. The $20.00 provided a strong incentive for students to come back for the second session and to be motivated in a math test with time limits. Participants were randomly assigned to samples VW and WV. Table 1 summarizes each sample's demographic characteristics and ACT Math scores. The result

Table 1

Summary of Demographic Characteristics and ACT Math Scores in Samples VW and WV

Variable	Subsample VW	Subsample WV
Sample size	93	94
Percent women	60.2	37.2
Median age	20	20
Lowest age	18	18
Highest age	43	47
Percent Native 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

Semmes et al. (2009)

Note. ACT Math scores are reported on a scale ranging from zero 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.

showed that the two samples, VW and WV, were similar except their percentages of women: 60% in sample VW and 37% in sample WV.

Due to our small sample size (N=181), we have to set constraints on the model described above so we do not run out of degrees of freedom. Therefore, for each dimension, the discrimination parameters for all the items were set equal if they were administered under the same condition. For the self-paced tests, two factor loadings were

estimated, λ_1- of the Level dimension and λ_2- of the Speed dimension; for the experimenter-paced tests, those are λ_3 and λ_4.

Measures

First, to create the tests, seven items were selected from the Differential Aptitude Tests (DAT) (Benett, Seashore, and Wesman, 1982, 1990). Then the researchers selected 69 SAT quantitative items and 16 GRE quantitative items from published editions of the SAT (College Board, 2003) and GRE (Educational Testing Service, 2002). Because only a small percentage of Midwestern high school students take the SAT, the researchers thought that few of them would have seen these items or remembered them. To prevent guessing and simplify the models, all the multiple-choice items were changed to a constructed-response format.

Finally, 92 numerical reasoning items were chosen for this study. Form V and Form W were composed of 40 items each. Based on the item properties provided, these two forms were created to be similar in content and difficulty. The remaining 12 items were assembled into Form C, which was designed as a linking test for the purpose of equating Form V and Form W.

Setting Time Limits

For the experimenter-paced testing, a time limit was set for each item using data from a pilot study. This was a very important decision to make because if the time limit is too lenient, one might not be able to discriminate among those who could answer the item correctly under self-paced administrations; if it is too stringent, it might discourage the examinees early and affect their performance through the whole test. Intuitively, the ideal

situation would be, for those who can answer the item correctly under self-paced conditions, half of them would get the item correct when the time limit was imposed, and half of them would not. For each item, the researchers computed the median response time among the examinees that got it correct without time limits from the pilot study. Then they rounded it to the nearest five seconds. The rounded median response time then became the time limit for the timed testing. However, as data will show, the time limits were not as discriminating as desired; more than half the proportion passing without limits answered items correctly in timed testing. For Form V self-paced, the proportion correct was .68. When administered in the experimenter-paced format, the proportion dropped to .49. For Form W, from unspeeded condition to speeded condition, the proportion correct dropped from .63 to .47 (Davison et al., 2009).

Apparatus

Testing stations were installed in each of the five assessment rooms. In each room, there was a long table, a chair, and six types of equipment on the table: (a) two personal computers placed side by side, (b) a speaker, (c) a five button response box, (d) a microphone, and (e) a null modern cable connecting the two PCs.

A participant was seated in front of the PC located on the left side of the table. On this PC, A software program called E-Prime was used to control the item display and RT recording process (Schneider, Eschman, & Zuccolotto, 2002). This will be called the E-Prime PC. The response box was connected to the E-Prime PC and was placed where the keyboard would normally reside. A speaker was also connected to this PC and was positioned toward the rear of the table on the same side of the response box as the

microphone. If the participant was right-handed, the microphone and speaker were placed to the left of the response box, and, if the participant was left-handed they were placed to the right of the response box. The speaker was used to announce each item number before that item was displayed on the E-Prime PC's screen. The microphone was to record the item number announced by the speaker and the participant's answers to the items.

The microphone was connected to the second PC located on the right side of the table. The LabVIEW software was used in this PC to record the participant's answers in a digital file (National Instruments Corporation, 2003). This PC will be called the LabVIEW PC in the descriptions below.

Procedure

As stated earlier, examinees participated in both the self-paced testing (Session 1) and experimenter-paced testing (Session 2). The self-paced session always preceded the participant's experimenter-paced session, and Session 2 always occurred at least one day after the participant's Session 1. The longest interval between Session 1 and Session 2 for a participant was 16 days. Sample VW took only Form V in Session 1, while sample WV took only Form W in Session 1. Then in Session 2, sample VW took the easiest 48 items obtained by merging Form W and Form C. Sample WV took the easiest 48 items obtained by merging Form V and Form C.

Session 1's Procedure: Self-paced Testing

First the researcher told the participants that the purpose was to find out which type of test, self-paced or experimenter-paced, provided better information about

examinees. Then the test administrator described the experimental apparatus. The participants were told that the questions would be presented one at a time and that there would be no time limits for them to answer any question or for them to finish the entire test. The test administrator told the students to take their time in order to maximize their test performance. They were also told that questions would be presented in random order with respect to difficulty, and they could not return to any items that had been answered. Blank paper and two pencils would be provided but no calculator was allowed. However, an answer sheet with 52 numbered entries would also be given to each of them on which they could write down revised answers to previously answered questions at any time. The participants were asked to turn off their cell phones in the testing rooms to prevent distraction.

After the students were seated in the testing rooms, they began to enter their demographic information, including gender, age, handedness, college of enrollment, college test taken, and whether English was their second language. Next, a series of practice items was provided for them to familiarize themselves with the apparatus. When the test administrator was satisfied with the participant's use of the testing apparatus, the testing session began.

In Session 1, the E-Prime program was written to randomize the order of the items presented for each participant. The sequence of events for administering items and recording responses was as follows. The presence of each item was preceded by a display of a pause screen that told the examinees that they could rest in the testing room, use a restroom or leave the room to get a drink. When the participant was ready to proceed, he

or she could press the "Next" button on the response box. After pressing that button, a signal would be sent to the PC to enter sound recording mode for three seconds. The E-Prime PC speaker would announce the item sequence number, which was recorded by the LabVIEW PC to label the participant's spoken item response. Next, a test item would appear on the E-Prime PC's screen with the start of the response time clock; the clock was not shown to the examinees. Students were not told that their self-paced item response times would be recorded.

When the examinee solved the problem, he or she would press the "Answer" button and then the test item would vanish from the screen, the response time clock would stop counting and the LabVIEW PC would enter sound recording mode for five seconds to record the spoken item response. After five seconds, a pause screen would appear on the E-Prime PC's screen. After the final question was administered, a message would appear to notify the participant the test was over.

Session 2's Procedure: Experimenter-paced Testing

The test administrator told the participants that each item would be administered with a time limit and if no answer were given before the time limit expired, this item would be scored as incorrect. Other conditions were kept the same as in Session 1. Before the actual testing began, students answered eight practice questions to help them become more familiar with the experimenter-paced condition.

The procedure for administering the test in Session 2 was quite similar to that in Session 1, with two important differences. First, the display of each item was preceded by a screen stating the time limit for answering this item but no countdown clock appeared.

Second, there was no “pause screen” for the examinees to control the time between items. As soon as a participant’s answer to a given item was recorded, the time limit screen for the next item appeared.

Test items were administered in order from low difficulty to high difficulty that was determined by the frequency of incorrect responses given by the pilot study. The rationale for this order was to reduce the chance of discouraging the examinees early in their tests. It was hoped they could maximize their performance even under this limited time condition. If the time limit expired before the participant could give an answer, the E-Prime PC speaker would give out a bell-like sound, and the item would vanish from the screen. After the first 24 items were administered, participants were given a mandatory 10-minute break during which they could leave the testing room.

CHAPTER FOUR

RESULTS

Results of Model Fitting

All models were fitted using MPLUS version 5.0 (Muthen & Muthen, 1998-2007) with mean structure to include means, thresholds and intercepts in the model. Maximum likelihood method was applied to estimate the latent scores. To evaluate our hypotheses, first we assessed the two submodels of our proposed model. That was testing the two unidimensionality assumptions of the self-paced responses and the self-paced RTs. Then three two-dimensional models were fitted to the full data composed of self-paced responses, self-paced response times, and experimenter-paced responses including our hypothesized model and two alternative models in which experimenter paced responses loaded on the Level dimension or Speed dimension. Using the Bayesian information criterion (BIC), these models were compared. Recent research (e.g. Kang, Cohen, & Sung; Kang & Cohen, 2007) suggests that, of the various information statistics, the BIC more often identifies the best model among several competing psychometric models. The chi-square could not be computed for models involving categorical item responses because the contingency table was too large. Next, parallel forms reliabilities for both dimensions were estimated for the best fitting model. Finally, the validity of the Speed dimension was explored by regressing Level and Speed onto ACT Math scores.

Variables that were used in this analysis included item responses of 181 participants under both self-paced and experimenter-paced conditions, and standardized log response times (Zln(RT)) under self-paced administration. All the items were split

into two nearly equal blocks so that we would not run out of degrees of freedom with a small sample size (N=181). We assigned the first item to Block 1, the second item to Block 2, the third item to Block 1, the fourth item to Block 2, and so forth. A paired t-test was conducted between and average raw scores of Block 1 and Block 2. The results showed that there is no significant mean difference between the two blocks (mean difference= -.0046, p= .482). In essence, blocks 1 and 2 can be considered alternative forms. Item W17 in session 1 and Item W02 in session 2 were excluded from the analysis because everybody got these items correct.

Fit Indices to Evaluate Model Fit

1. Tucker-Lewis Index (TLI)

Proposed by Tucker and Lewis (1973), TLI is calculated as

$$TLI = \frac{\chi_b^2/df_b - \chi_{H_0}^2/df_{H_0}}{(\chi_b^2/df_b) - 1} \quad (17)$$

Where df_b and df_{H_0} are the degrees of freedom for the baseline and the hypothesized models respectively. TLI can exceed the 0 to 1 range. Hu and Bentler (1999) recommended a cutoff value of TLI close to .95.

2. Comparative Fit Index (CFI)

Bentler (1990) proposed CFI in 1988. CFI has the advantages of having a 0-1 range and smaller sampling variability. CFI is calculated as

$$CFI = 1 - \frac{\max\left[\left(\chi_{H_0}^2 - df_{H_0}\right), 0\right]}{\max\left[\left(\chi_{H_0}^2 - df_{H_0}\right), \left(\chi_b^2 - df_b\right), 0\right]} \quad (18)$$

Hu and Bentler (1999) recommended a cutoff value of CFI close to 0.95.

3. Root-mean-square Error of Approximation (RMSEA)

Proposed by Steiger and Lind (1980) and Browne and Cudeck (1993), the RMSEA for continuous outcomes is defined as

$$RMSEA = \sqrt{\max\left[\left(\frac{2F(\hat{\theta})}{d} - \frac{1}{N}\right), 0\right]} \quad (19)$$

F is the maximum likelihood fitting function and d is the degrees of freedom of the model. Browne and Cudeck (1993) suggested that RMSEA values less than 0.05 are indicating a close fit.

4. Akaike's information criterion (AIC)

Proposed by Akaike in 1974, AIC is calculated as

$$AIC = d + 2p \quad (20)$$

The first component d (deviance), is $-2 \times \log(maximum\ likelihood)$. A smaller deviance indicates a better fit. The second component 2p is intended as a penalty for the model complexity where p is the number of estimated parameters. The model with the smallest AIC should be selected.

5. Bayesian information criterion (BIC)

Proposed by Schwarz in 1978, BIC is calculated as

$$BIC = d + p\ (\log N) \quad (21)$$

N is the sample size. We can infer from the equation that BIC gives a higher penalty to the number of parameters. Similarly, the model with the smallest BIC should be selected.

Table 2: Fit Measures of the Unidimensional Model of Self-paced Responses

Fit Measures	Block 1	Block 2
# of free parameters	38	37
Loglikelihood H0 value	-1782.20	-1582.68
AIC	3640.40	3239.36
BIC	3761.95	3357.70

Assessing the Unidimensionality of Self-paced Responses and Self-paced RTs

1. Level-only Model

First we fitted a unidimensional model to all the self-paced responses. Our proposed model posits that a single dimension can account for the self-paced responses. Only self-paced responses were included in this analysis. The item discrimination was set equal for all the items so we would not run out of degrees of freedom. Table 2 shows the number of parameters and fit measures. The unidimensional model was fitted separately to the items of block 1 and block 2. Some of the usual fit statistics were not computed because with more than eight binary items and maximum likelihood estimation, chi-square and related fit statistics are not available because means, variances and covariances are not sufficient statistics for model estimation (Muthen, 2010). With a huge number of cells in the multi-way frequency table of binary items, we were bound to have many cells with small expected values so that the chi-square approximation is not entirely valid. So we checked the bivariate standardized residuals for every pair of items. These standardized residuals should be approximately normally distributed z-scores. They are computed as:

Table 3: Assessing the Unidimensionality of the Self-paced Response Times

Fit Measures	Block 1	Block 2
Chi-square (*df*)	436.32 (341)	363.49 (288)
# of free parameters	77	71
P-value	0.001	0.0017
CFI	0.87	0.85
TLI	0.87	0.85
Loglikelihood H_0 value	-4541.21	-4268.130
H_1 value	-4323.05	-4086.382
AIC	9236.413	8678.26
BIC	9482.697	8905.35
RMSEA	0.039	0.038

$$Z = \frac{O-E}{\sqrt{E}\times\sqrt{\frac{1-E}{n}}} \qquad (22)$$

where O and E are the observed and expected counts for a pattern in the categorical data. Thus at the 5% significance level, values exceeding +/- 1.96 suggest a poor fit. The results showed that in Block 1, only 5.7% of all the patterns showed misfits. In Block 2, 13.3% of the patterns showed misfits.

2. Assessing the Unidimensionality of Self-paced RTs:

Then we fitted a unidimensional model to the Zln(RT). Our proposed model posits that self-paced responses can be accounted for by a single dimension. Again, the item discrimination was set equal for all items. Table 3 shows the number of parameters

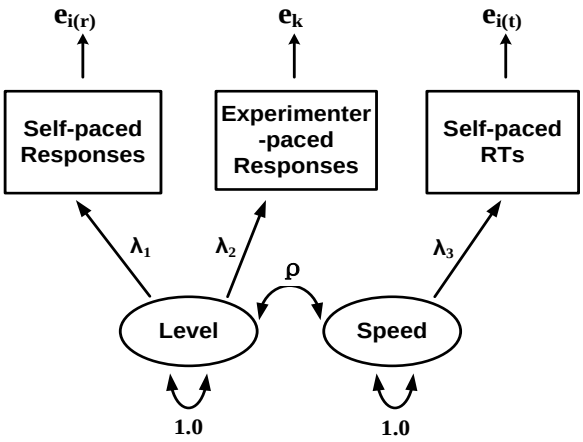

FIGURE 3: Model 1

and fit measures for the Speed-only model for Block 1 and Block 2. The chi-square measures for the Speed-only model were 436.32 (*df* =436.32, *p* =.001) and 363.49 (*df* = 288, *p* = .0017) respectively. The Tucker-Lewis Indices equaled .87 and .85; the RMSEA= .039 and .038, suggesting a close fit. Then factor scores of RTs were estimated for later investigation of the within-person relationship between RTs and Speed.

Two-dimensional Models

Self-paced and experimenter-paced responses and Z_{lnRT} were then analyzed together with different constraints for each model:

Model 1: The self-paced responses with equal factor loadings λ_1 and experimenter-paced responses with equal loadings λ_2 loaded on Level; the Z_{lnRT} loaded on Speed with equal loadings λ_3. See Figure 3 for illustration.

Model 2: The self-paced responses with equal factor loadings λ_1 loaded on Level; the experimenter-paced responses with equal loadings λ_2 and Z_{lnRT} with equal loadings λ_3 loaded on Speed . See Figure 4 for illustration.

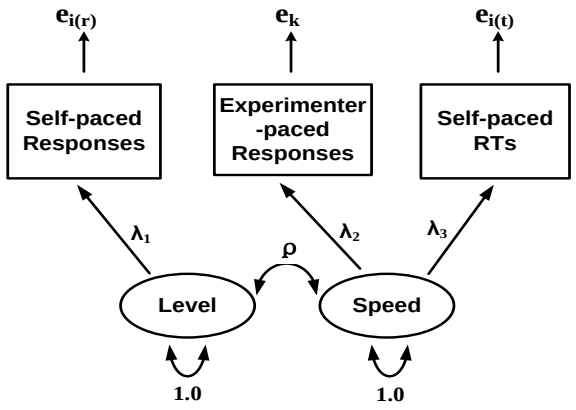

FIGURE 4: Model 2

Model 3: The experimenter-paced responses not only loaded on Level with equal loadings λ_3 but they also loaded on Speed with equal loading λ_4; the self-paced responses loaded only on Level with factor loading λ_1 and $Z_{\ln RT}$ loaded only on Speed with equal loading λ_2. See Figure 2 for illustration. It was our hypothesis that Experimenter-paced responses depend on both the Level and Speed factor; therefore Model 3 would fit better than either Models 1 or 2.

Table 4 shows the number of parameters, fit measures and the correlations between the two factors. The results indicated that Model 3 had the lowest BIC and the lowest AIC in both block 1 and block 2. We also conducted likelihood ratio tests between Model 1 and Model 3, and between Model 2 and Model 3. Table 5 shows that the log likelihood ratio differences were highly significant for each pair of models in both blocks, $p < .001$, suggesting Model 3 was the best fitting model. For this model, the Level and Speed dimensions were weakly correlated, r = .24 for both blocks. Figure 5 contains the scatter plots showing the relationship of the Level factor score and the Speed factor score

Table 4: Fit Measures of the Two-dimensional Models

Fit Measures	Block1			Block2		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
# of free parameters	159	159	160	152	152	153
Loglikelihood H_0 value	-8661.07	-8832.15	-8645.68	-8248.82	-8363.35	-8236.18
AIC	17640.15	17982.30	17611.36	16801.64	17030.69	16778.36
BIC	18148.71	18490.86	18123.12	17287.81	17516.86	17267.73
ρ(Level,Speed)	0.42**	0.73**	0.24**	0.39**	0.69**	0.24**

** indicates correlation is significant at .01 level (2-tailed)

Table 5: Log Likelihood Ratio Test Comparing Models 1-3

	Block 1		Block 2	
	M_3-M_1	M_3-M_2	M_3-M_1	M_3-M_2
-2 log likelihood ratio difference	30.78	372.94	25.28	254.34

Note. Each -2LLRD was tested against $\chi^2_{(1)}$ because the number of free parameters increased from 159 to 160 for block 1 and from 152 to 153 for block 2.

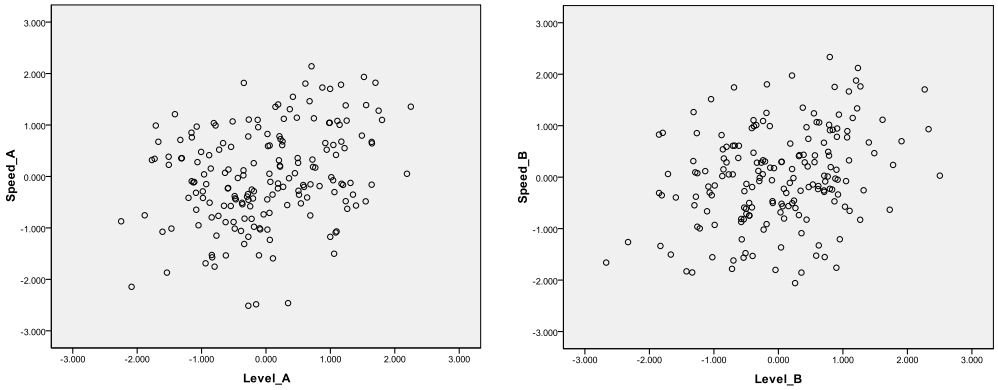

FIGURE 5: Scatter plots of the Level scores and the Speed scores of the Level-Speed Model

for each block. We can see a weak positive linear relationship between the Level scores and the Speed scores in both blocks. In addition, the correlation estimated from Model 3 was smallest among the three models. One plausible explanation is that by loading experimenter-paced item responses on both dimensions, we obtained purer measures of both dimensions, which are more relatively distinct by nature.

In each block, we computed two factor scores (IRT θ scores) for everybody, a Speed score and a Level score. Since blocks 1 and 2 constitute alternative forms, the correlations of the two Level scores constitute reliability estimates for the Level score. Likewise, the correlation of the two Speed scores provides an estimate of the reliability of the Speed score. For the Level dimension, the correlation between blocks 1 and 2 was 0.89. For the Speed dimension, the reliability estimate was 0.86.

Table 6: Correlations of Level, Speed, and ACT Scores

	Block1		Block2	
	Level	Speed	Level	Speed
ACT Math	.78**	.42*	.73**	.39**

** indicates correlation is significant at .01 level (2-tailed)

* indicates correlation is significant at .05level (2-tailed)

Correlations of Speed Scores with ACT scores

Table 6 shows the correlation of the Level and Speed scores with several criterion variables in our study. The focus here will be on the correlation of the speed scores. We expected there would be positive correlations between Speed scores and ACT mathematics scores but the correlation would not be very high because the ACT Mathematics test has a relatively lenient time limit. Averaging over the two blocks, the correlation between ACT Math score and Level Score was 0.78; while for the Speed factor, the correlation was 0.42.

Regression showed that, for block 1, Level accounted for 61.3% of the variation in ACT scores. Speed accounted for 2.7% of the variation after controlling for Level. The R^2 change was significant ($p < .05$). For block 2, Level accounted for 53.2% of the variation in ACT Math scores. Speed accounted for an extra 2.1% of the variation. The R^2 change was also significant ($p < .05$).

Within-person Relationship between RTs and Speed

The correlation between the factor scores from the RTs-only Model 1and the Speed factor scores from Model 3 was estimated for each block. For Block 1, r= -.967;

for block 2, r=-.979. These almost perfect correlations suggested that in this study, self-paced RTs are good reflections of Speed, but in a reciprocal way- the longer the RTs, the lower the Speed scores. The alternative forms reliability of the factor scores defined by RTs only was .85.

Gender differences were statistically significant for both Speed and Level scores. For block 1, males had higher mean Level scores on average than females (mean difference = .42 units, t (1, 178) = 3.02, p <.05). *Cohen's d* = .49 indicated a moderate effect size; on the Speed dimension, males also had higher average scores than females (mean difference = .37 units, t (1, 178) = 2.69, p <.05). *Cohen's d* was .43. Similarly, for block 2, on the Level dimension, males had higher scores on average than females (mean difference= .48, t (1, 178) = 3.16, p < .05). *Cohen's d* was .60; on the Speed dimension, males also had higher average scores than females (mean difference= .27, t (1, 178) = 2.00, p <.05). *Cohen's d* was .33. Previous researchers have shown that in mathematical reasoning tasks, males' performances usually have larger variance and males typically outperform females significantly among high-scoring individuals (Hedges & Nowell, 1995; Bielinski & Davison, 1998).

Whether the person is right-handed or left-handed seemed to have no association with any factor scores. Also, there was no significant difference between the native participants and ESL participants.

In summary, results suggested that, other than the Level dimension, a second Speed dimension was needed to account for variation in numerical reasoning under experimenter-paced administration. After including RTs in the model, we saw a

significant increase in the Speed reliability estimate compared to prior research with this same data, but estimating speed scores using only the experimenter-paced responses (Semmes et al., 2009). The validity of the Speed dimension was supported by its unique contribution to the prediction of ACT scores after controlling for the variation accounted for by the Level dimension.

CHAPTER FIVE

CONCLUSION AND DISCUSSION

The two goals of our study seemed to be accomplished: First, we added one more piece of evidence that a speeded test with non-trivial difficulty has two dimensions- Level and Speed; second, using three sources of information (self-paced responses, self-paced RTs and experimenter-paced responses), a reliable measure of Speed was obtained successfully. The correlations of Speed and Level across examinees were small, suggesting that in this particular task, these two factors are relatively distinct.

As mentioned before, this is a reanalysis of the data. Earlier, the researchers proposed and tested a Level-Speed model using the same data (minus the RT variables) with an HLM approach (Semmes et al., 2009). They ended up with two random effects for each examinee- a Level random effect and a Speed random effect. However, RTs were not included in their analysis. The reliability estimates for the Level and Speed effects were .87 and .39 respectively. They concluded that the Speed reliability was very low and not suitable for high stakes application. We correlated our speed factor scores with their speed random effects and obtained a correlation around .5. It seems that these two factor scores, although both are named "Speed", are not capturing exactly the same information. We believe that the inclusion of RTs in our current analysis would be a more hopeful approach if we believe that Speed is a relatively stable trait.

Our results also suggested that RTs are almost perfect manifestations of Speed in this study. Does this suggest a simpler way to measure Speed? For example, under certain circumstances, we can obtain a reliable measure of Speed through recording participants'

RTs under self-paced conditions, just like many reaction time studies in psychology. However, as van der Linden (2009) argued that RT and Speed are not equivalent because the time intensity of an item should be considered, thus, whether RTs are a perfect reflection of Speed may depend on the properties of items (e.g., content and psychometric properties) and test administrations (e.g., test instructions).

In real world testing, where under most conditions tests have to be administered with time limits for convenience or special purposes, more thought should be given to time limit setting because time limits have different effects on different cognitive tasks (Morrison, 1960).

Besides the measurement of Speed, RTs provide us other useful information as well. It can help us to improve the design of adaptive tests by selecting the items using the examinee's RTs on the previous items in addition to item responses (van der Linden, 2009). Schinipke and Scrams (1997) developed a model using RTs to detect rapid guessing. Van der Linden et al. (1999) proposed an item-selection algorithm to reduce the speededness of the test for those who would otherwise suffer from time limits in computerized adaptive testing. RT information can even be used in personality assessment. Ferrando et al. (2007) applied a model that incorporates the RTs to binary personality items. The results showed that there was some increase in the precision and reliability of trait estimates. An example would be, if a respondent's trait location is near the item location, he/she may hesitate between the two options and the RT is expected to be longer than normal and the response is more likely to change in a retest (Ferrando et

al., 2007). With the prevalence of computerized testing, the fact that RTs can be obtained without additional cost but provide more information seems to be promising.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6), 716-723.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1982). *Differential Aptitude Tests (Forms V and W)*. New York: The Psychological Corporation.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin,107,* 238-246.

Bielinski, J. & Davsion, M. L. (1998). Gender Differences by Item Difficulty Interactions in Multiple-Choice Mathematics Items. *American Educational Research Journal, 35,* 455-476

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University.

College Board (2003). *10 Real SATs* (3rd Ed.). New York: College Entrance Examination Board.

Davsion, M. L., Semmes, R., & Close, C. N. (2009). *Measuring average speed of numeric reasoning.* Technical report for United States Army Research Institute for the Behavioral and Social Sciences

Educational Testing Service (2002). *GRE: Practicing to take the General Test* (10th Ed.). Princeton, NJ: Educational Testing Service.

Ferrando, P.J., Lorenzo, U., & Molina, G. (2001). An item response theory analysis of response stability in personality measurement. *Applied Psychological Measurement, 25*, 3-17.

Furneaux, W. D. (1961). Intellectual abilities and problem-solving behavior. In H. J. Eysenck (Ed.), *Handbook of abnormal psychology (pp.167-192).* New York: Basic Books.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston (MA): Kluwer Academic Publishers.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269,* 41-45.

Hu, L., & Bentler, P. M. (1999). Cutoff criterion for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Kang, T. & Allan S. Cohen (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.

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

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

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp.157-171). Amsterdam: North- -Holland.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hamleton (Eds.), *Handbook of modern item response theory* (pp.187-208). New York: Springer.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: a new method of measuring speededness. *Journal of Educational Measurement, 34,* 213-232.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Steiger, H. H., & Lind, J. M. (1980, June). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Thissen, D. (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.* New York: Academic Press.

Thorndike, E.L., Bregman, E.O., Cobb, M.V., Woodyard, E. (1926). *The measurement of intelligence.* New York: Bureau of Publications, Teachers College, Columbia University.

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika, 2,* 249-254.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Van Breukelen, G.J.P. (1989). *Concentration, Speed and Precision in Mental Tests: a Psychonometric Approach*. PhD thesis, Nijmegen University, The Netherlands.

Van Breukelenm 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., Scrams, D.J., & Schnipke, D. L. (1999*). Using response-time constraints to control speedness in computerized adaptive testing. Applied Psychological Measurement, 23,* 195-210.

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.(2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46,* 247-272.

van der Linden,W.J.(2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33,* 25-41.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items, *Psychometrika, 75,* 120-139.

Verhelst, N. D., Versralen, H. H. F. M., & Jansen, M. G. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hamleton (Eds.), *Handbook of modern item response theory* (pp.169-185). New York: Springer.

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

White, P. O. (1973). Individual differences in speed, accuracy, and persistence: A mathematical model for problem solving. In H. J. Eysenck (Ed.), *The measurement of intelligence* (pp.246-260). Baltimore: Williams & Wilkins.

APPENDIX

Unstandardized Factor Loadings: Unidimensional Speed Model for Self-paced Standardized Log Response Time Variables (SP Response Time) and Two-dimensional Speed-Level Model for SP Response Time, Self-paced Accuracy (SP Accuracy) and Experimenter-paced Accuracy (EP Accuracy)

Unidimensional Speed Model for Self-paced Standardized Response Times				
	Block 1		Block 2	
Items	Level Factor	Speed Factor	Level Factor	Speed Factor
SP Response Time	---	-.51	---	-.48
Two-dimensional Model for SP Response Times, SP Accuracy, and EP Accuracy				
	Block 1		Block 2	
	Level Factor	Speed Factor	Level Factor	Speed Factor
SP Response Time	---	-.51	---	-.48
SP Accuracy	1.12	---	1.23	---
EP Accuracy	1.05	.37	.91	.32