KLARIC III, JOHN S., Ph.D. Relationships between Examinee Pacing and Observed Item Responses: Results from a Multi-factor Simulation Study and an Operational High Stakes Assessment. (2009)
Directed by Professor Richard M. Luecht. 155 p.

The use of response time in testing has a relatively long history, ranging from concerns over test speededness to using response times as performance indicators (e.g., speed and accuracy). This model-based investigation examined the relationship between item response times and examinee performance, focusing on semi-partial covariance between time indices and residual errors of measurement. Residual errors were estimated as deviations between observed item response scores on a multiple-choice test and item response theory (IRT) model-based expected response scores. In the first study, simulation was used to determine whether this relationship is detectable with either semi-partial correlation coefficients or with a measure of local item dependence, Q_3 statistics. The impact of this relationship on recovery of proficiency score estimates was studied with root mean square error (RMSE) statistics. Simulation results indicated that mean item semi-partial correlation coefficients were low, but increased as temporal manipulations increased in strength. Variability systematically decreased. Impacts on recovery of EAP proficiency estimates were small, with slight increases in estimate recovery as temporal manipulations increased in strength. In a companion study, simulation results were validated with results from an operational online assessment.

RELATIONSHIPS BETWEEN EXAMINEE PACING AND OBSERVED ITEM RESPONSES: RESULTS FROM A MULTI-FACTOR SIMULATION AND AN OPERATIONAL HIGH STAKES ASSESSMENT

by

John S. Klaric III

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2009

Approved by

Richard M Luecht, Ph.D.
Committee Chair

To Hope

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair Richard M. Luecht, Ph.D.

Committee Members Terry A. Ackerman, Ph.D.

Robert A. Henson, Ph.D.

John T. Willse, Psy.D.

Scott J. Richter, Ph.D.

June 25, 2009
Date of Acceptance by Committee

June 25, 2009
Date of Final Oral Examination

ACKNOWLEDGMENTS

Many individuals assisted and supported this project, particularly the Chair of my Supervisory Committee, Professor Richard Luecht. I am deeply grateful to Ric for his many suggestions and contributions, but also for his managing to get me through this process despite my best efforts to the contrary. Thanks are due to all Committee members, including Drs. Terry Ackerman, Robert Henson, Scott Richter, and John Willse, for significant contributions and examining innumerable drafts. All remaining errors are mine alone. Much gratitude is owed to the ever-available Dr. Morgan, with his tremendous supply of wit, good will, and energy. I hope our collaborations continue.

The University of North Carolina at Greensboro's Department of Educational Research Methodology, with Dr. Ackerman as Chair, together with its Graduate School, supplied significant graduate study support, both academic and financial. Additionally, the ERM Department supported me during studies at the University of Twente in The Netherlands. Some UNCG graduate work was funded by a grant to Rick Morgan from the College Board.

The North Carolina Department of Public Instruction advanced this dissertation by allowing access to examinee data and with substantial financial support. Thanks are due to Dr. Laura Kramer, who provided item response and timing data from North Carolina's Online Computer Skills Assessment. Without the indulgences of Drs. Lou Fabrizio and Gary Williamson (Accountability Services Division), this dissertation would have floundered.

My partner in life, Susan Klaric, and I have accomplished much together − forging relationships, very literally pouring foundations, and climbing red rocks. Her steadfast support enabled me to commit large amounts of time and energy in pursuing this goal. Susan played pivotal roles in the completion of this endeavor, initiated long ago on the midwestern prairie.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTATION

Symbol	Description
α_i	Item temporal discrimination
β_i	Item temporal intensity, equivalent to the amount of time required by examinees to correctly solve an item
δ_i	Measurement error, observed responses
ε_i	Measurement error, observed response times
θ_k, where $k = 1,2$	Examinee latent traits tested in simulation study
$\theta, \theta_j, \theta_{1j}$	Proficiency examinee latent trait (examinee j)
θ_{2j}	Pacing examinee latent trait (examinee j)
σ^2_{ti}	Variance due to item response time for item i
σ^2_{ei}	Variance due to residual for item i
τ	Examinee pacing (speededness) parameter, equivalent to θ_{2j}
a_1	Direct effect of examinee proficiency (θ_1) on an observed item response, Factor 2
a_{1i}, where i=1,2,3	Levels of a_1 in simulation study
a_2	Direct effect of examinee pacing (τ) on an observed item response, Factor 3
a_{2i}, where i=1,2,3,4	Levels of a_2 in simulation study
$\mathbf{a_i}$	Matrix of item discrimination parameters (multidimensional models)
a_i	IRT item discrimination parameter (unidimensional models)
b_i	IRT item difficulty parameter (unidimensional models)
c_i	IRT item asymptote parameter
d_i	IRT distance parameter (multidimensional models)
e_j, e_{ij}	Residual for item i (examinee j)
N	Number of observations contributing to a statistic
Q_3	A statistic indicating extent of local item dependence
P_i	θ_1 conditional probability of response to item i
$P(\theta)_j, P(\theta_1)_{ij}$	θ_1 conditional probability of response to item i (examinee j)
$\rho(\theta,\tau)$	Population-based correlation between latent traits θ and τ
$r\theta_1\tau$	Sample-based correlation between latent traits θ_1 and τ, indirect effect of pacing on an observed item response (Factor 4)
$r_{ee'}$	Correlation between item residuals for any given item pair
$r_{e_{ij}t_{ij}}$	Semi-partial correlation between item i's residual and observed response time for examinee j
t_i, t_{ij}	Observed response time for item i (examinee j)
u_i, u_{ij}	Observed response to item i (examinee j)

CHAPTER I

INTRODUCTION

Subjects' observed response times after independent stimulus presentations have been used to explain empirical phenomena in a range of scientific fields (see Luce, 1986). For example, temporal latencies between subjects' detection of a presented stimulus and response selection, together with errors made during that selection process, have been used in clinical screening for the potential presence of certain childhood psychiatric disorders (Epstein, Connors, Goldberg, & March, 1997). The investigation of response time measures has been facilitated by the proliferation of computer-based tests, where response times can be captured and recorded unobtrusively and accurately (Bartram, 2006; Kong, Wise, & Bhola, 2007; Schnikpe & Scrams, 1999). It is therefore not surprising that response-time research has been extended to psychometrics (Schnikpe & Scrams, 1999; van der Linden, 2006; van der Linden, 2007).

The sequence of response latencies to a series of administered test items, each of which with statistically-defined temporal characteristics, has been used to estimate an examinee characteristic that has been called *speededness* (van der Linden, 2006; 2007), or equivalently, *pacing*. This tempo of response generation has been modeled primarily through simulation or with assessment data from older subjects (e.g., the Arithmetic Reasoning Subscale of the Armed Services Vocational Aptitude Battery [ASVAB]; van der Linden, 2005, 2006). Prior to more widely applying this psychometric temporal model, it would be advantageous to examine the nature of the empirical relationship between examinee performance, as modeled by item response theory (IRT), and temporal parameters from a time-oriented model.

Theoretical Background

Unidimensional IRT and Response Time Modeling

To render these generalities more specific, consider a relatively common unidimensional IRT model, the three-parameter logistic model:

$$\Pr\left(u_i = 1 \middle| \theta_j\right) = E\left(u_i \middle| \theta_j\right) \equiv P_i = c_i + \left(1 - c_i\right) \frac{\exp\left[a_i\left(\theta_j - b_i\right)\right]}{1 + \exp\left[a_i\left(\theta_j - b_i\right)\right]} \tag{1}$$

In Equation 1, u_i is the observed dichotomous (0/1) response to item i, θ_j is a continuous latent variable describing examinee proficiency, a_i is an item discrimination parameter for the item response characteristic function (ICRF) that is proportional to the slope of the function at its inflection point, b_i is a location (item difficulty) parameter associated with the inflection point on the ICRF, and c_i is a lower-asymptote parameter (see Lord, 1980). Figure 1 represents observed dichotomous responses to a hypothetical test composed of three items. The observed responses are conditional on examinee proficiency ($u_i \middle| \theta_{1j}$). If responses to items are locally independent, that is, the responses are uncorrelated when θ_j is fixed, then θ_j, a_i, b_i, and c_i are sufficient to explain the observed response, u_i.

Insert Figure 1 about here

Now consider the observed amount of time required for a given examinee to respond to an item. Van der Linden (2006, 2007) has proposed a response time model (RTM) that relates the cumulative time spent on an item to three parameters: τ, a person speededness or pacing parameter; α_i, a temporal slope (discrimination) parameter; and β_i, an item time-intensity parameter describing the average amount of time required for correct item solution. The model assumes that, if independence of the time cumulants holds, τ, α_i, and β_i fully explain the observed response time, given the functional relationship:

$$f\left(t_{ij}\right)=\frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i\left(\ln t_{ij}-\left(\beta_i-\tau_j\right)\right)^2\right]\right\} \tag{2}$$

In Equation 2, α_i is an item slope parameter and t_{ij} represents the time spent on a particular item *i*, by examinee *j*. If response time, t_{ij}, and observed performance, u_{ij}, are independent, these two models can operate without confounds. However, should time and performance be related – for example, if more time-intensive items are also more difficult, or if higher proficiency examinees tend to work more quickly, potential confounds may emerge. Partialing out the impact of τ and θ on u_{ij} essentially "purifies" the measure of θ (and the item characteristics *a*, *b*, and *c*).

Local Item Dependence

In practice, stable estimates of θ are obtained by maximum likelihood or Bayesian estimation (i.e., mean or modal estimates of a posterior likelihood function). Both methods of estimation require the relatively strong assumption that item responses are locally independent (Hambleton & Swaminathan, 1985). When neither the observed responses themselves nor their errors are correlated at specific θ levels, the items satisfy the conditional local independence assumption (Lord & Novick, 1968; Lord, 1980; Hambleton & Swaminathan, 1985; Yen, 1993; Huynh, Michaels, & Ferrara, 1995; Iramaneerat, Myford, & Yudkowsky, 2006). Any residual correlation or covariance between observed responses implies that one item response is dependent on the occurrence of another and is referred to as *local item dependence* (LID). As Lord and Novick note (1968, p. 361), positive correlations among responses are expected across a total group of examinees if the test is measuring a single proficiency. Instead, LID must be evaluated conditionally on θ, the latent proficiency.

More specifically, under the usual assumption of local independence (and its corollary, the unidimensionality assumption), all residuals are mutually uncorrelated. A weak test of this is

$$E_{\theta} = \left(\varepsilon_i, \varepsilon_j \middle| \theta\right) = \operatorname{cov}\left(\varepsilon_i, \varepsilon_j \middle| \theta\right) = 0$$

for items $i \neq j$, and where $\varepsilon_i = u_i - P_i$. In this equation, the expected value of the covariance between two item residuals is zero, implying that the residuals are not correlated. The local item independence assumption is also reflected in the classical test theory assumption of mutually uncorrelated errors. A corollary assumption from classical test theory (e.g. Allen & Yen, 1979) is that $\operatorname{cov}(e_i, T_2) = 0$, where T_2 is some secondary true score such as τ.

As noted above, there are several ways to detect residual covariance. A standard test for LID is Yen's (1993) Q_3 statistic, which examines the conditional covariance between the residuals of pairs of item responses—that is, $E\{[u_i - P_i(\theta)]\}\{[u_j - P_j(\theta)]\}$, where, for two items *i* and *j*, u_k is the observed response and $P_k(\theta)$ is the model-based expected response function for an examinee with proficiency θ, $k \in \{i, j\}$. In simulation studies, Reese (1995) has shown that, when the local independence assumption is violated to a high degree (e.g., Q_3 approximates 0.3 or higher), examinee latent trait distributions are no longer normal in shape and the examinee proficiency scores may no longer be invariant. Iramaneerat (2006) similarly showed that violations of the essential item independence assumption could result in artificially low variance in responses, leading to inaccuracies in estimating proficiency under an IRT model.

Examples of violations of the local item independence assumption can be readily found in item responses on operational tests. The obvious case is where one item response cues responses to other items. A more subtle example is where responses to items associated with a specific reading passage cue responses to other items about that passage (Lee, 2000). Or, examinees may respond slowly to items on a test with rather stringent time limits, but as time limits are approached use a predetermined response strategy (e.g., a rapid-guessing strategy).

Stout (1987, 1990) showed that such a strict assumption concerning θ-conditional item independence can, in most IRT applications, be replaced by a weaker assumption, θ-conditional "essential item independence." This assumption can be satisfied when the mean of θ-conditional covariances between items is small in magnitude. However, this does not imply that minor amounts of conditional covariance are necessarily ignorable.

One source of potential violations of this weaker item independence assumption affecting proficiency score estimation is the speed-accuracy tradeoff phenomenon, used traditionally to explain varying within-subject response accuracies in perceptual discrimination tasks (Luce, 1986). This phenomenon has been applied to psychometric data to explain decrements in examinee θ estimates associated with examinee pacing strategies (Schnikpe & Scrams, 1999; van Breukelen, 2005; van der Linden, 2006; van der Linden, 2007). On the other hand, properly accounting for this source of response variability (when the context validly allows for it) may increase precision of examinee θ estimates (Schnikpe & Scrams, 1999; van der Linden, Scrams, & Schnipke, 2003). The important point is that nonrandom errors will almost assuredly distort any unidimensional IRT modeling procedures (van der Linden, 2005, p. 191; van der Linden, Breithaupt, Chuah, & Zhang, 2007).

Examinee pacing is not the only potential confound when attempting to accurately estimate proficiency. Examinee motivation is clearly confounded with performance (Wise, Kong, & Pastor, 2007; Wise & DeMars, 2006). Although it is usually assumed that responses to a psychometric instrument in specific testing situations represent examinees' optimal performances (Kong et al., 2007), motivation may be a significant source of nuisance variance—especially in low-stakes testing situations (Wise & DeMars, 2006; Wise & Kong, 2005). In fact, Wise and DeMars have explicitly used response times as surrogate indicators of diminished motivation among test takers.

Statement of the Problem

A number of researchers have shown that time-orientated characteristics of items and examinees can affect ability or proficiency estimation under IRT (Lord, 1980; Schnipke & Scrams, 1997; Thissen, 1983; van Breukelen, 2005; van der Linden, 2006; van der Linden, Scrams, & Schnipke, 2003; Wise & DeMars, 2006). To maintain conditional local independence, discarding response data from those examinees exhibiting speededness based on the proportion of items answered has been recommended (Lord, 1980, p. 182). Other options exist—including replacing items that may involve rapid guessing with simulated or expected responses. The present study does not propose any specific solutions, merely an exploration of the phenomenon and methods to detect that phenomenon. However, this previous research has not explicitly investigated varied relationships between speed (pacing) and accuracy, nor the role of those relationships in moderating performance, and ultimately, their effect on obtaining accurate estimates of proficiency.

This investigation specifically considers possible relationships between two latent variables of interest: one variable is an examinee's pacing trait, τ_j. The second latent variable is θ_j , a latent proficiency measured by some test or assessment. Potential relationships are depicted in Figure 2. The scored item response, u_i, is directly caused by the latent proficiency θ_j. However, a pacing trait τ, in addition to causing each item's response time, t_{ij}, may also influence that item's scored response, u_i. This influence may be exerted either directly via a_2 or indirectly via the correlation between τ and θ_j.

Insert Figure 2 about here

Research Questions

Two questions arise from the theoretical framework in Figure 2, addressing fundamental questions not previously investigated in the psychometric literature. These relationships are examined in the current research with two methodological approaches.

1. *Can item semi-partial correlations (* $r_{e_{ij}t_{ij}}$ *, where* e_{ij} *is the residual error of measurement [item* i*, person* j*] and* t_{ij} *is the response latency) be used to detect a confounded relationship between observed response accuracy and* t_{ij}*?* When $r_{e_{ij}t_{ij}}$ is of sufficient magnitude, this may serve to indicate a potential confound between e_{ij} and the nuisance variable t_{ij}. Should the magnitude of the confounded relationship be shown to vary due to experimental manipulations, it may be that the variance in item response time ($\sigma^2_{t_i}$) uniquely explains a portion of the variance in the item residual ($\sigma^2_{e_i}$), as shown by the intersection of *A* and *B* in Figure 3. This question will be dealt with in simulation studies, and generalized to a real-data situation.

Insert Figure 3 about here

----

2. *If these semi-partial correlation coefficients can indeed detect possible speed-accuracy confounds, when and under what conditions examined in this research do these confounds substantially impact EAP* θ_j *values?* This

question was addressed by manipulating relationships shown in Figure 2 in a four-factor computer simulation. At three levels of test length (20, 30, and 60 items), item semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ were evaluated in a 3 (mean Item Discrimination [a_1]) X 4 (direct influence of the pacing parameter τ on u_i [a_2]) X 5 (indirect influence of τ on u_i via the correlation between the latent variables τ and θ) fully-crossed factorial arrangement of treatments. All manipulations were conducted with the capabilities of *MIRTGEN 2.0 with Response Times* (Luecht, 2008).

It is expected that as levels of the Test Length and Item Discrimination a_1 factors increase in magnitude, increments in measurement precision indicated by decreased residual error [$e(i)=u(i)-P(\theta_j)$] would occur. The impact of temporal manipulations on recovery of true θ_j proficiency values was studied with root mean square error (RMSE) statistics.

CHAPTER II

REVIEW OF THE LITERATURE

This chapter summarizes three relevant themes. The first theme concerns the general problem of item-response dependencies in examinees' proficiencies. The second theme amounts to a review of popular IRT models for dichotomously scored items and implications of encountering dependencies in the conditional distributions of residuals. The final theme involves approaches to modeling response time. These three themes provide the necessary background for understanding the focus of this dissertation on exploring the direct and indirect influences of examinee pacing on observed, dichotomously scored, item responses.

Dependencies among Observed Test Scores

Examinee Latent Trait Estimation

Classical true-score theory (CTT) offers a useful model that has guided the estimation of unobservable trait levels in examinees based on their test performance since E. L. Thorndike's seminal work in 1904 (Allen & Yen, 2002). However, CTT estimation of examinee trait levels is based on hypotheses that cannot or are unlikely to be falsified by the available data (Lord, 1980). Hambleton, Swaminathan, and Rogers (1991) further note the following limitation of CTT: An examinee's true score is defined as the mean of total scores earned by that examinee on a specific test administered an infinitely large number of times. But estimates of classically defined test and item characteristics, such as item *p* values, are population-dependent, varying with mean examinee ability levels. Taken together, these statements imply that different true score values for the same examinee are likely to be obtained when she/he is tested in different populations varying in ability levels.

One mathematical modeling method that counters these dependencies is item response theory (IRT; Lord & Novick, 1968; Lord, 1980; Hambleton & Swaminathan, 1985; Hambleton, 2006). Given a sufficient number of examinees assessed, IRT has challenged the dominant role of CTT in estimating trait levels over the past several decades. Unobservable trait levels are estimated iteratively by IRT methods that account for both observed examinee item performance and statistical item characteristics (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). IRT-based measurement offers the capability of predicting unobservable examinee trait levels from observed examinee test behavior, through statistical mechanisms that are both test- and sample-independent (Lord, 1980).

Dichotomous Unidimensional IRT Models

The mathematics underlying IRT were developed to predict observed examinee test responses from an unobserved trait level (equivalently, proficiency or ability level), in combination with at least one statistical item parameter. This θ is a random variable estimated by Equation 1. When items are dichotomously scored (i.e., scored using binary right/wrong categories), the three-parameter logistic IRT model (3PL) is often employed (Birnbaum, 1968). Although not technically a pure logistic model (van der Linden & Hambleton, 1997), the 3PL model is used for the estimation of item parameters in several large-scale testing programs (Lord, 1980; Samejima, 1988; van der Linden, Scrams, & Schnipke, 1999): The 3PL model is directly reducible to a 2PL logistic model by substituting $c_i = c = 0$ for all items, $i = 1, \ldots, n$. A further simplification ($a_i = a = 1.0$) yields the 1PL, or Rasch (1960), IRT model (Hambleton & Swaminathan, 1985).

A key IRT assumption is that of local item independence. Local independence of item responses appears in numerous forms in test theory. In classical test theory, it is assumed that errors of measurement are uncorrelated given the true score of an examinee (Lord &Novick, 1968; Yen, 1984). In IRT, a set of items is considered locally independent with respect to the assumed model, if after conditioning on an examinee's proficiency, the joint probability distribution of all items is equal to the product of the univariate probability distributions of each item (Hambleton & Swaminathan, 1985; Lord, 1980). Formally, this is the strong definition of local independence and is stated mathematically in Equation 3:

$$L(U|\theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} \left[1 - P_i(\theta)\right]^{1-u_i} \qquad (3)$$

In Equation 3, U is the vector of observed responses for n items for a random test-taker with ability θ. A 1PL, 2PL, or 3PL response function (e.g., Equation 1 for the 3PL IRT model) can be assumed. A weaker definition of local independence is often used to investigate the appropriateness of this assumption. Weak independence states each pair of items has a joint probability equal to the product of corresponding marginal distributions after accounting for each examinee's ability. This independence can be expressed as:

$$P\left(u_i = 1, u_j = 1 \middle| \theta\right) = P_i\left(\theta\right) P_j\left(\theta\right),\ i \neq j \tag{4}$$

As the label implies, weak local independence is a less stringent requirement that is necessary but not sufficient for strong local independence (Stout, 1990). However, it is reasonable to assume that if variables are pair-wise independent, higher order dependencies, though possible, are highly implausible (McDonald, 1997). If Equation 4 holds for all item pairs, the trait proficiency (θ) accounts for all of the information relevant for each examinee, thus allowing the items to be evaluated independently (Yen, 1993).

This idea can be expressed in terms of conditional covariance as well. If item responses u_i and u_j are locally independent, they will have a conditional covariance of zero. That is,

$$\operatorname{cov}\left(u_i, u_j \middle| \theta\right) \propto E\left\{\left[u_i - P_i\left(\theta\right)\right]\right\}\left\{\left[u_j - P_j\left(\theta\right)\right]\right\} = 0,\ i \neq j \tag{5}$$

if the responses are conditionally independent. Non-zero covariances indicate that there may be one or more additional factors that explain the remaining variance (Yen, 1993). These additional factors are potential sources of LID that may or may not be relevant to the trait or behavior being measured.

Item Response Time Parameters

Two-Choice Discrimination Paradigm

Experimental studies of subjects' mean reaction times (MRTs) after presentation of perceptual stimuli have been used successfully and productively in psychophysics for several centuries (for a review, see Luce, 1986; Swets, Tanner, & Birdsall, 1961; van

Breukelen, 2005). Methodologically in two-choice discrimination studies, subjects trained or verbally instructed in patterns of response production are presented with multiple series of experimenter-controlled perceptual trials. Following stimulus discrimination during trial *i*, subject *j* selects an appropriate response from his/her behavioral repertoire based on a response criterion established by that subject (Swets et al., 1961). The latency during trial *i* required for stimulus discrimination, response choice, and producing an observable response is recorded as a response time measure (Wenger, 2005, p. 384). Although pioneering introspectionists discredited some earlier concepts and methods, the objective investigation of temporal parameters surrounding behavioral responses after stimulus presentation forms the basis of hypotheses for some elementary cognitive processes (Luce, 1986; Sternberg, 1966). Characteristics of typical MRT distributions in two-choice discrimination tasks have been well documented (Luce, 1986). To briefly summarize some of the empirical findings, statistical characteristics of MRT distributions from these tasks include unimodality and positive skewness. Distributional aspects are independent of sensory modality (Luce, 1986), and the use of response time measures has been extended to performance on cognitive tasks such as answering items on psychometric instruments (Schnipke & Scrams, 1997).

Several perceptual and cognitive processes compose sequential steps for the completion of a two-choice discrimination task. In initial processing steps, purely perceptual operations are involved in stimulus detection. Subsequent processing steps (stimulus discrimination, response choice and determining mode of response production) require higher order perceptual and cognitive operations with additional cognitive requirements leading to increases in MRT in reaction time tasks (Maris, 1993; Schweizer, 1998; Swets et al., 1961). Maris (1993) demonstrated that the increases in MRT on a task involving cognitive rotation of stimuli are directly related to the increase in cognitive requirements of the task. After stimulus detection through perceptual operations, however, higher order processing can be circumvented with examinee pacing strategies that increase the speed of response generation (Schnikpe & Scrams, 1999).

To model a given experimental condition with a two-choice discrimination paradigm where response times are measured in milliseconds, a "two-stage mixture model" has been developed (Luce, 1986). This same model has been applied to account for predetermined item response strategies (Schnipke & Scrams, 1997). According to this model, if the probability that subjects use a slow-paced strategy for responding to the given experimental condition is $p(s)$, then the distribution of observed response times (t) is a function of the response time distributions characterizing those subjects using either slow- [$(G_1)(t)$] or fast-paced [$(G_0)(t)$] response patterns:

$$F(t|s) = p(s)(G_1)(t) + [1-p(s)](G_0)(t) \tag{6}$$

This two-stage mixture model has been applied to the study of response pacing by examinees on psychometric tests with response latencies collected during the administration of non-adaptive computer-based tests (Schnipke & Scrams, 1997; Wise & DeMars, 2006).

Whether it is appropriate to apply a two-stage mixture model developed for a two-choice perceptual discrimination to item-solving tasks substantially greater in cognitive complexity with much longer response times (several orders of magnitude greater during achievement tests), has been investigated. Schnipke and Scrams (1997) applied the two-stage model while investigating fast-paced responding on a psychometric assessment. Two versions of a linear computer-administered nationwide test (the Graduate Record Examination, GRE) were analyzed. After showing that the probabilities of correct item responses approached chance levels during episodes of rapid item responding, they compared response time values predicted by the two-stage mixture model with empirical response times. They studied 1) whether predicted and observed item response time distributions in the presence of rapid item guessing were generally of the same form, and 2) whether the occurrence of rapid guessing behavior was dependent on an item's serial position. They also examined response accuracy as a function of response latency.

Schnipke and Scrams (1997) found that when rapid item guessing was used as a response strategy, predicted and observed item response time distributions were of approximately the same shape; in the exemplar items shown, the largest deviations of predicted values from observed findings were near the lower asymptote. This finding, noted earlier by Yamamoto (1995), supports the modeling of response times by a two-stage mixture model. The authors also found that serial position played a role in occurrence of rapid item guessing, but that occurrence of this phenomenon was also affected by item position within an item set. The relationships among response accuracy and response latency were complex. Although accuracy did improve as response latencies increased, response accuracies reached plateaus at longer latencies for all items shown. Among item latencies distributed in approximately the upper quartile, variability in accuracy increased markedly. Importantly, Yamamoto (1995) noted that, in the region of this increased temporal variability, correct responses no longer fit a standard IRT model.

Response Time Modeling Within an IRT Context

Potentially deleterious effects of attempting to model true-score distributions, when examinee pacing strategies or severe time limits result in some examinees failing to complete all of the test items, was suggested several decades ago (Lord, 1980, p. 242). When examinees can be clearly identified as running out of time and failing to complete certain items, Lord suggested the expedient of excluding such examinees from IRT analyses.

Thissen (1983) went a step further. By linearly combining temporal and weight-corrected IRT parameters, he proposed a mixture model for predicting log response times outside of the classical two-stage mixture model. The temporal parameters included log mean response times, as well as person and item slowness parameters:

$$\log (t_{ij}) = v + s_j + x_i - rz_{ij} + e_{ij}, \text{ with } e_{ij} \sim N(0,\sigma^2), \quad (7)$$

where $\log (t_{ij})$ is the response time of person j on item i, v is the overall mean log response time, s_j is a person parameter indicating slowness to respond, x_i is a parameter indicating time requirements for completion of an item, and z_{ij} is the exponential term from the 2PL item model (see Equation 1, with c_i=c=0 for all items). In the original (Thissen, 1983, equation 9.2), parameters composing z_{ij} were defined as in Bock's nominal IRT model. r is a regression weight parameter indicating the relationship between item easiness and examinee ability, and e_{ij} is an error term. This model implies that parameters underlying speed and accuracy are linear and additive, implications that, while granting simplifying mathematical assumptions, are not solidly supported by available evidence (Fischer & Kisser, 1983).

Other methods of incorporating temporal parameters into the 1PL (Rasch) model have also been developed. These methods usually involve directly introducing an examinee time variable as an additional parameter in an IRT model; this is shown in the parameterization of the Rasch model by Roskam (1997, p. 193). Roskam (1997) assumes that examinee response times are described appropriately by a Weibull distribution (Verhelst, Verstralen, & Jensen, 1977). Roskam (1997) notes however that an adequate goodness of fit test does not exist for his 1PL (Rasch) model. Moreover, the limited empirical data available support this model as well as a competing Rasch model (Verhelst, Verstralen, & Jensen, 1977). The direct incorporation of a temporal examinee parameter in an IRT model appears to confound, and not disentangle, relationships with the underlying latent traits under consideration.

An Effort Moderated Response Time Model

An "effort-moderated" IRT model was developed that followed the two-stage mixture model approach (Wise & DeMars, 2006). The effort-moderated model was developed from observations from speeded, high-stakes, testing situations, but is most applicable to low-stakes situations (Kong et al., 2007; Wise et al., 2007).

This model was motivated by the concept that more accurate θ estimates may be obtained by correcting for item responses performed with the fast-paced response mechanism (Yamamoto, 1995). This increased accuracy may result by decreasing variability in response vectors not necessarily related to the latent variable (Wise et al., 2007); this is observed during the later portions of test administrations.

In the effort-moderated response time model, item response time thresholds (T_i) are established describing the intersection of response time distributions for item *i* from two samples of examinees. In the formulation by Wise and colleagues, T_i is the point on the response time distribution where the response time distribution of "rapid item guessers" intersects with the distribution of those exhibiting "solution behavior." Whereas examinees showing rapid item guessing may respond according to a predetermined strategy (e.g., "always pick c") or after skimming the item stem and options for keywords, those exhibiting solution behavior carefully peruse each item and attempt to solve the puzzle, responding as accurately as possible (Schnipke & Scrams, 1997). The value of the threshold between these groups, T_i, determined by visual inspection of the bimodal item RT frequency distribution from the entire sample of examinees, was used to establish the value of a binary indicator showing whether solution behavior was exhibited by examinee *j* on item *i*:

$$\begin{cases} Solution\ Behavior_{ij} = 1 \ if \ RT_{ij} > T_i \\ 0,\ \ otherwise \end{cases} \tag{8}$$

In the two-stage mixture model (see Equation 6), *p(s)* is the realized value of the dichotomous variable *Solution Behavior*$_{ij}$. When *Solution Behavior*$_{ij}$ equals one, $G_1(t)$ determines the probability of a correct response given the 3PL IRT model shown in Equation (1). When *Solution Behavior*$_{ij}$ equals zero, $G_0(t)$ is defined as a guessing constant equivalent to the reciprocal of the number of options for each item. Use of this model resulted in decreased test information functions most evident at θ between –2 and +2 compared to the standard 3PL model, a finding that may have been expected due to the decrease in estimates of item discrimination parameters (a_i in Equation 1). The validity of this modeling approach, however, was demonstrated for college students by correlating θ estimates from a low-stakes test of student information literacy with SAT Verbal and Quantitative subscores. These correlations were significantly higher with the effort-moderated IRT model compared to θ estimates from the standard 3PL IRT model (Wise & DeMars, 2006).

Lognormal Modeling of Response Times

Thus far, two distinct types of methodologies have been presented for the modeling of response time parameters in psychometric tests. The first method models temporal parameters interacting with standard IRT parameters in a regression model, which is seen with the Thissen (1983) model. A second method, the effort-moderated model, uses a dichotomy based on item response time as a vehicle to characterize responses as providing evidence either for demonstrations of solution behavior or for rapid item guessing. For the item responses characterized by solution behavior, the probability of correct response conditioned on θ is estimated using standard IRT procedures. For item responses characterized by rapid item guessing, conditional correct response probabilities are estimated by the reciprocal of the number of response options (Wise & DeMars, 2006).

Van der Linden (2006, 2007) proposes one more possible model, one in which two separate person parameters are estimated corresponding to those components implied by the speed-accuracy tradeoff phenomenon. At the level of the fixed person, response accuracy is estimated by θ using standard IRT parameters. Examinee pacing, or speededness, is independently estimated by the following relationships based on a lognormal response time distribution. The RTM shown earlier in Equation 2 is replicated below:

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

The logarithmic relationship posited in the second term puts the RTM in the lognormal family of functions. In that equation, t_i ($t_i > 0$) is a random variable representing the response time of a fixed person on item *i,* and τ ($-\infty < \tau < \infty$) is the temporal pacing skill of the examinee, where a greater value of τ indicates that the person tends to complete a given set of items more quickly (i.e., complete given tasks with smaller response latencies). The RTM specifies two item temporal parameters. β_i is the amount of time required to solve item *i*. Van der Linden (2006) refers to β_i as the time intensity of item *i*. Since α_i ($\alpha_i > 0$) is defined as the reciprocal of the response time distribution's standard deviation, a larger dispersion results in a smaller α_i factor, implying that the item's ability to provide a precise estimate of τ_j (temporal pacing) is decreased. With a larger α_i corresponding to a smaller standard deviation, the item's capability to improve the estimation precision of τ is increased. Used in this way, α_i is a temporal discrimination factor.

Van der Linden (2006) addressed the issue of model fit by studying four conditions with a 2 (Model Type) X 2 (α_i, Slope Constraint) factorial arrangement of treatments. Response time data were gleaned from a random sample of over 38,000 individuals taking a nationwide test of Arithmetic Reasoning (a subscale of the ASVAB). The fit of the data was assessed with two different model types: the lognormal RTM shown above (see also Equation 2) and its normal analog. Fit was also assessed under two α_i constraint conditions: when α_i was permitted to vary freely across items; and when α_i was fixed to a common value across all items (i.e., constraining $\alpha_i=\alpha$). Based on visual evidence from fit plots, it appeared that the lognormal model fit the response time data better than the normal variant. Moreover, the fit of the lognormal model was similar under both α_i constraint conditions (van der Linden, 2006). Findings from these data indicate that constraining α_i , a parameter used to measure the temporal discrimination of an item, only slightly effects the fit of the RTM to empirical data. Because model fit under both α_i conditions was similar, constraining slopes to a common value may be justified in future research when selecting the structure of an IRT model.

This chapter provided a discussion of the general case where response dependencies, conditional or not, may occur. Some implications of parameter estimation with unidimensional IRT models in the presence of conditional response dependencies are also summarized. Several methods of modeling item response times, models from both a historical perspective and those within an IRT framework, were presented.

CHAPTER III

METHODOLOGY

This chapter provides the methodologies that were used in this research to investigate direct and indirect influences of examinee pacing on observed item responses. In a simulation study, several of the theoretical relationships shown in Figure 2 were manipulated as random factors in a fully-crossed research design. The generalizability of these simulation findings was partially shown by examining relationships between item response times and examinee performance on a dichotomously-scored, computerized but not adaptive, operational assessment.

Simulation Study

Data Source

In the following simulation, one N x n dichotomously scored response matrix was generated for each of 10 replications within each treatment condition with *MIRTGEN 2.0 with Response Times* (Luecht, 2008). These matrices of scored responses were generated with a multidimensional three-parameter logistic IRT model (equation 11); two latent trait dimensions (θ_k=2) were modeled: θ_1=ability and θ_2=τ_j. Each scored response, u_{ij}, could be influenced either by θ_1 only or by both θ_1 and τ_j.

$$\Pr\left(u_i \middle| \boldsymbol{\theta}_\mathbf{k}; \mathbf{a}_\mathbf{i}, d_i, c_i\right) \equiv P_{ij} = c_i + \left(1 - c_i\right) \frac{\exp\left[\mathbf{a}_\mathbf{i}{}' \boldsymbol{\theta}_\mathbf{k} + d_i\right]}{1 + \exp\left[\mathbf{a}_\mathbf{i}{}' \boldsymbol{\theta}_\mathbf{k} + d_i\right]} \qquad (11)$$

For each replication, a *N x n* matrix of item response latencies was generated by *MIRTGEN 2.0 with Response Times* (Luecht, 2008) using van der Linden's response time model (2005; equation 2), where the strength of the relationship between t_{ij} and τ_j was parametrically determined. τ_j was also allowed to influence observed responses, u_{ij}, through parametric manipulations in the generating model (i.e., either via the direct influence of τ_j on u_{ij}, or indirectly via the correlation between τ and θ).

In summary, two data points were generated for each simulee (j=1,…,N) by item (i=1,…,n) transaction: (1) a dichotomously scored test response, u_{ij}; and (2) an item response latency, t_{ij}. This was accomplished for each of the 10 replications per treatment condition by first creating a *N x k* (N=1000, k=2) matrix of multivariate normal random deviates with a specified correlation between the k=2 dimensions. Second, a *n x* 6 item parameter matrix was made containing multi-dimensional IRT and temporal item characteristics.

Research Design

Three possible influences of the θ_k latent traits on observed item responses (u_i) were manipulated, as shown in Figure 2. These included manipulations of mean item discrimination (a_1), the mean direct influence of τ on u_i (a_2), and the indirect influence of τ on u_i expressed as the linear correlation between the latent traits ($r_{\theta\tau}$). These manipulations were systematically varied at three fixed test lengths (20, 30, and 60 items). Three target levels of the mean item discrimination factor (a_1) were used: 0.5, 0.75, and 1.0. The mean direct influence of τ on u_i (a_2) was modeled at 4 target levels: 0.0, 0.25, 0.5, and 0.75. The mean indirect influence on u_i by the $r_{\theta 1\tau}$ correlation was modeled at 5 target levels: -0.2, 0.0, 0.2, 0.4, and 0.6. A tabular presentation for the three (Test Length) by three (Item Discrimination, a_1) by four (Direct τ Effect on responses, a_2) by five (Indirect τ Effect, $r_{\theta I \tau}$.) research design is shown in Table 1. This design included ten replications, each with N=1000 simulees.

Insert Table 1 about here

Experimental Procedures

Item Parameters

For each level of the Test Length condition, deviates from a normal distribution were sampled with replacement to make three levels of the item discrimination factor (a_{1i}): $a_{11} \sim N(0.5,0.15)$, $a_{12} \sim N(0.75,0.15)$, and $a_{13} \sim N(1.0,0.15)$. Within each a_{1i} treatment condition, deviates were sampled with replacement to make four levels of the a_{2i} condition specifying the direct τ effect on observed item responses: $a_{21} \sim N(0.0,0.15)$, $a_{22} \sim N(0.25,0.15)$, $a_{23} \sim N(0.5,0.15)$, and $a_{24} \sim N(0.75,0.15)$. These were fully crossed with the five levels of the indirect τ effect, $r_{\theta I \tau}$.

Across all treatment conditions, values for d_i (a multidimensional IRT item parameter for all items i analogous to the unidimensional b_i item difficulty parameter [Luecht, 2008]) were obtained by sampling with replacement pseudo-random deviates from a normal distribution ($d_i \sim N(0.0,1.0)$). Constant c_i values (0.15) were maintained for all items i in all treatment conditions. Values for item temporal parameters α_i and β_i (van der Linden's [2005] item temporal discrimination and item time intensity parameters, respectively) were sampled with replacement from normal (α_i, β_i ~ N(0.0,1.0)) distributions. Lognormal variates for β_i were generated from these values.

Examinee Characteristics

For each of the 10 replications in each treatment condition, a 2 x N matrix of deviates from a multivariate normal distribution (N=1000) was generated. Elements in each column vector of these matrices were realizations for one of the θ_k latent traits. This procedure produced true τ (Pacing) estimates (Figures 4 to 6). These figures show that τ estimates did not vary across levels of the factors manipulated in this simulation.

Insert Figures 4, 5, and 6 about here

Unidimensional 3PL IRT Calibration

Dichotomous item response matrices were generated for all replications in each treatment condition according to these item and examinee characteristics. A standard 3PL unidimensional IRT model was fit to these data with BILOG MG (Zimowski, Muraki, Mislevy, & Bock, 2002). A maximum of 125 E-M cycles with a maximum of 75 Newton-Raphson (maximization) iterations was specified for each local calibration. A convergence criterion of 0.001 was used. If the local calibration of the item response matrix converged to a unique solution, unidimensional item parameters (a_i, item discrimination; b_i, item difficulty; and c_i, pseudo-guessing) were estimated and retained for further analyses. A sample BILOG MG program for the simulation study is shown in Appendix A. Based on these item parameters and simulated examinee θ_j EAP estimates, response probabilities ($P[\theta_j]$) were calculated using equation (1). Item residuals were calculated as the difference between u_{ij} and $P(\theta_j)$, that is: $e_{ij}=u_{ij}-P(\theta_j)$.

Data Analysis

Analyses for Each Replication

Item independence. After convergence to a solution was confirmed, Yen's Q_3 statistics were obtained for each unique pair of item residuals. Yen's correction [-1/(n-1), where n is the total number of test items] was then applied to these correlations. For each replication within a treatment condition, descriptive statistics of the corrected Q_3 statistics (mean and standard deviation) were obtained.

Item semi-partial correlation calculations. For each item in a given replication, unstandardized item residuals ($e_{ij}=u_{ij}-P(\theta_j)$) for each examinee response, u_{ij}, were correlated with that item's response times, t_{ij}. This specific semi-partial correlation ($r_{e_{ij}t_{ij}}$) expressed the relationship between these variables after the effect of θ on the scored response had been statistically removed. For each item, therefore, $r_{e_{ij}t_{ij}}$ expressed the "purified" relationship between these two variables. For each replication within a treatment condition, descriptive statistics of item $r_{e_{ij}t_{ij}}$ (mean and standard deviation) were obtained.

Recovery of EAP θ_j estimates. Root mean square error (RMSE) statistics were used to assess the extent to which unidimensional EAP θ_j estimates accurately recovered "true" θ_j values generated by *MIRTGEN 2.0 with Response Times.* RMSE, the standardized difference between expected and true values, was obtained with equation (12) for each replication r in all treatment conditions:

$$RMSE_{jr} = \sqrt{\frac{\Sigma\left(\hat{w}_j - w_j\right)^2}{k}} \tag{12}$$

where $RMSE_{jr}$ is a measure of standardized error in θ estimation for simulated examinee j at replication r in any given treatment condition, $\hat{w}_j$ is that examinee's EAP θ estimate, and w_j is the true value of that estimate for examinee j. k is the total number of observations in replication r. In the next phase of analyses for each treatment condition, the mean RMSE of the θ_j estimates across all converging replications r was found for simulates in that treatment condition. This method is a slight modification to that found in previous reports (Kaskowitz & De Ayala, 2001; Schnipke & Scrams, 1997).

Procedures Summarizing Replications Within Each Treatment Condition

Data analyses were conducted for all converging replications within each treatment condition. For each replication, separate datasets were built containing item semi-partial correlation coefficients, item pair Q_3 statistics, and squared deviations between EAP θ estimates and true θ values. The distributions of the first two dependent variables were summarized for each replication (i.e., mean semi-partials and Q_3, as well as the standard deviation of the dependent variables in each replication). RMSE for each replication was computed from mean square error. Concurrently, datasets summarizing mean item parameters (a_i, b_i, and c_i) and mean simulee statistics, including correlations between τ and EAP θ estimates, were built.

A treatment dataset was made with statistics describing the distributions of replication mean semi-partial correlations and mean Q_3 statistics (n, mean of means, average standard deviations). For mean RMSE and mean latent trait correlation coefficients, measurement error was estimated as the standard error of replication means.

Interpretation of Simulation Results

To guide interpretation of main and interaction effects for several of the dependent measures (mean semi-partial correlation coefficients, mean Q_3, and mean RMSE), a four-way analysis of variance from the general linear model was conducted; the 1741 converging replications were used as "subjects." All main effects and possible interactions were included as terms in the general linear models; only effects with alpha levels less than 0.0001 were considered statistically significant. Because each of the 10 replications in the 180 treatment conditions had data from 1000 simulees, statistical power was such that conservative criteria for declaring statistical significance were used. η^2, the ratio of treatment sum of squares to the total sum of squares from the analysis of variance, was also calculated as a measure of effect size and used to assess practical significance. Because these univariate analyses both revealed that standard deviations of least-square means were uniformly low for the dependent measures and they provided a method for adequate interpretation, subsequent multivariate analyses were not performed.

Real Data Study

Source Data

Operational test data from the Fall 2005 administration of the Online Computer Skills Assessment (OCSA) by North Carolina's Department of Public Instruction

Accountability Services/Test Development Section were used for this investigation. Data were from over 100,000 8th graders in this statewide, computer-based (non-adaptive) assessment. Motivation was not a serious issue. This assessment is a part of efforts to "prepare North Carolina students for 21st Century opportunities" (North Carolina Department of Public Instruction, 2008); this assessment is high-stakes because its successful completion is currently a high school graduation requirement. Eight different OCSA test forms were analyzed. The forms had been randomly assigned to examinees (i.e. spiraled within schools) during the operational test administration. The OCSA (3rd Edition) was administered completely online beginning in Fall 2005. The test is 54 items in length, with approximately half being multiple-choice (MC) items. The remaining items are performance-based and are arranged into problem-based item sets. MC items have up to four distractors, including the keyed answer response. All test items were dichotomously scored (right/wrong).

The 3rd edition of the OCSA is composed of items in six content-related strands: Societal/Ethical Issues (12-14 percent of the items), Database (22-25 percent), Spreadsheet (22-25 percent), Keyboard Utilization/Word Processing/Desktop Publishing (18-20 percent), Multimedia and Presentation (10-12 percent), and Telecommunications and Internet (10-12 percent; North Carolina Department of Public Instruction, 2008).

Data Analysis

General Procedures

All data checking, dataset manipulations, and item scoring were performed with PC-SAS (version 9.1). Similarly, item *p*-values, estimates of internal consistency reliability with coefficient α, estimates of item temporal parameters (α_i and β_i), and semi-partial correlation coefficients were obtained programmatically with PC-SAS modules. Item – total score correlation coefficients were found, and 3PL IRT modeling conducted, with BILOG MG (see Appendix A for the BILOG MG listing and three sample SAS program listings). Graphics were produced with Systat (version 7.0, Systat, Inc.).

Preliminary Data Checks and Dataset Manipulations

The total amount of time that each item was presented to each examinee was determined during the operational test administration. Of 106,583 examinees in the Fall 2005 administration, 16 examinees had incomplete item response time records due to mechanical or related reasons and were removed. 508 records with duplicate student identifiers were also removed. 56 records with the total presentation time missing were also excluded from analyses; visual inspection revealed that no examinees in this group responded to any item. Records containing 40 or more missing responses (approximately 75 percent of the test items) were removed (86 examinees). The "complete" sample of students administered the OCSA was composed of 105,917 examinees.

Not all possibilities for duplicate records were checked. Particular students may have been assigned multiple unique identifiers by different administrative units (schools or Local Education Areas). However, the subset of students with multiple identifiers that 1) attended all administrative units as 8th graders, 2) were administered identical forms of the Computer Skills assessment, and 3) were assigned identifiers not modified during original dataset cleaning was considered minimal in size.

All item presentation times for all examinees were cumulated and rounded to the nearest second (i.e., sum of time spent on the item, including review). Because times of item presentation for each item for every examinee were non-negative in the raw records, the following 2 X 2 table (Table 2) was constructed to interpret individual item presentation times.

Insert Table 2 about here

The total amount of time that items were presented was then determined for each examinee. When an item had a non-zero presentation time, but no overt mouse or keyboard response was made, presentation time was recoded as missing. Total presentation time, in seconds, was the sum of presentation times for items in which overt keyboarding responses were made. Total presentation time, defined in this way, is a more accurate alias for total item response time required for test completion. For almost all examinees, this differed substantially from the total amount of time recorded for a complete test administration. The official total amount of time included administration of test instructions, tutorials on screen navigation, and item presentation times not terminating with an overt keyboarding response.

Descriptive statistics on the newly-obtained total presentation time revealed that some examinees were allowed lengthy amounts of time to complete the test. To reduce influence of time-dependent effects (Glickman, Gray, & Morales, 2005), students taking more than 2.5 standard deviations above the mean to complete the test, as indicated by total presentation time, were removed from the dataset. This time-truncated dataset contained 103,751 students, and is used for the remainder of the analyses.

IRT Calibration

A 3-PL IRT local calibration was conducted in BILOG-MG using the time-truncated dataset. Proficiency scores (θ_I) were estimated for individual examinees using the expected a posteriori (EAP) estimation algorithm in BILOG-MG. Priors for the c_i parameters (based on a beta-binomial distribution) were consistent with the priors actually used by the North Carolina Department of Public Instruction to calibrate the items (Table 3). The convergence criterion was set at 0.0001. A maximum of 125 E-M cycles were specified, which included 75 Newton-Raphson iterations. An acceleration constant, used during the E-M cycles to speed up convergence, was set to 1.0. Rescaling of the proficiency scores, θ, to a unit distribution was suppressed.

θ_I-conditional response probabilities were calculated using 3PL item parameter estimates. Residual probabilities for each item i and individual j were determined by subtracting the item response probability from the dichotomized item score, that is $e_{ij}=u_{ij}-P_i(\theta_{Ij})$, where θ_{Ij} is an EAP estimate. Appendix B contains the complete residual variance-covariance matrix; note that the median item covariance approximates 0.0. To estimate the variance in these residuals accounted for by item response times, semipartial correlation coefficients were calculated between the individual response times, t_{ij}, and the residuals, e_{ij}.

Insert Table 3 about here

Local Item Independence

Summary statistics based on Yen's (1993) Q_3 statistic were used to assess possible violations of IRT's local item independence assumption (Reckase, Ackerman, & Carlson, 1988). The expected value of the summary Q_3 statistic (-0.02), indicating local item independence, was calculated as (-1/(n-1), a correction factor derived by Yen (1987). Q_3 statistics were then determined for each of the 1431 unique item pairs. The residuals for each pair ($e_{ij}=u_{ij}- P_i(\theta_{1j})$) were summed over examinees. If the correlation between a given pair of item residuals exceeded 0.05, prior research has suggested that this magnitude of residual covariance might indicate a possible violation of the local item independence assumption (Pommerich & Segall, 2008). The same criterion ($r_{ee'} > 0.05$) was used to indicate whether the mean residual correlation of an item pair exhibited a possible violation of the local item independence assumption.

Variance in Item Residuals ($e_{ij}=u_{ij}- P_i(\theta_{1j})$*) Explained by Item Response Times*

Across all examinees, the correlation between each item's residual as calculated above and all 54 item response times was determined. A semi-partial correlation coefficient was determined between an item's residual and its response time ($r_{e_{ij}t_{ij}}$) that provided a "purified" measure of the relationship. The variance in each item's residual uniquely explained by examinees' item response times was calculated as the square of the semipartial correlation coefficient. This provided an estimate of the amount of variance in the θ_j-conditional residual explained by the variance in that item's response time.

CHAPTER IV

RESULTS

Chapter IV presents results from the simulation and real data studies that attempt to answer the research questions concerning detection and impact. Across all items with the EAP θ estimates used here, results indicate that when the direct influence of τ on observed item responses is increased in strength, mean item semi-partial correlation coefficients ($r_{e_{ij}t_{ij}}$) detect the manipulation. The $r_{e_{ij}t_{ij}}$ increase in magnitude with a moderate effect size (η^2=0.45). Mean item semi-partial correlations are apparently not as sensitive to manipulations of the mediated indirect influence of τ on observed responses. Importantly, direct τ influences on observed responses are not reliably detected with a measure of local item independence, Q_3. Neither mean Q_3 estimates nor their variability change substantially with temporal manipulations.

Temporal and psychometric manipulations systematically impact EAP θ_l estimation, but only to a small extent. Slight decreases in RMSE statistics, revealing increased accuracy of EAP θ_l estimation, are obtained when both Item Discrimination (a_1) increases, and when the Direct τ Effect on observed item responses increases. In simulations, expected increases in Pearson correlations between τ and EAP θ_l estimates due to manipulation of the Direct τ Effect (a_2) are obtained. Compared to baseline a_2

conditions, linear increases in these correlations are observed as a_2 increases in magnitude.

In a real data study, classical and IRT item parameters from an operational assessment are shown. Also, analyses of local item independence and relationships between item response times and IRT residuals using semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ are shown.

Simulation Study

Experimental Checks

Results from several experimental checks, performed to examine whether response generation programs were functioning as intended, are shown in Table 4. This table indicates that the mean empirical values of several factors (mean Item Discrimination [a_1] and the direct influence of τ on u_i [a_2]) closely approximate the targets set as desired factor levels. The extent to which dichotomously-scored response matrices converged to a solution for the 3PL IRT unidimensional model is examined for each treatment condition in the 3 X 4 X 4 X 5 design. This is shown in table 5; of the 1800 individual response matrices, almost 97% (1741/1800) converged to a solution. Mean item parameters and θ_I estimates of treatment conditions in that design after calibration with a unidimensional 3PL model are provided in Table 6.

Insert Table 4 about here

Insert Table 5 about here

Insert Table 6 about here

Results from a final manipulation check, performed to determine if the indirect influences on u_i by $r_{\theta l \tau}$ correlations were being modeled as expected, are shown in Figure 7. This figure shows that when the target direct τ effect (a_2) is null, the resulting $r_{\theta l \tau}$ correlations approximate the target levels of the indirect τ effect ($r_{\theta l \tau}$ = -0.2, 0.0, 0.2, 0.4, and 0.6). Moreover, as the direct effect of τ on observed responses (a_2) increases in magnitude, large increases in correlations between EAP θ and τ estimates are found. Change in these correlations decreases as the basal target levels of $r_{\theta l \tau}$ correlation increase, reflected by functions with slopes diminishing in acceleration. This is expected due to the ceiling imposed by the maximum value of the correlation coefficient. After a four-factor analysis of variance was performed and overall statistical significance established, a Scheffé test comparing multiple treatment means reveals that every level of the a_2 factor significantly differs from every other (all ps<.01).

Insert Figure 7 about here

Detection of Temporal Effects

One goal of this simulation study is to determine whether item semi-partial correlations ($r_{e_{ij}t_{ij}}$, where e_{ij} is the residual error of measurement [item i, person j] and t_{ij} is the response latency) could be used to detect a possible confounded relationship between observed response accuracy and t_{ij}.

Figure 8 shows that mean item semi-partial correlation coefficients increase in magnitude as the Direct τ Effect on observed responses is increased. Strikingly, the variability in mean item semi-partial correlation coefficients decreases as temporal manipulations are strengthened. The increments in magnitude of mean item semi-partial correlations reach statistical significance ($F(3,1561)=1347.06$, $p<.0001$). Moreover, the Direct τ Effect has a moderate effect size ($\eta^2=0.45$). The R^2 for the four-factor model approximates 0.83, an indication of good model fit. All other main and interaction effects reaching statistical significance (*p*s<.0001), such as Test Length, the Indirect τ Effect, and the Test Length x Direct τ Effect interaction, have small effect sizes as measured by η^2 (0.15, 0.12, and 0.02, respectively). Manipulations of Item Discrimination (a_1) used in this study have negligible effects on mean item $r_{e_{ij}t_{ij}}$. In both Figures 8 and 9, error bars are the average standard deviation across replications at the lowest mean value of the Item Discrimination factor (a_1).

Insert Figure 8 about here

Influences from temporal manipulations are not seen with a measure of local item dependence, the Q_3 statistic. Figure 9 shows that mean estimates of local item independence across all treatment cell replications, as measured with an index of Q_3, are relatively stable following manipulations of temporal parameters, with large average standard deviations. The sole factor attaining statistical significance ($p<.0001$) is Test Length, with a large effect size ($\eta^2 = 0.87$). All other effects attaining statistical significance have negligible effect sizes. The R^2 for the four-factor model including the robust Test Length factor approximates 0.90, an indication of good model fit. Scheffé mean comparisons at the .01 level of statistical significance indicate that mean Q_3 indices differ from each other at each level of Test Length tested. Examination of these means indicates that the shortest Test Length has a slightly higher mean Q_3 index than that seen at longer Test Lengths. The mean difference in absolute terms, however, is small (0.01).

Insert Figure 9 about here

Impact on EAP θ_l (Proficiency) Estimation

Impact on θ_l estimates was examined with root mean square error (RMSE) statistics to determine parameter recovery of true θ estimates (Figure 10). As Item Discrimination increases, mean RMSE and its standard deviation decrease (when Item Discrimination is 0.50, mean RMSE is 0.0513 (SD=0.0027); at the highest level of Item Discrimination tested, mean RMSE is 0.0480 (SD=0.0016)). The Item Discrimination factor is statistically significant ($F(2,1561)=677.98$, $p<.0001$); a Scheffé test comparing treatment means indicates that mean RMSE at the several levels of Item Discrimination significantly differs ($p<.01$). Although the effect size is moderate ($\eta^2 = 0.28$), the absolute mean difference in RMSE due to varying Item Discrimination levels is small.

Slight mean differences in RMSE are also observed after varying the Direct τ Effect of τ on observed responses. As the levels of the a_2 factor increase in magnitude, mean RMSE decreases significantly ($F(3/1561)=310.58$, $p<.0001$). As the Direct τ Effect increases, mean RMSE and its standard deviation decrease (when $a_2 = 0.00$, mean RMSE is 0.0510 (SD=0.0028); at the highest level of a_2 tested, mean RMSE is

0.0481 (SD=0.0017). Mean RMSE decreases significantly across levels of the a_2 factor, as indicated by a Scheffé test of treatment mean comparisons (p<.01). The size of the a_2 effect is very small ($\eta^2 = 0.19$).

The Item Discrimination a_1 factor significantly interacts with Test Length (F(4/1561)=23.81, p<.0001). The effect size, however, is negligible ($\eta^2 = 0.05$). The size of all remaining main and interaction effects approximates 0.00.

Insert Figure 10 about here

Real Data Study

Distributional Characteristics of Item Response Times

The distribution of total test response times, summed across all item presentation times when overt responses were made, is shown in Figure 11. These times, reflecting response latencies, are pooled across test forms; Figure 11 shows the non-normal distribution of total test response times for the complete sample. The first four statistical moments are shown in Table 7. Figure 12 shows the distribution of total test response times after removal of 2166 students with total item response times greater than the mean value plus 2.5 standard deviations (7062 seconds [approximately 2 hours]; Glickman et al., 2005). Statistical moments for this truncated sample are likewise reported in Table 7.

Insert Figure 11 about here

Based on skewness and kurtosis values, as well as visual inspection, the distribution of total test response times from the truncated sample appears normal in shape. Substantial effects on the total test response time distribution by outliers appear to be limited to the upper percentiles (Table 8).

Insert Table 7 about here

Insert Figure 12 about here

Insert Table 8 about here

The distribution of total test scores, calculated after dichotomous item scoring, is shown in Figure 13. That truncation based on total test response times does not have a substantial effect on the underlying score distribution is shown by examining descriptive statistics of the total test score distribution (Table 9). Moreover, over 97% of those tested with the Online Computer Skills Assessment during Fall 2005 remained after data checks and dataset truncation based on total test response time.

Insert Figure 13 about here

Insert Table 9 about here

Classical Test and Item Statistics

Classical test statistics (coefficient α, standard error of measurement) were obtained after pooling across test forms using the complete and time-truncated datasets using examinees with no missing responses (Table 10). Classical item statistics (item *p*-values, and test total-item correlation coefficients [point biserial, serial]) were likewise computed after pooling across test forms (Table 11). Similar test and item statistics were also obtained for each of the 8 test forms. Classical test statistics by form are shown in Table 12; classical item statistics by form are in Appendix B (Tables B1-B8).

Insert Table 10 about here

Insert Table 11 about here

Insert Table 12 about here

Item Response Time Summaries

The patterns of item response latencies, inferred from item presentation times, were examined with scatterplots. As described previously (Schnipke & Scrams, 1997), these patterns were examined separately for incorrect and correct responses. Distributions of response times by raw score are presented for five items in Figures 14 through 18. These individual items have the following characteristics: largest amount of variance in the residual explained by item response time (item 9, Figure 14), an approximately 50% response probability (item 7, Figure 15), very difficult in terms of response probability (item 14, Figure 16), comparatively easy (item 23, Figure 17), and greatest residual (item 45, Figure 18).

Insert Figure 14 about here

Insert Figure 15 about here

Insert Figure 16 about here

Insert Figure 17 about here

Insert Figure 18 about here

Bimodal distributions in item response times, providing evidence of "rapid item response" and "solution behavior" (Wise & DeMars, 2006), are shown only for items 9 (Figure 14) and 14 (Figure 16). The magnitude of the semi-partial correlation coefficient for Item 9 is the largest observed on this test. These figures also show that, whereas increased item response times may lead to increased frequency of correct observed responses (item 45), this relationship is not true for all items (item 14).

Item and Examinee Characteristics

Item IRT Parameters

Table 13 shows, for the first 20 items on the assessment, 3PL IRT estimates of a_i, b_i, and c_i parameters with their standard errors. For each examinee, response probabilities for every item were calculated using these parameter estimates and the θ estimate. Item fit to the IRT model is shown by item χ^2. The residual probability for each item, the difference between the dichotomized raw score and response probability, was also calculated for each examinee; the mean residual across all examinees is shown (Table 13). 3PL IRT item characteristics for the remaining items are shown in Table 14. A summary of IRT item characteristics for the Fall 2005 NC Online Computer Skills Assessment is provided in Table 15.

Insert Table 13 about here

Insert Table 14 about here

Insert Table 15 about here

Figure 19 is a scatterplot of the 54 mean item residuals and the mean 3PL response probabilities computed for all examinees in the time-truncated data. With the exception of three items, this figure shows that easier items, those with higher mean response probabilities, tend to have higher mean residuals (greater than 0.01); while items with lower residuals have a broad range of response probabilities. This heteroscedasticity implies that errors in determining response probabilities are greater for easier items than for more difficult items. It also suggests that variance in the residuals may be related to one or more explanatory variables (Cai & Hayes, 2008).

Insert Figure 19 about here

Item Temporal Parameters

Table 16 shows mean time intensity and a temporal discrimination parameter for each item. Mean time intensity (β) was calculated as the average amount of time elapsed (in seconds) prior to response production across all responding examinees. The temporal discrimination parameter (α) is calculated as the reciprocal of the standard deviation of that elapsed time.

Insert Table 16 about here

Examinee EAP θ_j Estimates

Table 17 summarizes EAP θ_j estimates from this assessment. When examinees with near-perfect scores and those achieving the lowest scores are excluded, the ability distribution appears normally distributed, with a mean EAP θ_j near 0.0 (0.013) and a standard deviation approximating 1.05.

Insert Table 17 about here

Local Item Dependence (LID)

The IRT assumption of local item independence was assessed with the Q_3 statistic (Reckase, Ackerman, & Carlson, 1988; Reese, 1995; Yen, 1984). A Q_3 statistic was calculated for each of the 1431 unique item pairs from the 54-item test, and a linear correction factor was implemented. As Table 18 shows, the mean Q_3 approximates 0 with a low standard deviation, evidence that responses satisfy the assumption of local item independence and are unidimensional. On closer inspection (Table 19), responses to item pairs exhibiting high LID levels (above the 90^{th} percentile; adjusted Q_3 greater than 0.028) do not appear to satisfy this assumption.

Insert Table 18 about here

Insert Table 19 about here

Semi-Partial Correlation Coefficients

For each item, the relationship between the deviation of observed raw scores and the expected 3PL response probability (the residual, e_{ij}) with that item's response time t_{ij} was then examined. Tables 20 and 21 present such semipartial correlation coefficients, using the deviation between observed and expected values (the item residual) as the dependent variable. Individual item response times t_{ij} serve as explanatory variables. The semipartial correlation coefficient $r_{e_{ij}t_{ij}}$ estimates the relationship between the item residual and that item's mean response time after removing effects from all other explanatory variables. When squared, the item semipartial correlation is an estimate of the amount of variance explained in the residual by that item's mean response time. Semipartial correlation coefficients for the first 20 items are shown in Table 20; for the remainder of the items, in Table 21.

Insert Table 20 about here

Insert Table 21 about here

Table 22 shows the linear relationships between item residuals and semipartial correlations ($r_{e_{ij}t_{ij}}$) with 3PL IRT item parameters (Discrimination [a_i], Difficulty [b_i], and Pseudo-guessing [c_i]). A moderate negative correlation is seen between item residuals and item difficulty, indicating that the magnitude of the residual is inversely related to the item difficulty parameter. Figure 20 shows a modest relationship between item difficulty as indexed by mean 3PL response probability for each item and that item's semipartial correlation ($r_{e_{ij}t_{ij}}$). This relationship is represented in Table 22 as a moderate linear correlation between squared semipartial correlations and 3PL IRT response probability, indicating that the amount of variance explained by the semipartial correlation ($r_{e_{ij}t_{ij}}$) may also be inversely related to the value of the item difficulty parameter.

Insert Table 22 about here

Insert Figure 20 about here

CHAPTER V

DISCUSSION

Relationships between Item Response Time and Response Accuracy

Detecting the occurrence of a speed-accuracy tradeoff in responses to items on a psychometric assessment, and then quantifying the magnitude of that phenomenon, is not easily performed without ambiguity. Results from the present investigation indicated that mean effects on item response accuracy due to manipulations of an examinee pacing parameter (τ) could be detected with item semi-partial correlation coefficients. These coefficients estimated the magnitude of the linear relationship between item response times and their residual errors, determined after response probabilities were estimated with a unidimensional 3PL IRT model.

As the direct effect of τ on observed responses was strengthened in simulations, item semipartial correlation coefficients systematically increased. In addition, a reduction in the variability of the semipartial correlation coefficients also occurred. Although the mean increase was small in absolute terms, the effect size was moderate (η^2=0.45). Although manipulation of the Direct τ Effect on observed item responses resulted in small mean changes in item semi-partials, such a relatively strong η^2 statistic indicates that the Direct τ Effect may be a useful heuristic in other, perhaps clinical, settings.

Manipulations of item quality by varying mean Item Discrimination, and of classical test reliabilities by varying Test Length, have effects negligible in size on these item semi-partial correlation coefficients.

The impacts on EAP θ_l estimation were studied in two ways, with RMSE statistics and with correlations between the examinee latent traits. Analysis of recovery of true θ estimates with RMSE statistics revealed a slight Item Discrimination (a_l) effect: As Item Discrimination increased, accuracy of EAP θ_l estimation increased with an effect size of 0.28. Actual decreases in mean RMSE were very small. Also, recovery of true θ estimates was influenced slightly by increases in the Direct τ Effect. Recovery as indexed by RMSE statistics increased slightly with an effect size of $\eta^2 = 0.19$; differences in mean RMSE due to these manipulations were again small.

Correlations between τ and EAP θ_l estimates increased systematically with increases in the τ direct effect. Because τ distributions were not substantially altered by temporal manipulations (Figures 4 to 6), this finding implies that the distributions of EAP θ_l estimates more closely approximated τ distributions as the Direct τ Effect was strengthened.

Results from the real data study indicated that in the absence of an overall speed-accuracy relationship, the semipartial correlation coefficient $r_{e_{ij}t_{ij}}$ at the item level could serve to indicate that responses to specific items are influenced by this relationship. For instance, item 9 on the NC Online Computer Skills Assessment had the greatest item semi-partial correlation in magnitude (-0.23). That some item residuals correlated moderately with item response times suggested that observed responses to particular items were influenced somewhat by temporal factors.

Relationships between Q_3 and τ Manipulations

Mean Q_3, a statistic used to assess the magnitude of local item independence, approached zero as Test Length increased, but these item-pair statistics were not meaningfully influenced by the other manipulations (Item Discrimination [a_1], the Direct τ Effect on observed responses [a_2], or the Indirect τ Effect due to the $r_{\theta l \tau}$, mediated through a_1). The a_2 x Test Length and Indirect τ Effect x Test Length interactions had negligible effect sizes. Indeed, nearly all of the variation in Q_3 was explained by the very robust Test Length factor: Nearly 99% of the total variation in Q_3, as indicated by R^2, was explained by a general linear model that excluded all temporal factors and their interactions.

The most important finding from this investigation was that the mean magnitude of item semi-partial correlation coefficients did vary with the Direct τ Effect, compared to effects of this factor on Q_3 magnitudes. Mean semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ did correlate highly with mean Q_3 statistics. However, the present results indicated that these two measures provided non-redundant information. Q_3 statistics assessed relationships between residual errors in each unique pair of items on a given test, providing an index of local item dependence. In this research, item semi-partial correlation coefficients assessed relationships between residual errors and item response time t_{ij}, a source of supplementary information causally related to an examinee latent trait, τ.

The aggregated Q_3 index used was not effective in detecting occurrences of local item dependence due to varying the magnitude of the Direct τ Effect on response accuracy. Even should this index have been efficacious, isolating and identifying the specific cause of LID would have been problematic. Influences on response accuracy due to the Direct τ Effect were detected with item $r_{e_{ij}t_{ij}}$ after unidimensional IRT modeling. Advantages to using item semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ in lieu of mean Q_3 statistics for the detection and subsequent interpretation of temporal effects include:

1. The item semi-partial correlation coefficient expresses the relationship between residual errors, e_{ij}, and item response time, t_{ij}. Because t_{ij} is causally related to the examinee parameter τ, this relationship serves to isolate an item-specific τ effect that explains at least a portion of the residual error, e_{ij} (Luecht, personal communication).
2. Q_3 statistics are based on residuals from responses to item pairs. Responses to both items in the pair would have to be affected by τ in the same direction in order for Q_3 statistics to detect τ effects.
3. Item semi-partial correlation coefficients can be estimated with common statistical packages, and are easily interpretable.

Real Data Results

A criterion was established for modeled semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ for the detection of substantive relationships between item *i*'s residual and temporal phenomena (mean $r_{e_{ij}t_{ij}} \geq |0.15|$). The usefulness of this criterion to detect the operation of temporal phenomena was examined in a real data study. Student responses to items on the NC OCSA were used. In this research, total test response times were calculated for each student, and students with extremely long total test response time (slowly-responding students) were systematically removed as outliers. This range restriction on total response time was imposed to reduce spurious relationships due to severe pacing effects.

Item 9 on the third edition of the NC OCSA from the Fall 2005 administration had the greatest semi-partial correlation coefficient $r_{e_{ij}t_{ij}}$ in magnitude, -0.23. Item 32 had the highest positive semi-partial correlation coefficient $r_{e_{ij}t_{ij}}$, 0.11. The magnitude of the median coefficient was far lower, indicating that responses to most items were without apparent temporal biases. That item responses were locally independent was supported by Q_3 statistics (Table 18). Moreover, in the real data study, the largest semipartial correlation was actually negative in sign (Item 9, Table 20); in simulations at the 60-item Test Length, strongly negative semipartial correlation coefficients were more likely obtained at the lowest levels of the a_2 direct τ effect with a low or negative correlation between the latent traits (Figure 8). A negative correlation between item residuals and pacing may partially explain the score histogram of Item 9 (Figure 14), which was a relatively easy item (b_9=-1.448). Students responding slowly to this item were less likely to answer correctly, and so had larger item residuals, than students responding quickly.

Variance accounted for was quantified by squared semipartial correlation coefficients. Item 9 was an exemplar item; a portion of the variance in item residuals, calculated as deviations between observed item responses and θ_j-conditional 3PL response probabilities, could be explained by variance in item response times.

Analysis of local item independence using the Q_3 statistic revealed little evidence for possible violations of this crucial IRT assumption. Correlations between pairs of item residuals did increase non-linearly in Q_3 percentiles greater than 90.

Assumptions and Limitations

IRT Modeling

Standard IRT procedures make strong assumptions, as elaborated by Hambleton and Swaminathan (1985). These include assumptions concerning local independence of item responses conditional on θ, and invariance of IRT parameters across populations and items. As a corollary to the local item independence assumption, these procedures also assume that responses to test items are unidimensional (Hambleton & Swaminathan, 1985), implying that item responses are not substantially influenced by nuisance factors such as examinee speededness (τ, van der Linden, 2005). Moreover, IRT models assume that a dichotomous item's ICC, showing the relationship between the probability of response to a specific item and θ, represents a normal ogive function only when that relationship is both normal in form and linear (Samejima, 1997b, p. 472). Empirical results from the real data study may be limited in their generalizability due to these strong underlying IRT assumptions.

Test Performance Assumptions

IRT models have been developed to produce item and person estimates that are useful and invariant across samples of examinees. At least two variables can contribute to inaccurate examinee estimates: 1) Differential speededness of examinees (Hadadi & Luecht, 1998; Lord, 1980; van der Linden, 2005), and 2) motivational factors (Kong et al., 2007). One method of controlling the potential temporal confound has been to omit responses of individuals providing evidence of examinee speededness from IRT calibrations (Lord, 1980). Because this method cannot distinguish between high-ability examinees responding quickly but inaccurately due to time pressures and examinees with low ability levels (Schnikpe & Scrams, 1999), inaccuracies in θ estimates may result. An omission method has been proposed for those item responses where less than optimal examine motivation, as indicated by item response times, is exhibited (Wise & DeMars, 2006). Omitting such responses from IRT calibrations is shown in the Solution Behavior model (Wise & DeMars, 2006). However, this method may also introduce inaccuracies in θ estimates: it also cannot distinguish between high-ability examinees responding quickly but inaccurately from those examinees with low ability levels.

Q_3 as a Measure of Local Item Dependence

Huynh et al. (1995) examined three methods of determining the magnitude of local item dependence. Yen's Q_3 statistic was one of these measures; all statistics performed similarly on an operational assessment (the Maryland School Performance Assessment Program) for the identification of item clusters showing a response dependency. The authors noted that these all assume that inter-item correlation matrices remain constant across all levels of θ.

Univariate Normality Assumptions

The traditional IRT ICC represents the normal ogive model given certain normality and linearity assumptions (Samejima, 1997b). She states that an item ICC will approximate the normal ogive only when the distribution of response probabilities given the latent trait being measured is normal in form, and the regression of response probabilities on θ is linear (Samejima, 1997b). Several authors have observed that statistical univariate normality assumptions may not be appropriate for the mathematical modeling of multi-dimensional aspects of human behavior (Maris, 1993; Samejima, 1997b, p. 490). This observation becomes relevant should latent temporal variables act as a source of nuisance variation with θ for the production of observed responses to test items (van der Linden, 2005).

Research Directions

There are several lines of inquiry that can be pursued based on this investigation. The first issue that might be studied is the differential measurement of underlying latent traits in examinees due to the presence of a pacing construct. This raises construct validity concerns because significant confounding of θ and τ can influence observed item response accuracy of examinees, and thus affect inferences drawn due to test performance. The magnitude of the confounded relationship between residual errors in θ estimation and τ, as reflected in item response times, is not obtained with standard unidimensional IRT models. The semi-partial correlation between residual errors of measurement and item response times can be used to quantify potential confounded relationships for individual test items. Items with semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ exceeding |0.15| should be closely examined for evidence of possible confounds due to examinee pacing.

A second possibility for future research concerns the investigation of populations characterized by abnormal response pacing. Whether response accuracy interacts with observed response time distributions in such populations, a possibility shown in Figure 2 and explored with normal distributions of simulees here, might provide additional information concerning these temporal processes. Because IRT models assume that underlying trait levels are normally distributed in the population (Samejima, 1997b, p. 472; Hambleton & Swaminathan, 1985), progress in uncovering inferred temporal processes underlying observed response production may be made primarily by studying examinee misfit to IRT models. Determining how cognitive diagnostic models might interface with temporal modeling would be an interesting endeavor.

A third line of inquiry involves following up empirical findings from the current research. In a state-wide operational test used in this study, a large number of item semi-partial correlation coefficients $r_{e_{ij}t_{ij}}$ were negative in direction; the items with the lowest IRT *b*-parameter had item semi-partial correlation coefficients of the greatest magnitude. Perhaps in another set of future simulations, tests could be developed with items characterized by lower *b*-parameters, and administered to groups of simulees characterized with moderately lower θ_l estimates and demonstrating longer item response times. The higher residuals may correlate negatively with the individual item response times, mimicking the results seen here. In a second set of simulations, median splits based on θ_l estimates might be used to determine differential τ effects. Other statistical modeling procedures might also be used.

A last line of possible inquiry would concentrate on item parameter estimates and amounts of statistical item information from unidimensional IRT models. For instance, it may be that a substantial part of the variance in item pseudo-guessing parameters is accounted for by temporal phenomena associated with response production. This is supported by the data in Table 6: a_i parameters were exaggerated by the unidimensional modeling procedure, suggesting that the amount of item information detected by the unidimensional modeling procedure had increased. Further, although item responses were generated with a constant c_i parameter of 0.15, expected c_i parameters from the unidimensional model were routinely almost twice that amount. As online testing becomes more prevalent, temporal item response measures may be a source of critical additional diagnostic information concerning examinee characteristics. Clearly, further research is needed to investigate these possibilities and their validity ramifications.

REFERENCES

Allen, M.J., & Yen, W.M. (2002). *Introduction to Measurement Theory.* Prospect Heights, IL: Waveland Press, Inc.

Bartram, D. (2006). Testing on the Internet: Issues, Challenges and Opportunities in the Field of Occupational Assessment. In D. Bartram and R.K. Hambleton (eds.), *Computer-Based Testing and the Internet.* Hoboken, NJ: John Wiley & Sons, Ltd.

Cai, L., & Hayes, A.F. (2008). A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics, **33***, 21-40.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

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

Epstein, J.N., Goldberg, N.A., Connors, C.K., & March, J.S. (1997). The effects of anxiety on continuous performance test function in an ADHD clinic sample. *Journal of Attention Disorders, **2***, 45-52.

Fischer, G.H., & Kisser, R. (1983). Notes on the exponential latency model and an empirical application. In H. Wainer and S. Messick (eds.), *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Glickman, M.E., Gray, J.R., & Morales, C.J. (2005). Combining speed and accuracy to assess error-free cognitive processes. *Psychometrika, **70***, 405-425.

Hadadi, A., & Luecht, R.M. (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine, **73**,* S47-S50.

Hambleton, R.K. (2006). Psychometric models, test designs and item types for the next generation of educational and psychological tests. In D. Bartram and R.K. Hambleton (eds.), *Computer-Based Testing and the Internet.* Hoboken, NJ: John Wiley & Sons, Ltd.

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

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Huynh, H., Michaels, H.R., & Ferrara, S. (1995). A comparison of three statistical procedures to identify clusters of item with local dependency. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Iramaneerat, C., Myford, C.M., & Yudkowsky, R. (2006). Item dependency in an objective structured clinical examination. Paper presented at the 25th International Objective Measurement Workshop, Berkeley, CA.

Kaskowitz, G.S., & De Ayala, R.J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, **25***, 39-52.

Kong, X.J., Wise, S.L., & Bhola, D.S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, **67***, 606-619.

Lee, Y-W. (2000). Examining passage-related local item dependence (LID) using Q3 indices in an EFL reading comprehension test. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28).

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Luce, R.D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization.* New York, NY: Oxford University Press.

Luecht, R.M. (2008). MIRTGEN 2.0 with Response Times. Greensboro NC: University of North Carolina at Greensboro.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, **58***, 445-469.

McDonald, R. (1997). Normal ogive multidimensional model. In W.J. van der Linden, and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

North Carolina Department of Public Instruction. (2008). Test of computer skills (Graduation – Requirement) [Electronic Version]. Retrieved September 12, 2008 from http://www.dpi.state.nc.us/accountability/testing/computerskills/

Pommerich, M., & Segall, D.O. (2008). Local dependence in an operational CAT: Diagnosis and implications. *Journal of Educational Measurement, **45***, 201-223.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M.D., Ackerman, T.A., & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, **25***, 193-203.

Reese, L.M. (1995). The impact of local dependencies on some LSAT outcomes. Law School Admission Council, LSAC Research Report Series, LSAC-R-95-02.

Roskam, E.E. (1997). Models for speed and time-limit tests. In W.J. van der Linden, and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

Samejima, F. (1988). Comprehensive latent trait theory. *Behaviormetrika, **24***, 1-24.

Samejima, F. (1997b). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika, **62***, 471-493.

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

Schnikpe, D.L., & Scrams, D.J. (1999). Exploring issues of test taker behavior: Insights gained from response-time analyses. Law School Admission Council, Computerized Testing Report, LSAC-R-98-09.

Schweizer, K. (1998). Complexity of information processing and the speed-ability relationship. *Journal of General Psychology, **125,*** 89-102.

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652-654.

Swets, J.A., Tanner, W.P., & Birdsall, T.G. (1961). Decision processes in perception. *Psychological Review, 68*, 301-340.

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, 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. (2005). *Linear models for optimal test design.* New York, NY: Springer.

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. (2007). A hierarchical framework for modeling of speed and accuracy on test items. *Psychometrika, **72***, 287-308.

van der Linden, W.J, Breithaupt, K., Chuah, S.C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, ***44***, 117-130.

van der Linden, W.J., & Hambleton R.K. (1997). Item response theory: Brief history, common models, and extensions. In W.J. van der Linden, and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, ***23***, 195-210.

van der Linden, W.J., Scrams, D.J., & Schnipke, D.L. (2003). Using response-time constraints in item selection to control for differential speededness in computerized adaptive testing. Law School Admission Council, Computerized Testing Report 98-03.

Verhelst, N.D., Verstralen, H.H.F.M., & Jansen, M.G.H. (1997). A logistic model for time-limit tests. In W.J. van der Linden, and R.K. Hambleton (eds.), *Handbook of Modern Item Response Theory*. New York, NY: Springer-Verlag.

Wenger, W.J. (2005). Models for the statistics and mechanisms of response speed and accuracy. *Psychometrika, **70***, 383-388.

Wise, S.L., & DeMars, C.E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, ***43***, 19-38.

Wise, S.L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, **18***, 163-183.

Wise, S.L., Kong, X.J., & Pastor, D.A. (2007). Understanding correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, **5***, 245-262.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, **8***, 125-146.

Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item independence. *Journal of Educational Measurement, **30***, 187-214.

Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model. Educational Testing Service, TR-95-02.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2002). BILOG-MG. Lincolnwood IL: Scientific Software International.

Appendix A. Tables and Figures

Table 1. Specifications for the Design of the Simulation Study

		Factor 1 (Test Length, 20 Items)												Factor 1 (Test Length, 30 Items)												Factor 1 (Test Length, 60 Items)											
F2: a_1		0.50				0.75				1.00				0.50				0.75				1.00				0.50				0.75				1.00			
F3: a_2		0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75
F4:	-.20																																				
$\rho[\theta_1 \tau]$	.00																																				
	.20																																				
	.40																																				
	.60																																				

Note: A 3X3X4X5 fully-crossed design is used in this simulation study (10 replications per condition, 1000 simulees per replication).
Factor 1 is Test Length (3 levels); Factor 2 is item discrimination (3 levels, a_1). Factor 3 is the mean direct effect of τ on u_i (4 levels, a_2).
Factor 4 ($\rho[\theta_1 \tau]$) is the indirect effect of τ on u_i through correlation with θ_1; there are 5 levels of correlation in this simulation.
True θ_1 and τ estimates were generated from 550 items with a constant item discrimination parameter (2.75), d_i values ranging from –3 to 3, and c_i at 0.0.
These estimates were generated with a 0.0 correlation between the latent traits, a_2 parameters of 0.0, and negligible β_i and α_i parameters.

Table 2. Interpretation of Reported Item Timing Data Based on Examinee Responses

	Reported Item Presentation Time	
	Time = 0	Time > 0
Response Present	Rapid item response rounding to 0 seconds, t_{ij}=0	Elapsed time for response production, t_{ij}=reported time
Response Absent	Item not presented during administration, t_{ij}=missing	No overt keyboarding response after item presented, t_{ij}=missing

Table 3. Priors for the Estimation of the Pseudo-Guessing (c_i) IRT Parameter During 3PL Modeling: Real Data Study

Item	Location (α)	Dispersion (β)	Item	Location (α)	Dispersion (β)	Item	Location (α)	Dispersion (β)
1	6.0	16.0	19	6.0	16.0	37	1.1	10000.0
2	1.1	10000.0	20	6.0	16.0	38	6.0	16.0
3	1.1	10000.0	21	1.1	10000.0	39	1.1	10000.0
4	1.1	10000.0	22	1.1	10000.0	40	1.1	10000.0
5	6.0	16.0	23	6.0	16.0	41	6.0	16.0
6	6.0	16.0	24	6.0	16.0	42	6.0	16.0
7	6.0	16.0	25	1.1	10000.0	43	1.1	10000.0
8	1.1	10000.0	26	1.1	10000.0	44	6.0	16.0
9	1.1	10000.0	27	6.0	16.0	45	1.1	10000.0
10	6.0	16.0	28	6.0	16.0	46	6.0	16.0
11	6.0	16.0	29	6.0	16.0	47	1.1	10000.0
12	6.0	16.0	30	6.0	16.0	48	6.0	16.0
13	6.0	16.0	31	1.1	10000.0	49	6.0	16.0
14	1.1	10000.0	32	1.1	10000.0	50	6.0	16.0
15	1.1	10000.0	33	6.0	16.0	51	1.1	10000.0
16	1.1	10000.0	34	1.1	10000.0	52	1.1	10000.0
17	1.1	10000.0	35	1.1	10000.0	53	1.1	10000.0
18	6.0	16.0	36	1.1	10000.0	54	6.0	16.0

Note: In this table, α and β designate location and dispersion parameters in a beta-binomial distribution.

Table 4. Mean Empirical Parameters from a Multidimensional IRT Model after Pooling across Treatment Conditions

Treat-ment Levels	N	Mean Item Discrimination (a_1)	Mean Direct Effect of θ_2 (a_2)	Mean d_i (Distance Parameter)	Mean c_i (Lower Asymptote)
Across levels of Test Length (Factor 1)					
20	60	0.75 (0.21)	0.38 (0.28)	0.05 (0.07)	0.15
30	60	0.75 (0.20)	0.38 (0.28)	-0.00 (0.05)	0.15
60	60	0.75 (0.21)	0.38 (0.28)	0.00 (0.04)	0.15
Across levels of Item Discrimination (a_1, Factor 2)					
0.50	60	**0.50 (0.01)**	0.38 (0.28)	0.02 (0.07)	0.15
0.75	60	**0.75 (0.01)**	0.38 (0.28)	0.01 (0.06)	0.15
1.00	60	**1.00 (0.01)**	0.37 (0.28)	0.02 (0.05)	0.15
Across levels of the Direct τ Effect ($a_{2,}$ Factor 3)					
0.00	45	0.75 (0.21)	**-0.00 (0.01)**	0.01 (0.06)	0.15
0.25	45	0.75 (0.27)	**0.25 (0.01)**	0.03 (0.05)	0.15
0.50	45	0.78 (0.20)	**0.50 (0.01)**	0.04 (0.06)	0.15
0.75	45	0.78 (0.21)	**0.75 (0.01)**	-0.01 (0.05)	0.15
Across levels of the Indirect Effect of the $\theta_1 \tau$ correlation (Factor 4)					
-0.20	36	0.75 (0.21)	0.38 (0.28)	0.02 (0.06)	0.15
0.00	36	0.75 (0.21)	0.38 (0.28)	0.02 (0.06)	0.15
0.20	36	0.75 (0.21)	0.37 (0.28)	0.01 (0.06)	0.15
0.40	36	0.75 (0.21)	0.38 (0.28)	0.02 (0.06)	0.15
0.60	36	0.75 (0.21)	0.38 (0.28)	0.02 (0.06)	0.15

Notes: Parameter values are means for the treatment conditions pooled across r all other factors. N is the number of treatment conditions in each factor level; each N is the item parameter mean from converging replications in each treatment. Each replication is composed of 1000 simulees.
The bold-faced statistics indicate that empirical values for Factors 2 and 3 are within rounding error of their respective targets.
Average standard deviations are in parentheses. Mean d_i is analogous to the item difficulty parameter in unidimensional models (Luecht, 2008).

Table 5. Number of Non-converging Response Matrices in Each Treatment Condition, Simulation Study

		Factor 1 (Test Length, 20 Items)											
F2: a_1		0.50				0.75				1.00			
F3: a_2		0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75
F4:	-0.2	1	0	0	0	1	0	0	1	1	0	0	0
$\rho[\theta_I \tau]$	0.0	0	1	1	1	0	1	0	0	0	0	0	0
	0.2	1	2	0	1	0	0	0	1	0	1	0	0
	0.4	0	0	0	0	0	0	0	0	0	1	0	0
	0.6	2	0	0	0	0	0	0	1	0	0	0	1

		Factor 1 (Test Length, 30 Items)											
F2: a_1		0.50				0.75				1.00			
F3: a_2		0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75
F4:	-0.2	1	0	0	0	0	0	0	0	0	0	0	0
$\rho[\theta_I \tau]$	0.0	0	0	1	0	0	1	0	0	0	0	0	0
	0.2	0	0	0	0	1	0	0	0	0	0	0	0
	0.4	0	1	0	0	0	0	0	0	0	0	0	0
	0.6	0	0	0	0	0	0	0	0	0	0	0	0

		Factor 1 (Test Length, 60 Items)											
F2: a_1		0.50				0.75				1.00			
F3: a_2		0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75	0.0	0.25	0.5	0.75
F4:	-0.2	2	1	0	0	0	1	2	0	0	0	1	0
$\rho[\theta_I \tau]$	0.0	3	0	2	0	0	0	0	0	0	0	0	0
	0.2	3	3	0	1	0	0	0	0	0	0	0	1
	0.4	3	1	0	0	1	0	0	0	0	0	0	0
	0.6	5	2	0	1	1	0	0	0	1	0	0	0

Notes: A 3X3X4X5 fully-crossed design is used in this simulation study (10 replications per condition, 1000 simulees per replication).
Factor 1 is Test Length (3 levels); Factor 2 is item discrimination (3 levels, a_1). Factor 3 is the mean direct effect of τ on u_i (4 levels, a_2).
Factor 4 ($\rho[\theta_I \tau]$) is the indirect effect of τ on u_i through correlation with θ_I; there are 5 levels of correlation in this simulation.
True θ_I and τ estimates were generated from 550 items with a constant item discrimination parameter (2.75), d_i values ranging from –3 to 3, and c_i at 0.0.
These estimates were generated with a 0.0 correlation between the latent traits, a_2 parameters of 0.0, and negligible β_i and α_i parameters.

Table 6. Empirical Means of IRT Parameters from Unidimensional Modeling after Pooling across Treatment Conditions

Treatment Levels	N	a_i	SE, a_i	b_i	SE, b_i	c_i	SE, c_i
		Across levels of Test Length (Factor 1)					
20	60	1.13	0.23	0.23	0.13	0.29	0.04
30	60	1.12	0.23	0.28	0.16	0.28	0.04
60	60	1.12	0.23	0.28	0.16	0.28	0.04
		Across levels of Item Discrimination (a_1, Factor 2)					
0.50	60	0.93	0.17	0.39	0.16	0.37	0.04
0.75	60	1.12	0.17	0.24	0.10	0.28	0.03
1.00	60	1.32	0.16	0.16	0.05	0.25	0.02
		Across levels of the Direct τ Effect (a_2, Factor 3)					
0.00	45	0.96	0.18	0.379	0.17	0.31	0.04
0.25	45	1.04	0.18	0.273	0.15	0.30	0.03
0.50	45	1.16	0.19	0.219	0.12	0.27	0.03
0.75	45	1.33	0.19	0.185	0.08	0.25	0.02
		Across levels of the Indirect Effect of the $\theta_1 \tau$ correlation (Factor 4)					
-0.20	36	1.02	0.18	0.30	0.13	0.30	0.04
0.00	36	1.08	0.19	0.28	0.15	0.29	0.04
0.20	36	1.12	0.22	0.27	0.17	0.28	0.04
0.40	36	1.17	0.24	0.24	0.15	0.28	0.04
0.60	36	1.22	0.27	0.24	0.15	0.27	0.04

Notes: Parameter values are means for the treatment conditions pooled across all other factors. N is the number of treatment conditions in each factor level; each N is the item parameter mean of converging replications in each treatment. Each replication is composed of 1000 simulees.

Table 7. Statistical Moments for Total Test Response Times during the Fall 2005 Administration of North Carolina's Online Computer Skills Assessment (NC OCSA)

	N	Mean	Standard Deviation	Skewness	Kurtosis
Complete sample	105917	3915.4	1258.6	1.0	3.4
Truncated sample	103751	3826.0	1094.3	0.3	0.2

Note: Item response times are rounded to the nearest second; means and standard deviations are based on these values.

Table 8. Percentiles of Total Test Presentation Time during the Fall 2005 Administration of the NC OCSA

Percentile	Complete Sample	Truncated Sample
1	1321	1305
5	2210	2201
10	2571	2560
50	3757	3729
90	5472	5308
95	6140	5831
99	7841	6650

Note: Item response times are rounded to the nearest second.

Table 9. Statistical Moments for Total Test Scores during the Fall 2005 Administration of the NC OCSA

	N	Mean	Standard Deviation	Skewness	Kurtosis
Complete dataset	105917	28.2	10.65	-0.12	-0.77
Truncated dataset	103751	28.3	10.66	-0.13	-0.77

Note: Means and standard deviations are from sums of dichotomized item scores (0=incorrect, 1=correct) across all 54 items. Items with missing responses are excluded from this calculation.

Table 10. Classical Test Statistics, Fall 2005 Administration of the NC OCSA

	N	Coefficient α	SEM
Complete dataset	67010	0.89	3.53
Truncated dataset	65541	0.88	3.69

Note: SEM is the standard error of measurement; only examinees with no missing item responses are included in *N*, the sample size.

Table 11. Classical Item Statistics: Time-Truncated Data from Fall 2005 Administration of the NC OCSA: *N*=103751

		Item-Total *r*				Item-Total *r*				Item-Total *r*	
Item	*p*	Pearson	Serial	Item	*p*	Pearson	Serial	Item	*p*	Pearson	Serial
1	62.5	0.225	0.287	19	61.8	0.378	0.482	37	73.2	0.555	0.746
2	72.5	0.495	0.663	20	68.9	0.374	0.489	38	60.9	0.337	0.428
3	27.6	0.423	0.565	21	34.1	0.425	0.549	39	54.7	0.475	0.597
4	59.9	0.463	0.587	22	77.8	0.349	0.488	40	79.8	0.358	0.511
5	48.8	0.274	0.344	23	82.4	0.438	0.644	41	57.5	0.438	0.553
6	61.6	0.282	0.359	24	42.3	0.242	0.305	42	43.5	0.278	0.350
7	51.5	0.313	0.392	25	56.0	0.418	0.526	43	73.9	0.553	0.747
8	62.7	0.374	0.478	26	31.6	0.502	0.656	44	65.7	0.425	0.549
9	76.8	0.347	0.480	27	63.2	0.467	0.598	45	80.5	0.460	0.660
10	69.8	0.372	0.490	28	74.8	0.271	0.369	46	50.8	0.457	0.573
11	43.2	0.303	0.382	29	74.4	0.349	0.473	47	29.1	0.540	0.715
12	75.2	0.169	0.231	30	41.3	0.222	0.280	48	33.9	0.273	0.354
13	23.4	0.165	0.229	31	36.1	0.430	0.552	49	64.3	0.282	0.362
14	10.6	0.301	0.506	32	29.0	0.447	0.593	50	60.5	0.445	0.565
15	22.4	0.374	0.521	33	41.9	0.329	0.415	51	16.5	0.365	0.545
16	66.1	0.475	0.615	34	33.6	0.572	0.741	52	39.7	0.490	0.621
17	29.7	0.446	0.589	35	37.2	0.448	0.572	53	43.3	0.631	0.795
18	48.8	0.329	0.412	36	49.3	0.578	0.724	54	52.1	0.288	0.361

Note: *p* is the percent responding correctly, *N* is the total number of examinees in the time-truncated dataset, and *r* is correlation

Table 12. Classical Test Statistics by Form, Fall 2005 Administration of the NC OCSA

Form	*N*	Coefficient α	SEM
1	10996	0.88	3.60
2	10878	0.89	3.59
3	10941	0.88	3.64
4	10947	0.89	3.63
5	8175	0.88	3.66
6	8274	0.88	3.65
7	2689	0.88	3.60
8	2641	0.89	3.58

Note: SEM is Standard Error of Measurement; only examinees with no missing responses are included in *N*, the number of examinees administered each form.

Table 13. 3PL IRT Item Characteristics: First 20 Items of the NC OCSA, Fall 2005 Administration									
Item	Discrim-ination (a_i)	SE, a_i	Difficulty (b_i)	SE, b_i	Pseudo-guessing (c_i)	SE, c_i	Item χ^2	Mean Probability ($P(\theta_{1j})$)	Mean Residual
1	0.709	0.022	0.794	0.025	0.444	0.006	29.2	0.626	-0.001
2	0.909	0.007	-0.862	0.007	<0.001	<0.001	244.5	0.720	0.024
3	0.791	0.007	0.969	0.007	<0.001	<0.001	103.5	0.280	0.002
4	0.736	0.006	-0.389	0.006	<0.001	<0.001	256.9	0.597	0.008
5	0.666	0.016	0.865	0.019	0.250	0.006	29.4	0.489	0.001
6	0.392	0.007	-0.658	0.053	0.040	0.015	23.0	0.612	0.007
7	0.767	0.016	0.669	0.016	0.258	0.006	54.7	0.516	0.002
8	0.555	0.005	-0.631	0.009	<0.001	<0.001	233.6	0.623	0.007
9	0.571	0.005	-1.448	0.014	<0.001	<0.001	452.0	0.760	0.015
10	0.809	0.015	-0.206	0.025	0.312	0.009	186.6	0.697	0.004
11	0.791	0.017	0.961	0.013	0.213	0.005	64.4	0.435	0.001
12	0.258	0.005	-2.555	0.063	0.024	0.011	796.7	0.748	0.010
13	0.811	0.027	2.094	0.028	0.154	0.003	116.2	0.236	-0.001
14	0.804	0.010	1.992	0.017	<0.001	<0.001	67.2	0.109	-0.003
15	0.733	0.007	1.286	0.010	<0.001	<0.001	155.8	0.227	0.001
16	0.806	0.006	-0.626	0.007	<0.001	<0.001	925.0	0.657	0.014
17	0.826	0.007	0.851	0.007	<0.001	<0.001	121.6	0.301	-0.002
18	0.770	0.016	0.694	0.015	0.224	0.006	56.2	0.490	0.002
19	0.712	0.012	-0.061	0.024	0.201	0.009	31.6	0.617	0.002
20	0.697	0.012	-0.421	0.030	0.217	0.011	88.4	0.687	0.005

Note: Residuals are computed as $e_{ij}=u_{ij}-P(\theta_{1j})$. Item means are computed from the truncated dataset. SE=standard error, 3PL=three parameter logistic IRT model, RS=dichotomized raw score.

Table 14. 3PL IRT Item Characteristics: Last 34 Items of the NC OCSA, Fall 2005 Administration									
Item	Discrim-ination (a_i)	SE, a_i	Difficulty (b_i)	SE, b_i	Pseudo-guessing (c_i)	SE, c_i	Item χ^2	Mean Probability ($P(\theta_{Ij})$)	Mean Residual
21	0.735	0.006	0.707	0.007	<0.001	<0.001	142.0	0.344	0.003
22	0.583	0.006	-1.500	0.014	<0.001	<0.001	639.9	0.771	0.012
23	0.895	0.012	-1.283	0.028	0.098	0.015	28.5	0.816	0.013
24	0.389	0.013	0.942	0.044	0.093	0.013	118.8	0.422	0.004
25	0.641	0.005	-0.249	0.007	<0.001	<0.001	304.1	0.558	0.006
26	1.093	0.008	0.679	0.005	<0.001	<0.001	191.7	0.321	0.004
27	0.905	0.013	-0.227	0.015	0.141	0.007	113.2	0.631	0.003
28	0.409	0.005	-1.687	0.027	0.014	0.006	363.7	0.742	0.012
29	0.581	0.011	-1.062	0.052	0.121	0.020	18.1	0.739	0.010
30	0.550	0.018	1.399	0.022	0.220	0.007	18.5	0.415	0.002
31	0.728	0.006	0.623	0.007	<0.001	<0.001	108.6	0.363	-0.001
32	0.863	0.008	0.866	0.006	<0.001	<0.001	798.6	0.293	<0.001
33	1.159	0.021	0.941	0.009	0.232	0.003	192.3	0.423	-0.001
34	1.366	0.010	0.566	0.004	<0.001	<0.001	257.6	0.341	0.004
35	0.760	0.006	0.559	0.006	<0.001	<0.001	281.6	0.375	0.001
36	1.180	0.008	0.073	0.004	<0.001	<0.001	175.7	0.495	0.019
37	1.184	0.009	-0.783	0.005	<0.001	<0.001	154.5	0.728	0.018
38	0.532	0.011	-0.295	0.042	0.110	0.014	33.4	0.607	0.006
39	0.763	0.006	-0.162	0.006	<0.001	<0.001	290.5	0.547	0.006
40	0.628	0.006	-1.556	0.013	<0.001	<0.001	474.9	0.790	0.017
41	0.744	0.010	-0.131	0.017	0.078	0.007	50.1	0.574	0.004
42	0.616	0.015	0.992	0.019	0.192	0.006	43.9	0.437	0.002
43	1.164	0.009	-0.816	0.006	<0.001	<0.001	207.0	0.734	0.026
44	0.676	0.008	-0.630	0.021	0.033	0.009	566.8	0.653	0.008
45	0.936	0.007	-1.255	0.009	<0.001	<0.001	923.7	0.797	0.039
46	1.063	0.015	0.325	0.009	0.154	0.004	51.8	0.510	0.001
47	1.373	0.011	0.713	0.004	<0.001	<0.001	447.7	0.297	0.003
48	1.419	0.028	1.210	0.008	0.211	0.002	328.1	0.343	-0.001
49	0.534	0.014	-0.013	0.048	0.277	0.013	18.8	0.642	0.009
50	0.939	0.014	-0.002	0.014	0.191	0.006	54.2	0.606	0.006
51	0.830	0.008	1.524	0.011	<0.001	<0.001	70.2	0.169	-0.003
52	0.891	0.007	0.425	0.005	<0.001	<0.001	263.3	0.399	0.016
53	1.617	0.012	0.254	0.003	<0.001	<0.001	259.3	0.438	0.011
54	0.588	0.014	0.599	0.026	0.225	0.008	29.2	0.522	0.007

Note: Residuals are computed as $e_{ij}=u_{ij}- P(\theta_{Ij})$; item means from truncated data. SE= standard error. 3PL=three parameter logistic IRT model. RS=dichotomized raw score.

Table 15. Summary of IRT Item Characteristics, 3PL Item Response Probabilities, and Item Residuals: Fall 2005 Administration of the NC OCSA

Statistic	Mean	Standard Deviation	Minimum	Maximum
Discrimination (a_i)	0.805	0.271	0.258	1.617
Difficulty (b_i)	0.094	0.987	-2.555	2.094
Pseudo-guessing (c_i)	0.088	0.112	<0.001	0.444
Response Probability	0.524	0.181	0.109	0.816
Residual ($e_{ij}=u_{ij}-P(\theta_{1j})$)	0.007	0.008	-0.003	0.039

Notes: Item means are computed from the truncated dataset. All calculations are performed to full precision.

Table 16. Item Time Intensity (β_i), RT Standard Deviation, and Temporal Discrimination (α_i): Fall 2005 Administration of the NC OCSA

Item	Time Intensity (β_i)		Discrimination (α_i)	Item	Time Intensity (β_i)		Discrimination (α_i)	Item	Time Intensity (β_i)		Discrimination (α_i)
	Mean	SD			(α_i)	SD			Mean	SD	
1	32.97	26.30	0.038	19	36.87	39.50	0.025	37	81.52	73.85	0.014
2	99.44	86.17	0.012	20	42.91	33.35	0.030	38	34.72	35.66	0.028
3	111.37	93.13	0.011	21	104.06	78.53	0.013	39	46.44	40.81	0.025
4	97.02	90.74	0.011	22	39.41	48.30	0.021	40	60.26	47.01	0.021
5	44.26	31.16	0.032	23	35.92	35.17	0.028	41	32.51	31.72	0.032
6	44.72	34.80	0.029	24	68.71	53.38	0.019	42	35.41	34.81	0.029
7	57.63	40.43	0.025	25	88.80	71.83	0.014	43	81.32	60.16	0.017
8	47.54	42.77	0.023	26	175.59	136.80	0.007	44	56.40	46.90	0.021
9	112.19	69.64	0.014	27	38.47	57.95	0.017	45	66.83	61.63	0.016
10	48.38	45.66	0.022	28	31.82	39.62	0.025	46	40.61	43.34	0.023
11	51.68	37.91	0.026	29	42.08	36.86	0.027	47	139.94	99.16	0.010
12	40.97	29.21	0.034	30	31.32	32.28	0.031	48	54.91	53.95	0.019
13	27.86	26.21	0.038	31	65.60	59.85	0.017	49	23.73	34.64	0.029
14	97.05	83.66	0.012	32	218.83	132.24	0.008	50	23.56	26.35	0.038
15	211.61	135.99	0.007	33	49.41	67.35	0.015	51	68.81	54.91	0.018
16	168.37	121.39	0.008	34	228.23	162.99	0.006	52	113.55	92.47	0.011
17	81.76	62.33	0.016	35	117.61	93.83	0.011	53	67.34	61.59	0.016
18	43.25	40.22	0.025	36	127.86	113.87	0.009	54	48.21	45.70	0.022

Note: Item descriptive statistics are computed from the truncated dataset. All calculations are performed to full precision. RT=item response time (sec), SD=standard deviation.

Table 17. Summary of Examinee Characteristics: Examinee EAP θ_{Ij} Estimates from Fall 2005 Administration of the NC OCSA

Statistic	Mean	Standard Deviation	Minimum	Maximum
EAP θ_{Ij} estimates, all examinees	-0.033	1.130	-4.00	+4.00
SE, EAP θ_{Ij} estimates of all examinees	0.392	0.941	0.236	9.00
EAP θ_{Ij} estimates, examinees without maximal SE.	0.013	1.051	-3.995	3.915
SE, EAP θ_{Ij} estimates, examinees without maximal SE	0.291	0.107	0.236	1.318

Note: 1210 students have EAP θ_{Ij} estimates with maximal standard errors (9.00). 1203 have θ_{Ij} estimates of -4.00; the number of correct responses for these students ranged from 1 to 12. 7 students have θ_{Ij} estimates of +4.00; these students all responded correctly to 53 or 54 items. EAP = Expected a posteriori, SE=standard error. Descriptive statistics from the truncated dataset (*N*=103,751; number of examinees with non-maximal SE's: 102,541.

Table 18. Descriptive Q_3 Statistics (Yen, 1984): Fall 2005 Administration of the NC OCSA

	Number of Item Pairs	Mean	Standard Deviation	Minimum	Maximum
Unadjusted Q_3	1431	-0.0153	0.0336	-0.1185	0.3677
Adjusted Q_3	1431	0.0036	0.0336	-0.0997	0.3866

Note: Q_3 statistics are computed for all unique item pairs (k_1, k_2), where k is an item identifier and $k_1 \neq k_2$. Unadjusted Q_3 values are computed from the truncated dataset; a correction factor ($-1/(n-1)$) is applied to the unadjusted Q_3 values, where n is the total number of test items (Yen, 1984). All calculations are performed to full precision.

Table 19. Q_3 Percentiles: Fall 2005 Administration of the NC OCSA

Percentile	Unadjusted Q_3	Adjusted Q_3
25	-0.029	-0.010
50	-0.016	0.003
75	-0.004	0.015
90	0.009	0.028
99	0.095	0.114

Note: Q_3 statistics are computed for all unique item pairs (k_1, k_2, $k_1 \neq k_2$). Unadjusted Q_3 values are computed from the truncated dataset; a correction factor (-1/(n-1) is applied to the unadjusted Q_3 values, where n is the total number of test items (Yen, 1984). All calculations are performed to full precision.

Table 20. Semipartial Correlation Coefficients ($r_{e_{ij}t_{ij}}$): First 20 Items from Fall 2005 Administration of the NC OCSA

Item	Semipartial Correlation ($r_{e_{ij}t_{ij}}$)	Squared Semipartial Correlation	Percent of Variance
1	-0.098	0.010	0.95
2	-0.180	0.033	3.25
3	-0.013	<0.001	0.02
4	-0.134	0.018	1.81
5	-0.075	0.006	0.57
6	-0.117	0.014	1.37
7	-0.075	0.006	0.57
8	-0.143	0.020	2.03
9	-0.230	0.053	5.28
10	-0.084	0.007	0.71
11	-0.058	0.003	0.34
12	-0.057	0.003	0.33
13	0.031	0.001	0.10
14	0.035	0.001	0.12
15	0.064	0.004	0.41
16	-0.152	0.023	2.30
17	-0.079	0.006	0.63
18	-0.054	0.003	0.29
19	-0.076	0.006	0.57
20	-0.007	<0.001	0.01

<u>Note:</u> Rounding is performed after calculating to full precision. RT = item response time.

Table 21. Semipartial Correlation Coefficients ($r_{e_{ij}t_{ij}}$): Last 34 Items from Fall 2005 Administration of the NC OCSA

Item	Semipartial Correlation ($r_{e_{ij}t_{ij}}$)	Squared Semipartial Correlation	Percent of Variance
21	0.018	<0.001	0.03
22	-0.053	0.003	0.28
23	-0.089	0.008	0.80
24	-0.018	<0.001	0.03
25	-0.133	0.018	1.78
26	-0.085	0.007	0.72
27	-0.046	0.002	0.21
28	-0.044	0.002	0.19
29	-0.072	0.005	0.52
30	-0.027	0.001	0.07
31	-0.029	0.001	0.09
32	0.112	0.013	1.25
33	-0.037	0.001	0.14
34	-0.068	0.005	0.47
35	-0.098	0.010	0.96
36	-0.134	0.018	1.80
37	-0.165	0.027	2.74
38	-0.076	0.006	0.58
39	-0.044	0.002	0.19
40	-0.077	0.006	0.59
41	-0.066	0.004	0.43
42	-0.007	<0.001	0.00
43	-0.103	0.011	1.05
44	-0.028	0.001	0.08
45	-0.046	0.002	0.21
46	-0.046	0.002	0.22
47	-0.119	0.014	1.42
48	-0.016	<0.001	0.03
49	-0.034	0.001	0.11
50	-0.041	0.002	0.17
51	-0.036	0.001	0.13
52	0.030	0.001	0.09
53	-0.117	0.014	1.36
54	0.010	<0.001	0.01

Note: Rounding is performed after calculating to full precision. RT = response time.

Table 22. Pearson Correlation Coefficients between Item Residuals and Semipartial Correlation Coefficients ($r_{e_{ij}t_{ij}}$): Items from Fall 2005 Administration of the NC OCSA

	IRT Parameter		
	Discrimination *(a_i)*	Item Difficulty *(b_i)*	Pseudo-guessing *(c_i)*
Item Residual	0.078	-0.686	-0.366
Semipartial Correlation ($r_{e_{ij}t_{ij}}$)	-0.114	0.498	0.151
Squared Semipartial	-0.094	-0.366	-0.319

Note: Rounding is performed after calculating to full precision. RT = item response time. Item Residuals: $e_{ij}=u_{ij}-P(\theta_{Ij})$.

Figure 1. Relationships Between Examinee Ability Level (θ_{1j}) and Observed Responses to Three Hypothetical Test Items

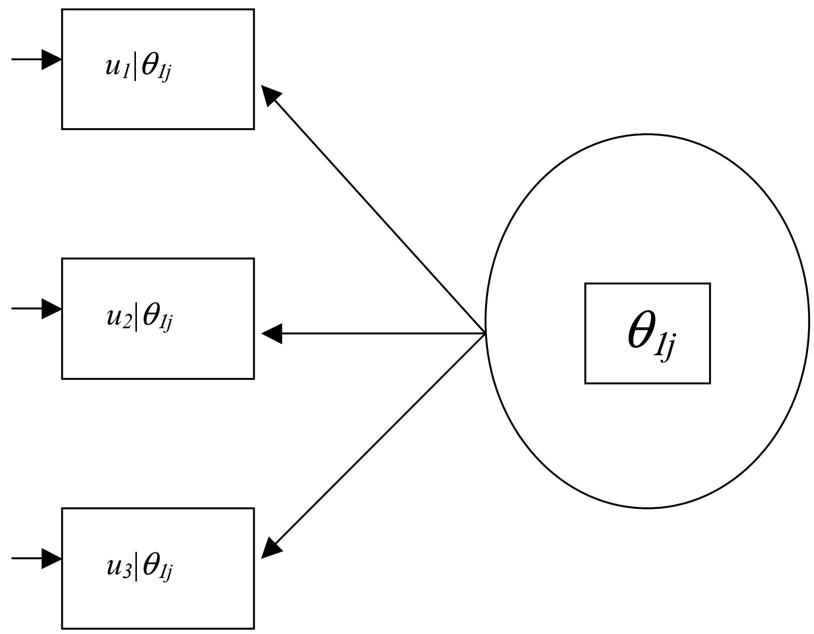


Figure 2. Postulated Relationships between Examinee Latent (θ and τ) and Observed Variables (Item Responses [u_i] and Item Response Times [t_i])

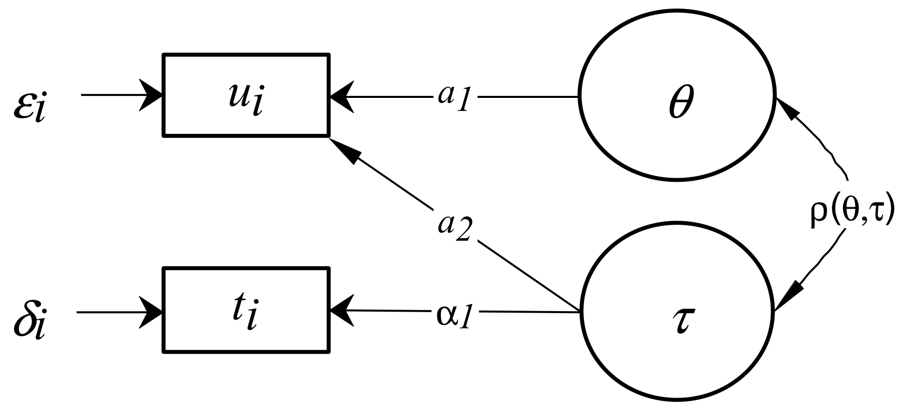
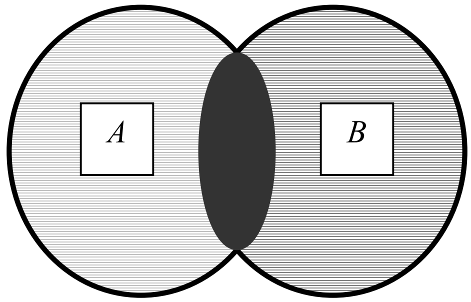


Figure 3. Relationship between Variance in an Item Residual (*A*) Explained by Item Response Time (*B*)



Circle *A*: Variance of Item Residual
Circle *B*: Variance of Item Response Time
Intersection: Portion of *A* explained by *B* through a semi-partial correlation

Figure 4. Relationships between an Examinee Pacing Parameter (τ) and the Indirect $\theta_I \tau$ Effect at Several Item Discriminations: Mean τ, Short Test Length (20 items)

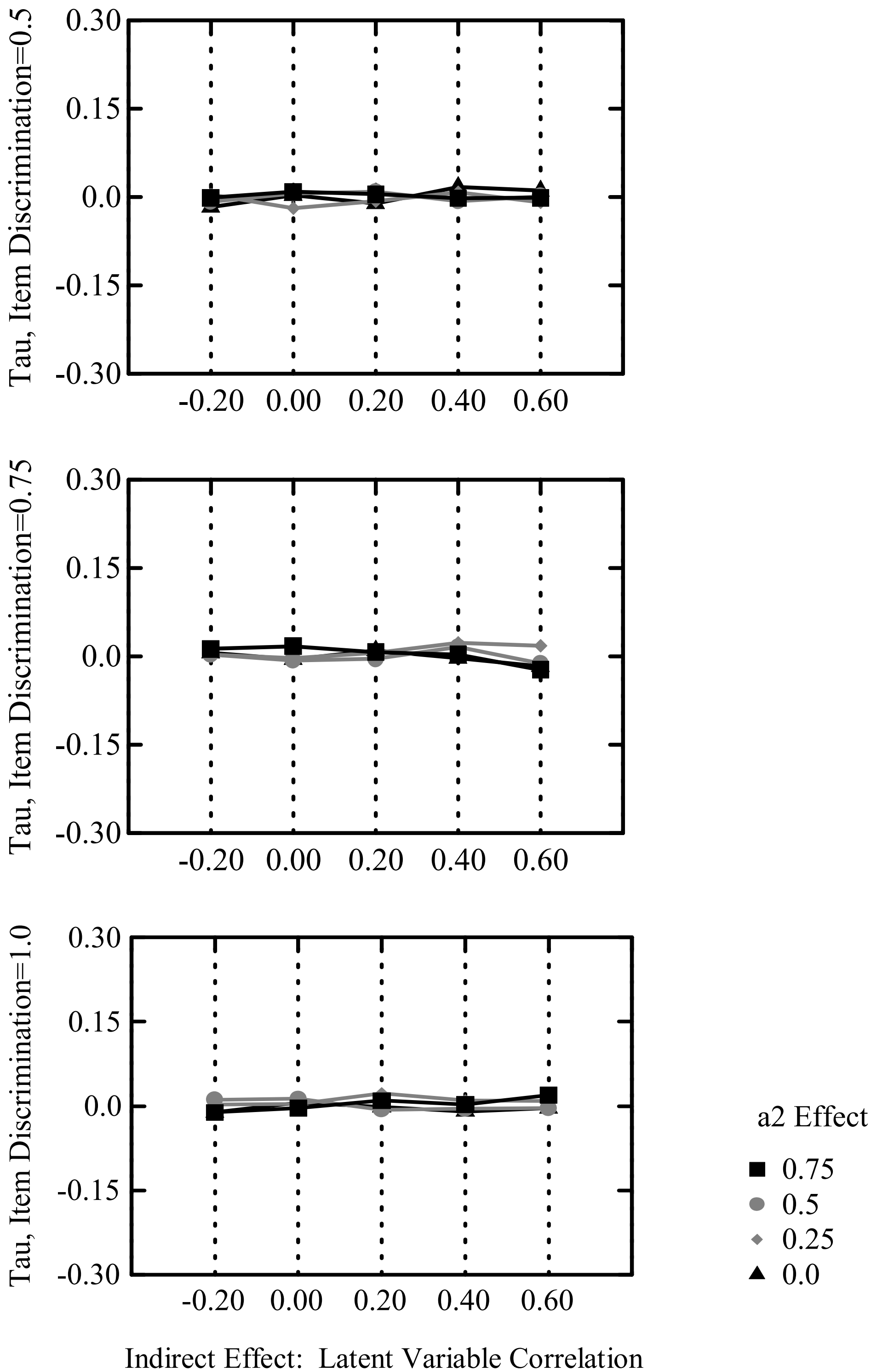


Figure 5. Relationships between an Examinee Pacing Parameter (τ) and the Indirect $\theta_I \tau$ Effect at Several Item Discriminations: Mean τ, Intermediate Test Length (30 items)

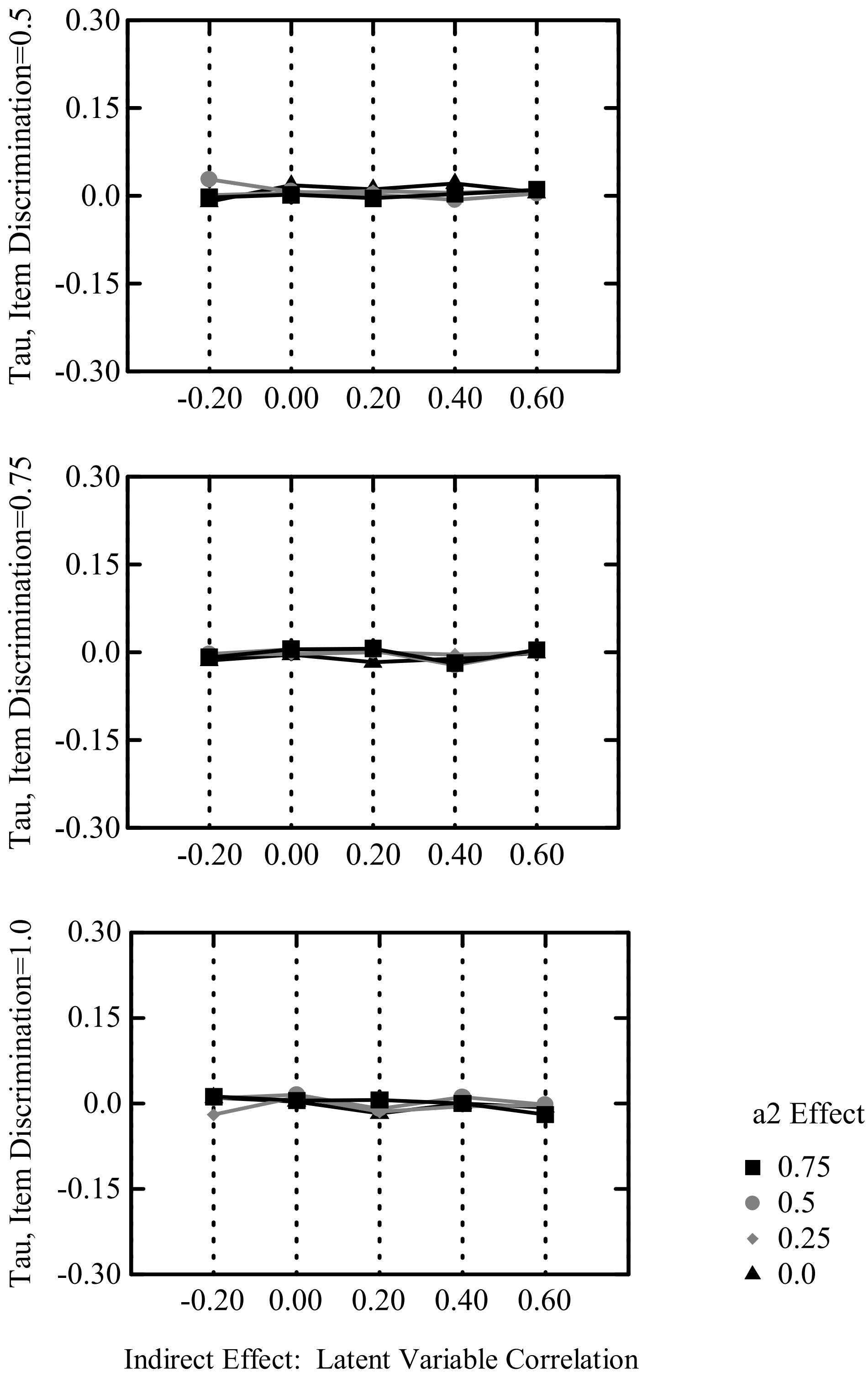


Figure 6. Relationships between an Examinee Pacing Parameter (τ) and the Indirect $\theta_I\tau$ Effect at Several Item Discriminations: Mean τ, Long Test Length (60 items)

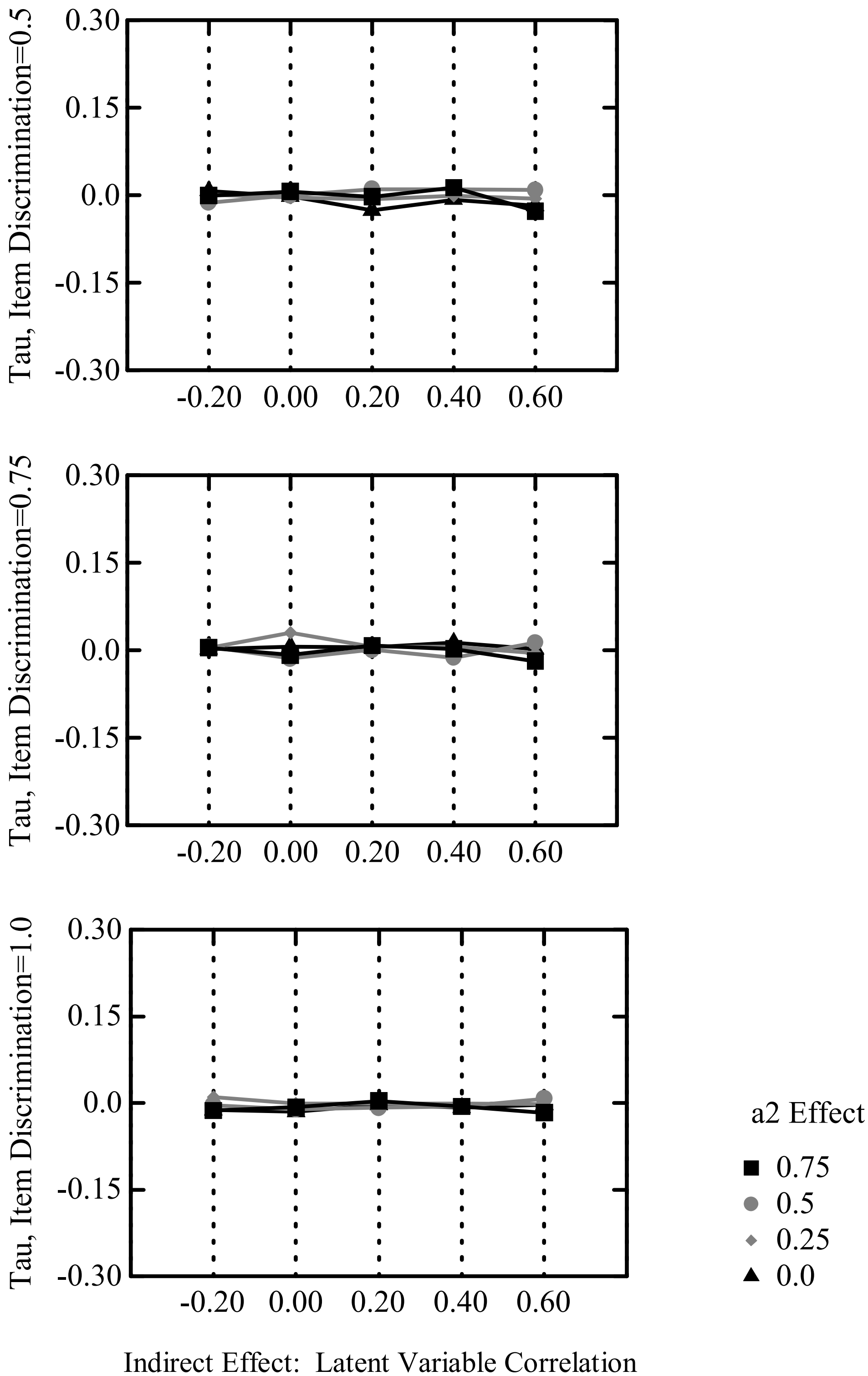


Figure 7. Mean Pearson Correlations (+ SEM) between EAP θ_1 and τ as a Function of 4 Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

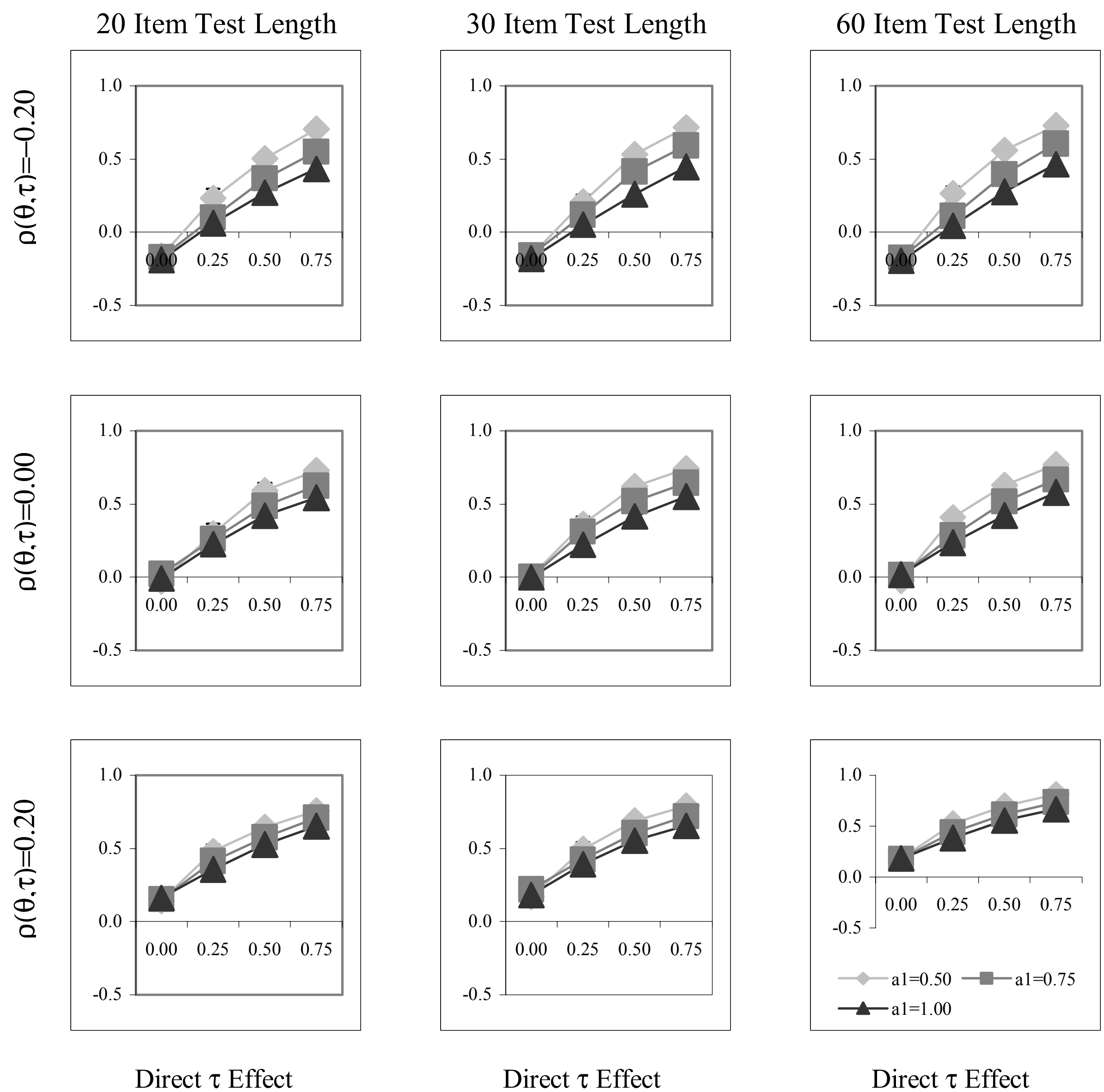


Figure 7 (continued). Mean Pearson Correlations (+ SEM) between EAP θ_1 and τ as a Function of 4 Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

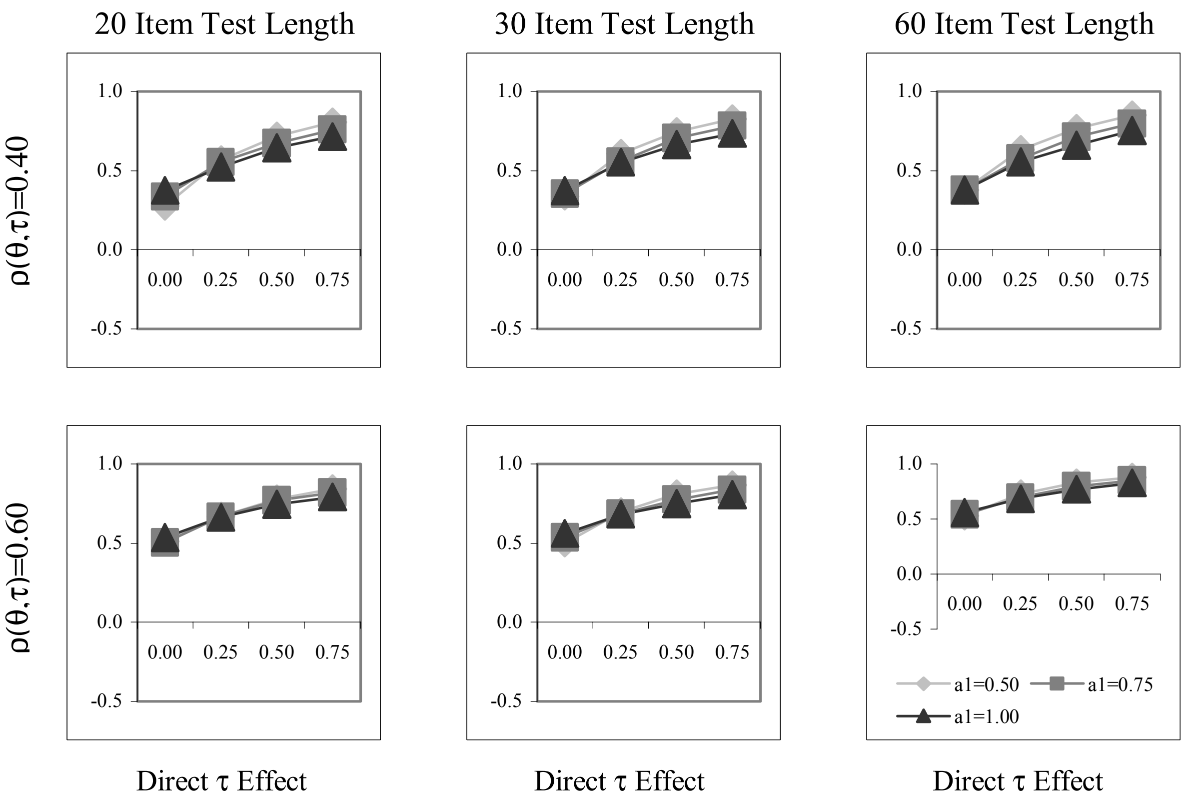


Figure 8. Mean Semi-Partial Correlation Coefficients (+ Average Standard Deviation) as a Function of Four Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

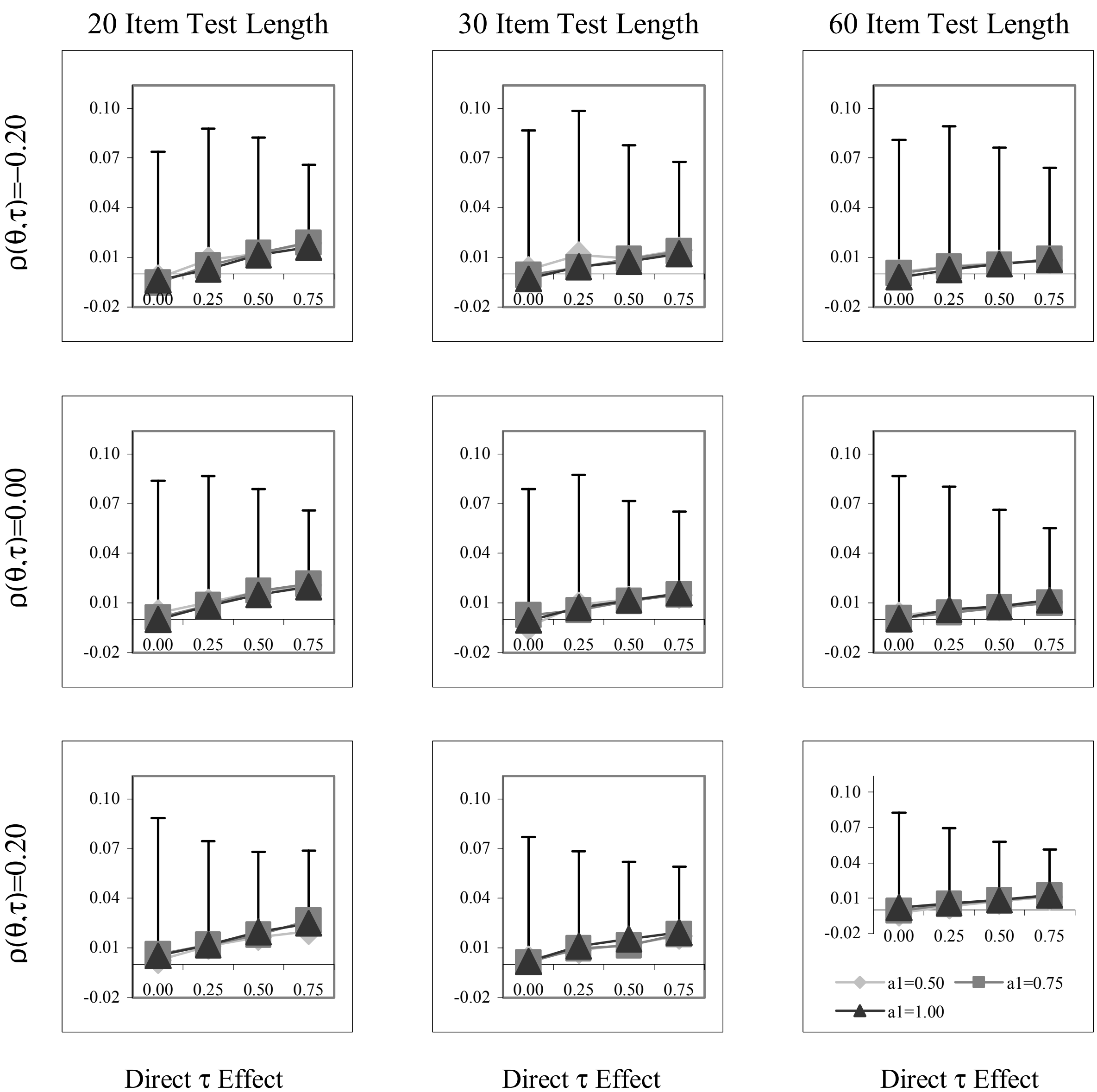


Figure 8 (continued). Mean Semi-Partial Correlation Coefficients (+ Average Standard Deviation) as a Function of Four Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

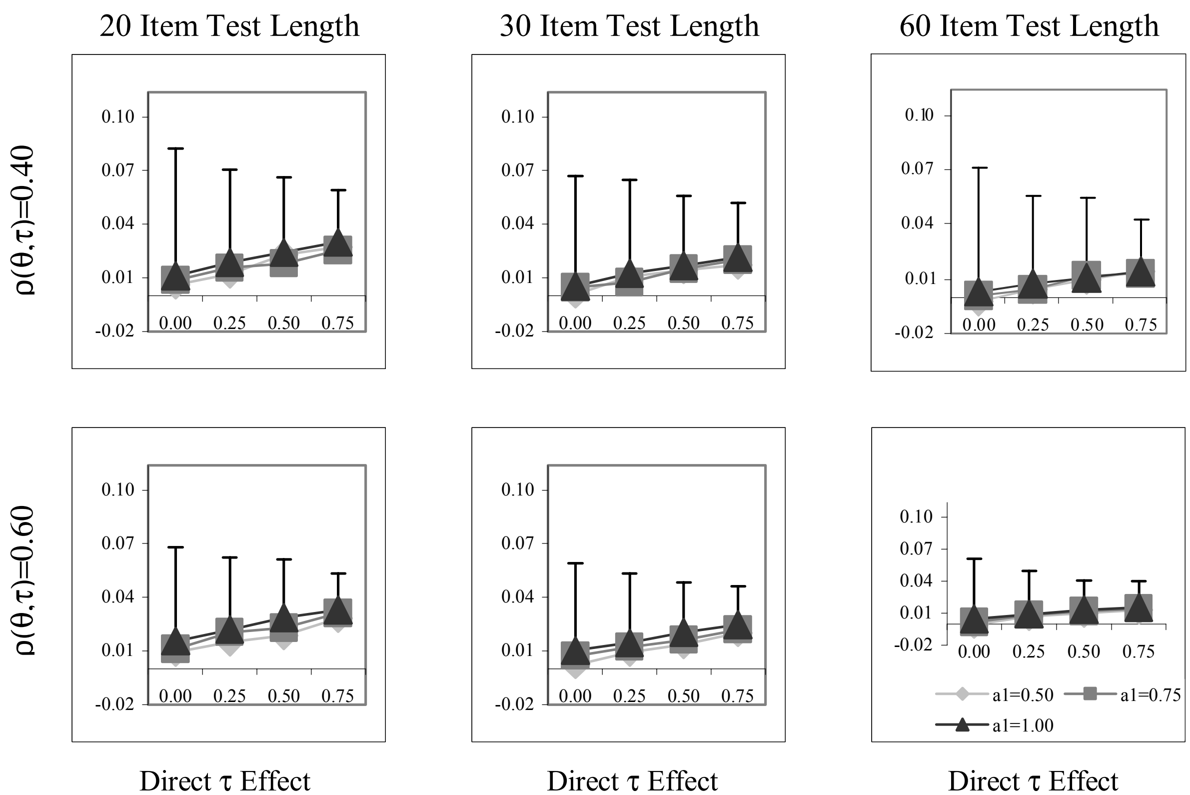


Figure 9. Mean Q_3 (+ Average Standard Deviation) as a Function of 4 Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

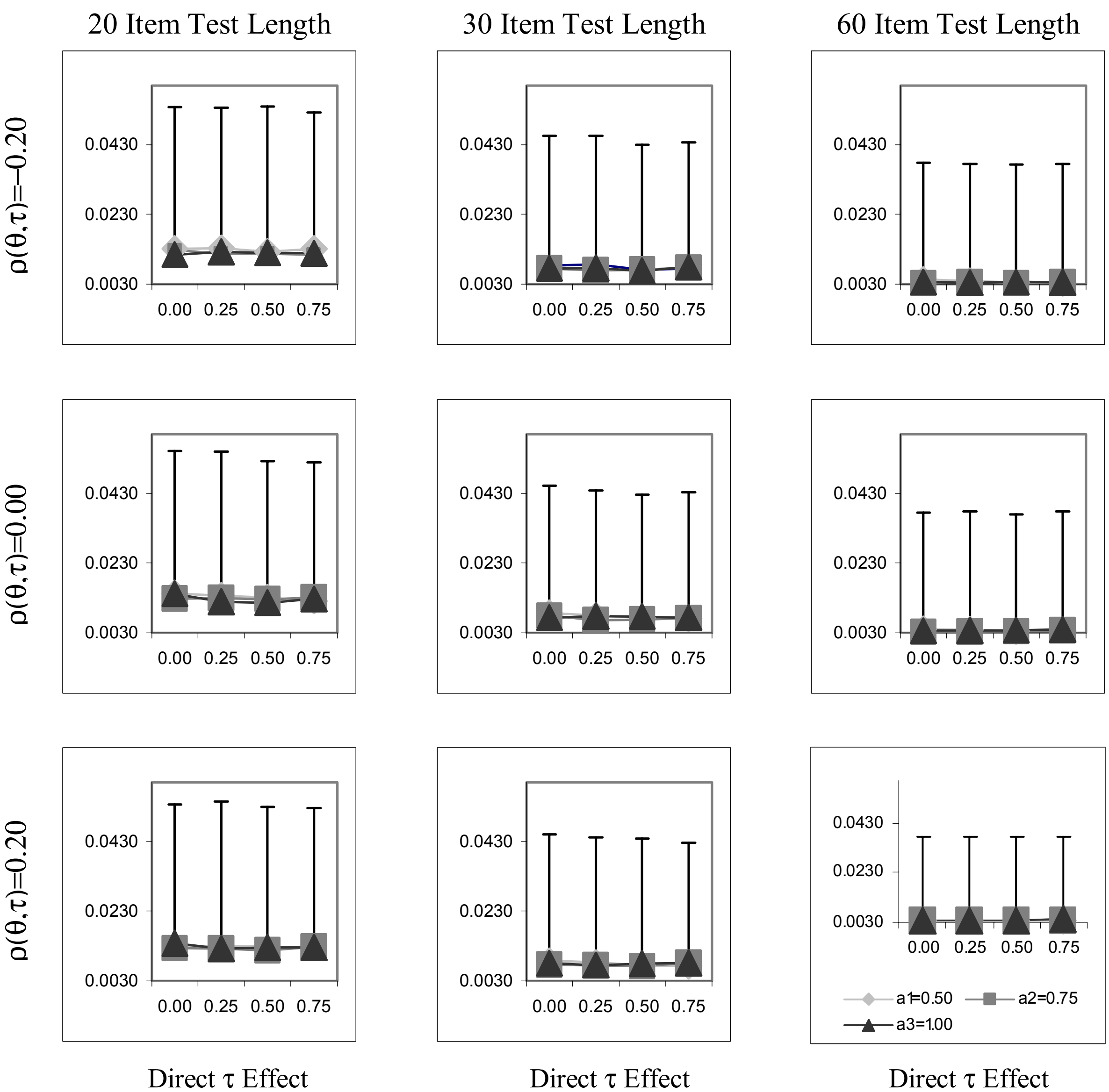


Figure 9 (continued). Mean Q_3 (+ Average Standard Deviation) as a Function of 4 Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

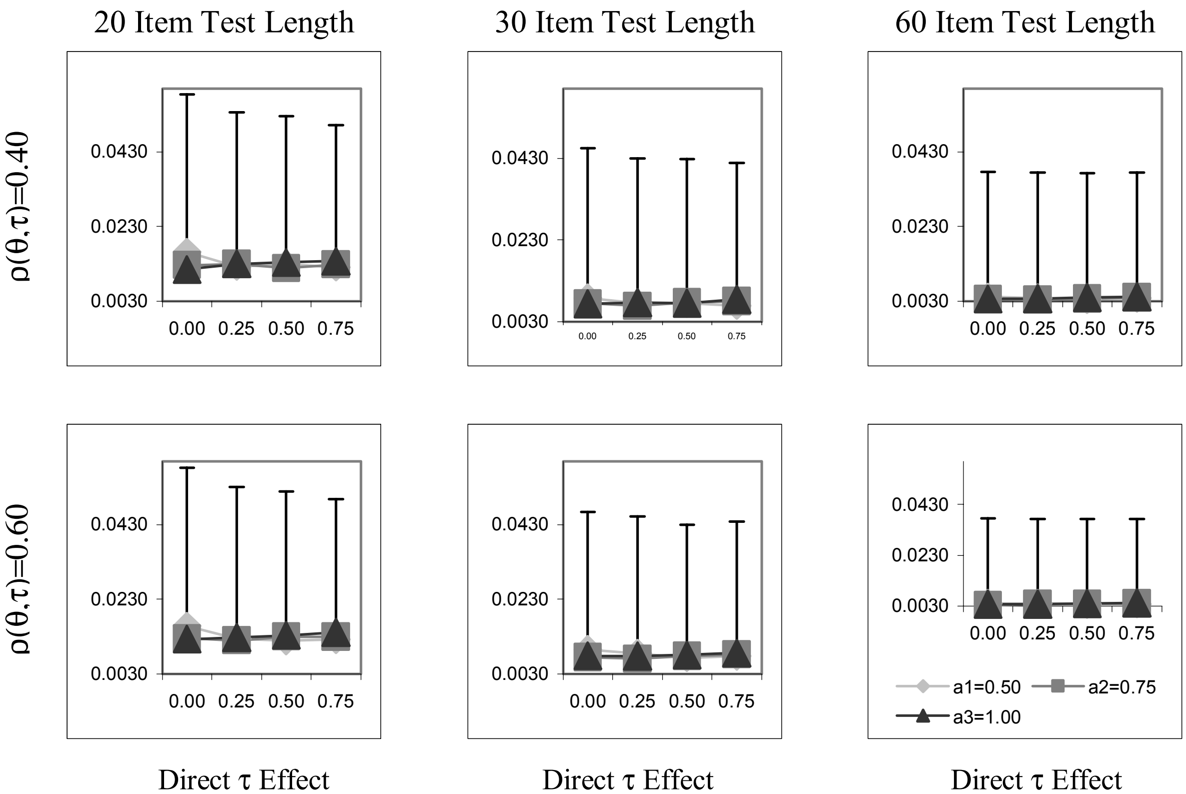


Figure 10. Mean RMSE (+ Standard Error of Measure) as a Function of 4 Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

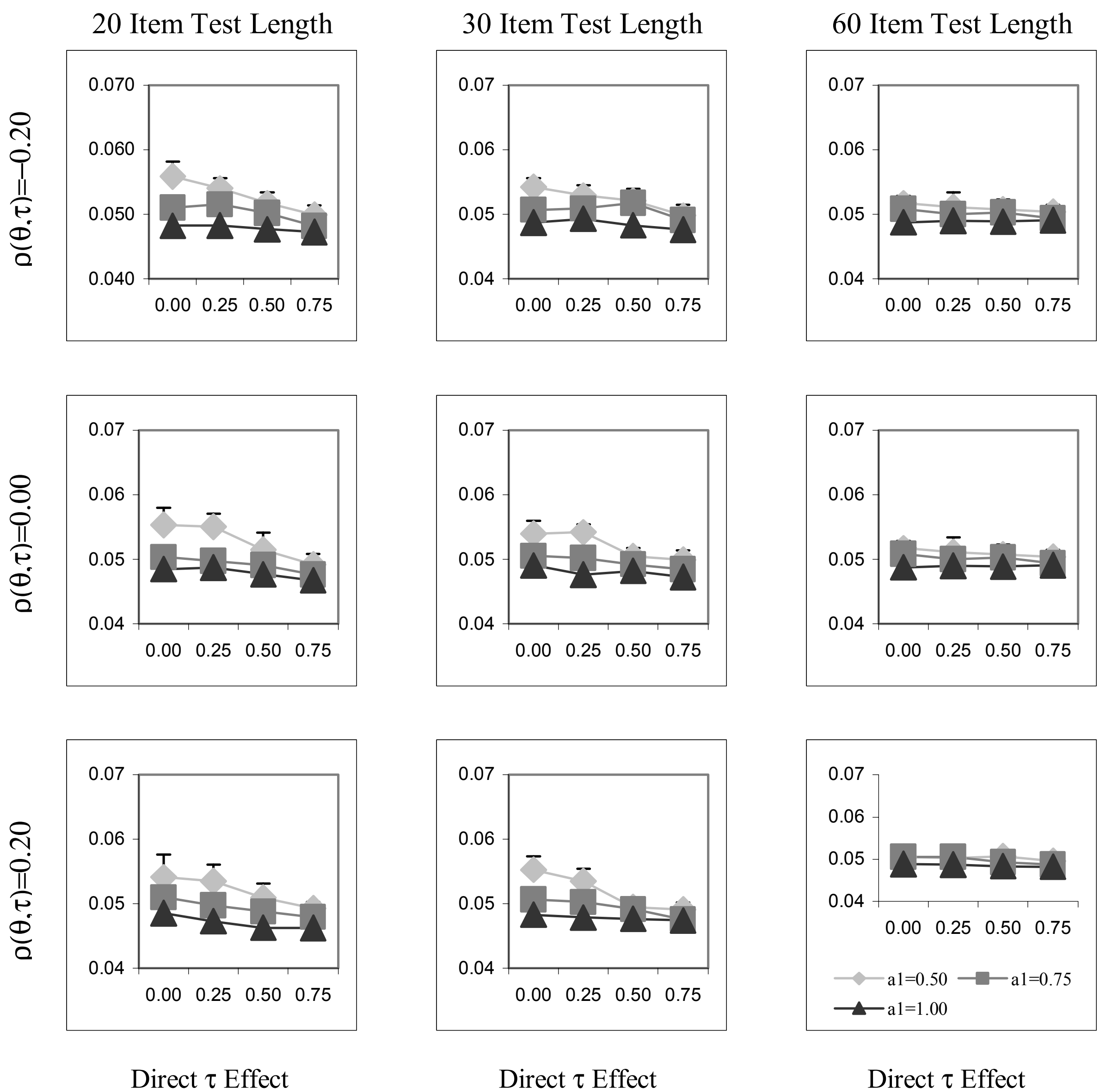


Figure 10 (continued). Mean RMSE (+ Standard Error of Measure) as a Function of 4 Factors: Item Discrimination (a_1), Direct τ Influence (a_2), Indirect τ Influence, and Test Length

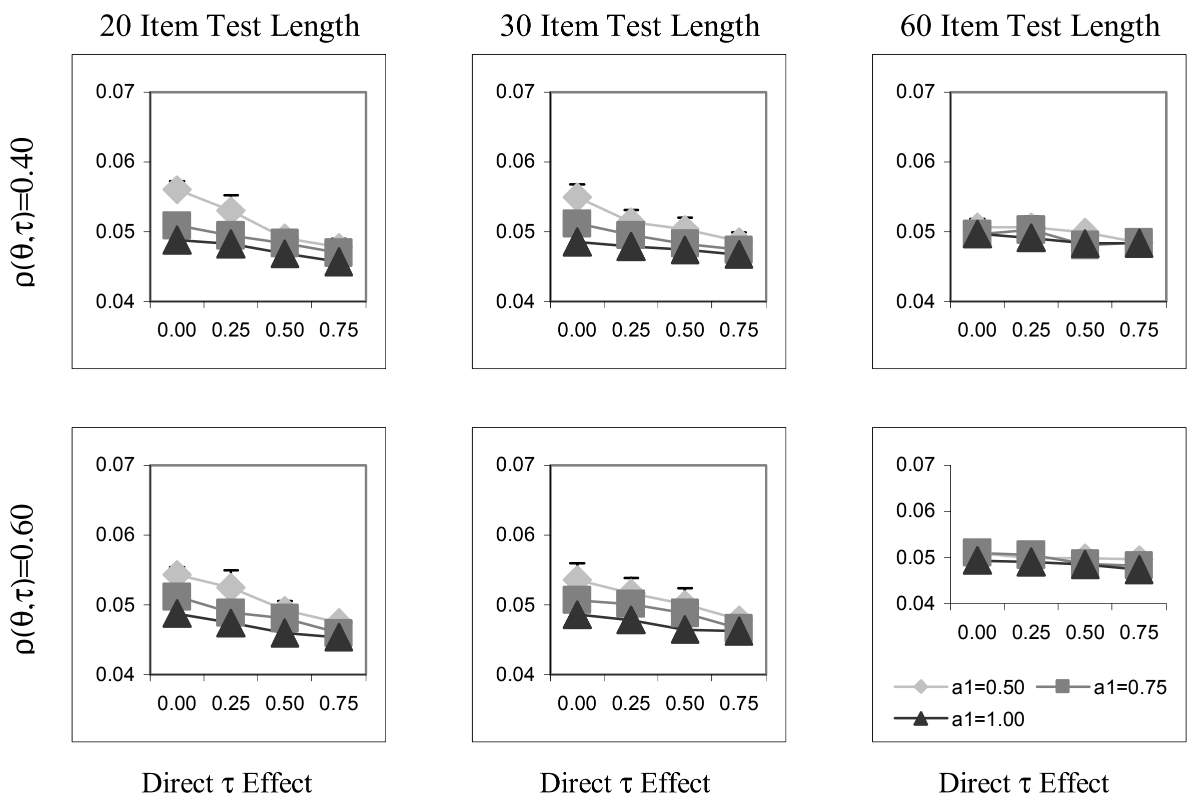


Figure 11. Complete Dataset (*N*=105917), Total Test Response Times: Fall 2005 Administration of Online Computer Skills Assessment

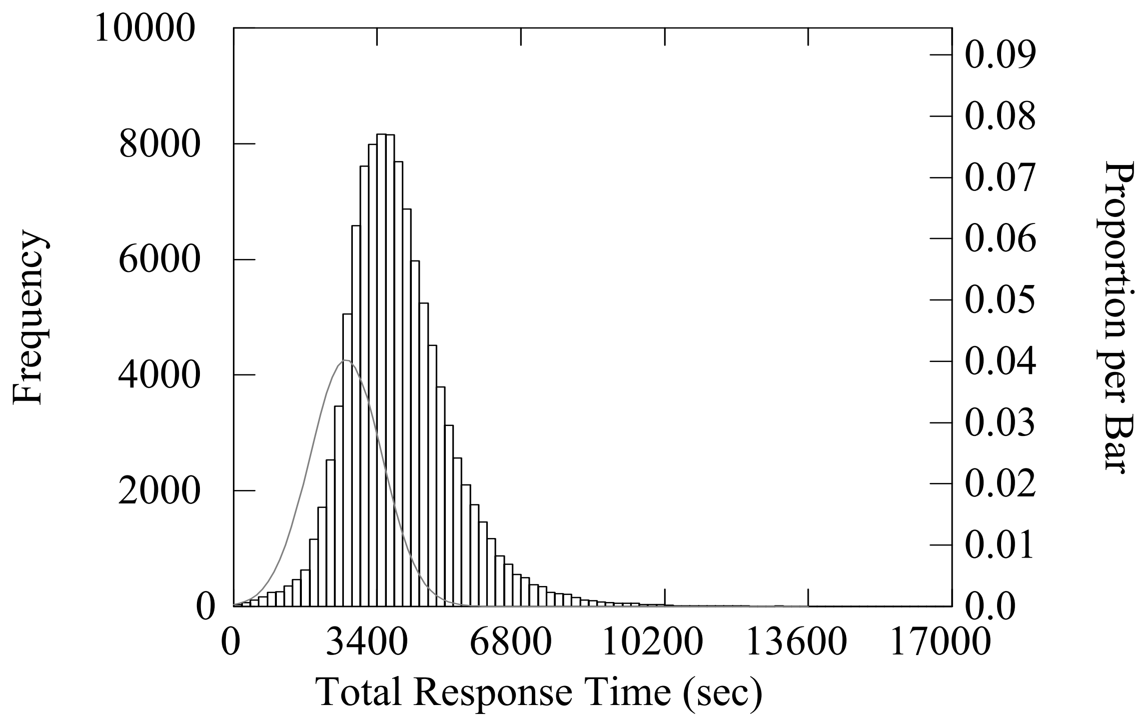


Figure 12. Time-Truncated Dataset (*N*=103751), Total Test Response Times: Fall 2005 Administration of Online Computer Skills Assessment

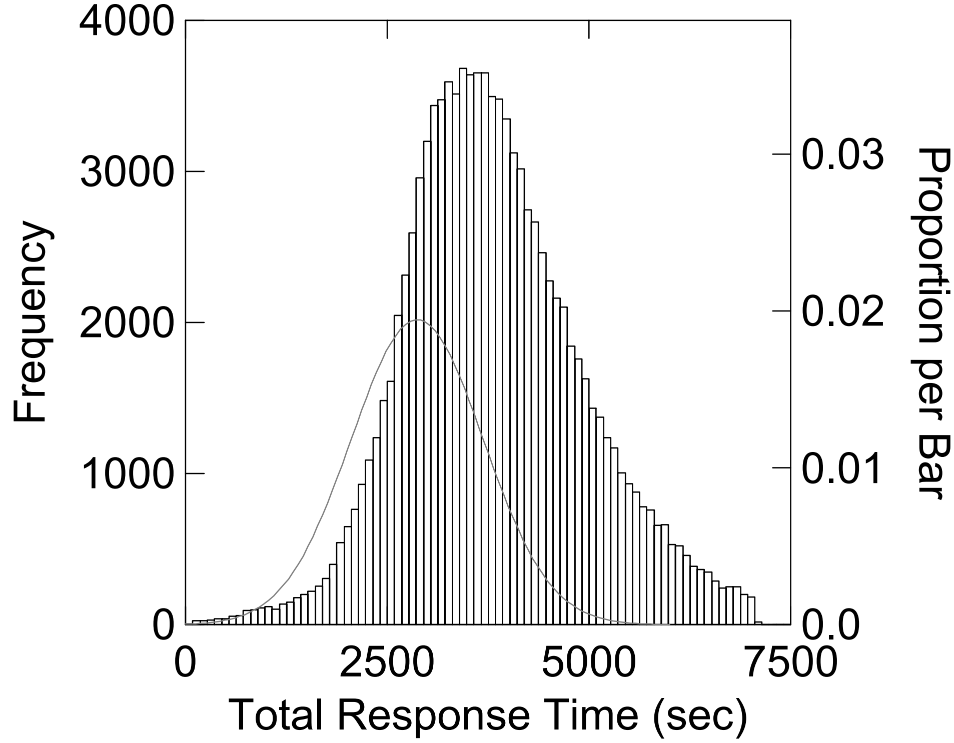


Figure 13. Time Truncated Dataset (N=103751), Total Test Score: Fall 2005 Administration of Online Computer Skills Assessment

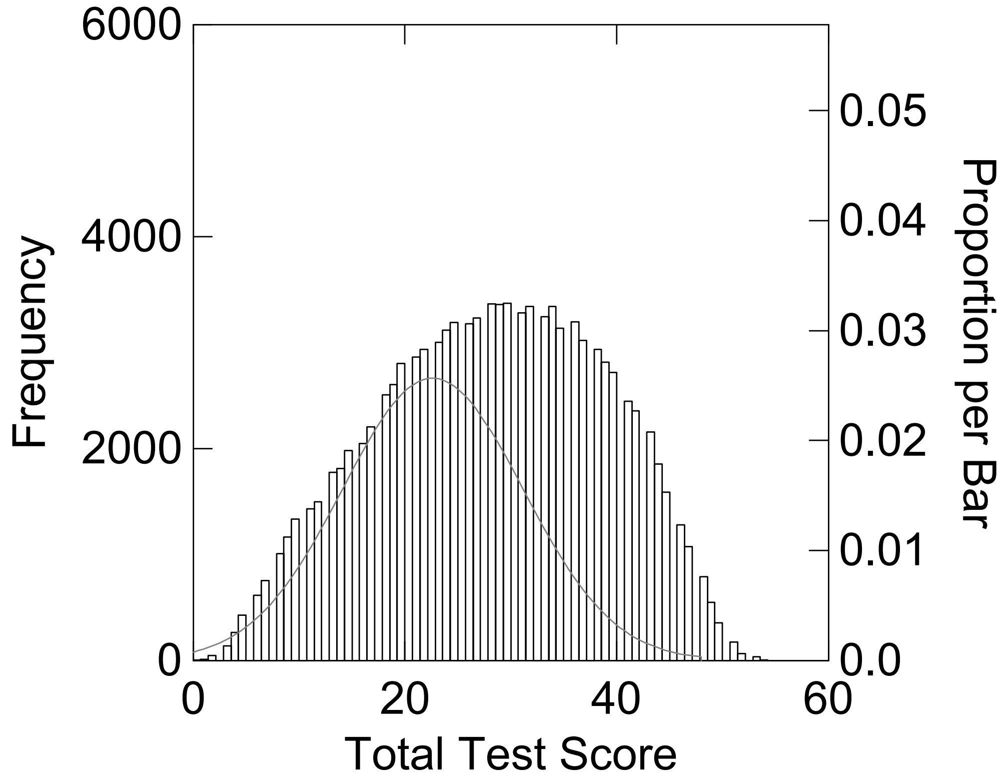


Figure 14. Time Truncated Dataset (N=103751), Response Times on Item 9: Fall 2005 Administration of Online Computer Skills Assessment

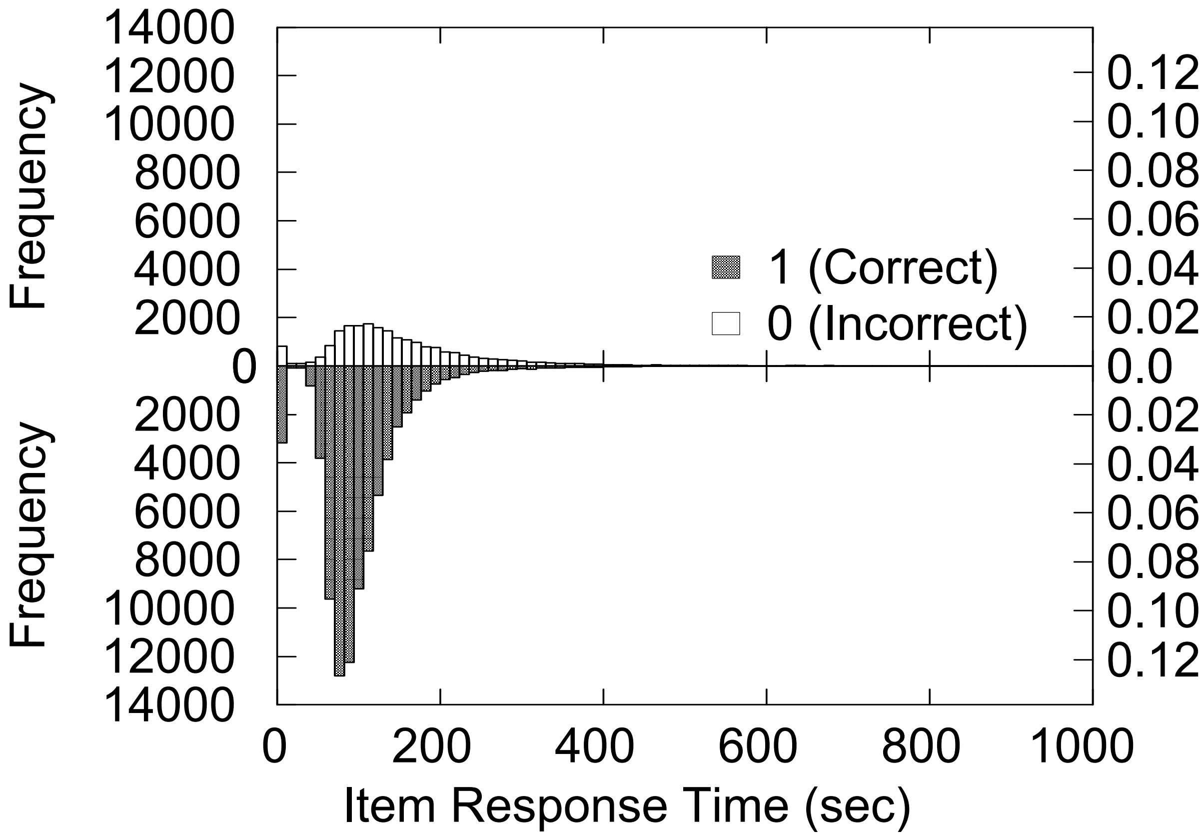


Figure 15. Time Truncated Dataset (N=103751), Response Times on Item 7: Fall 2005 Administration of Online Computer Skills Assessment

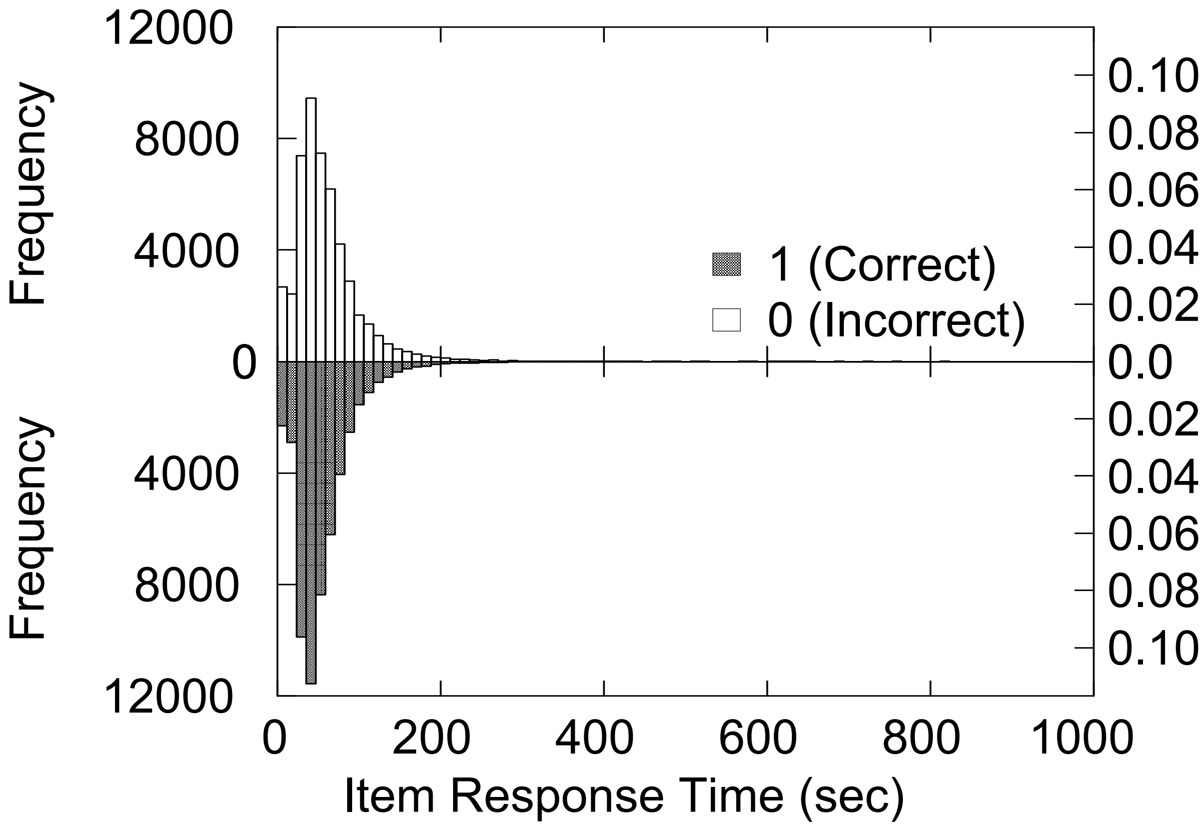


Figure 16. Time Truncated Dataset (N=103751), Response Times on Item 14:
Fall 2005 Administration of Online Computer Skills Assessment

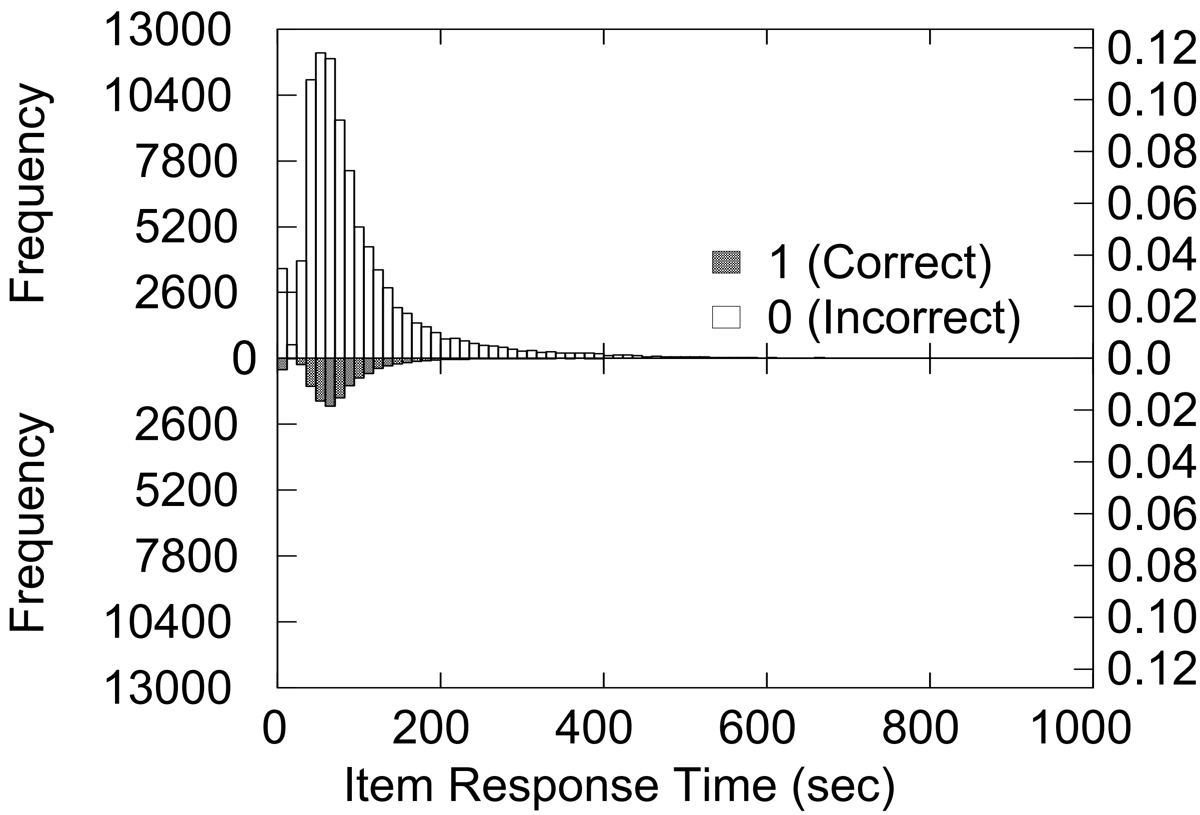


Figure 17. Time Truncated Dataset (N=103751), Response Times on Item 23: Fall 2005 Administration of Online Computer Skills Assessment

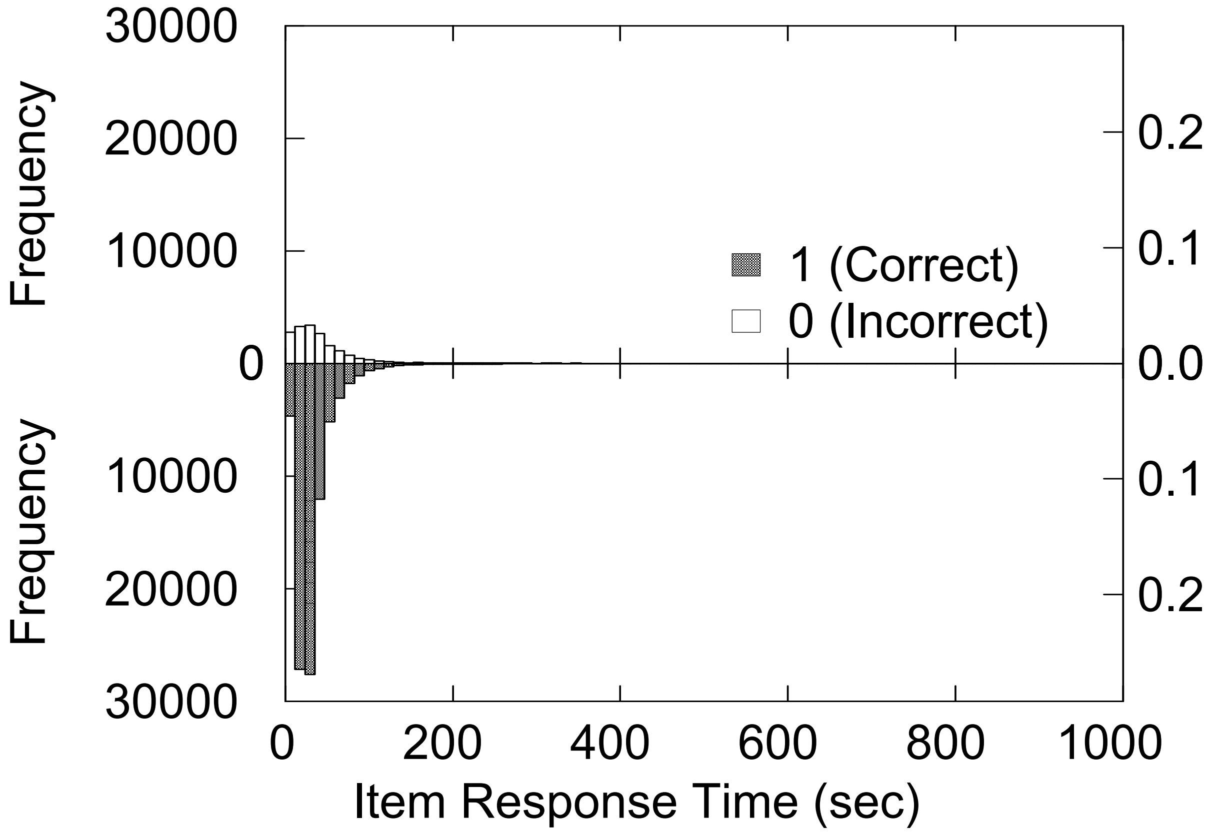


Figure 18. Time Truncated Dataset (*N*=103751), Response Times on Item 45: Fall 2005 Administration of Online Computer Skills Assessment

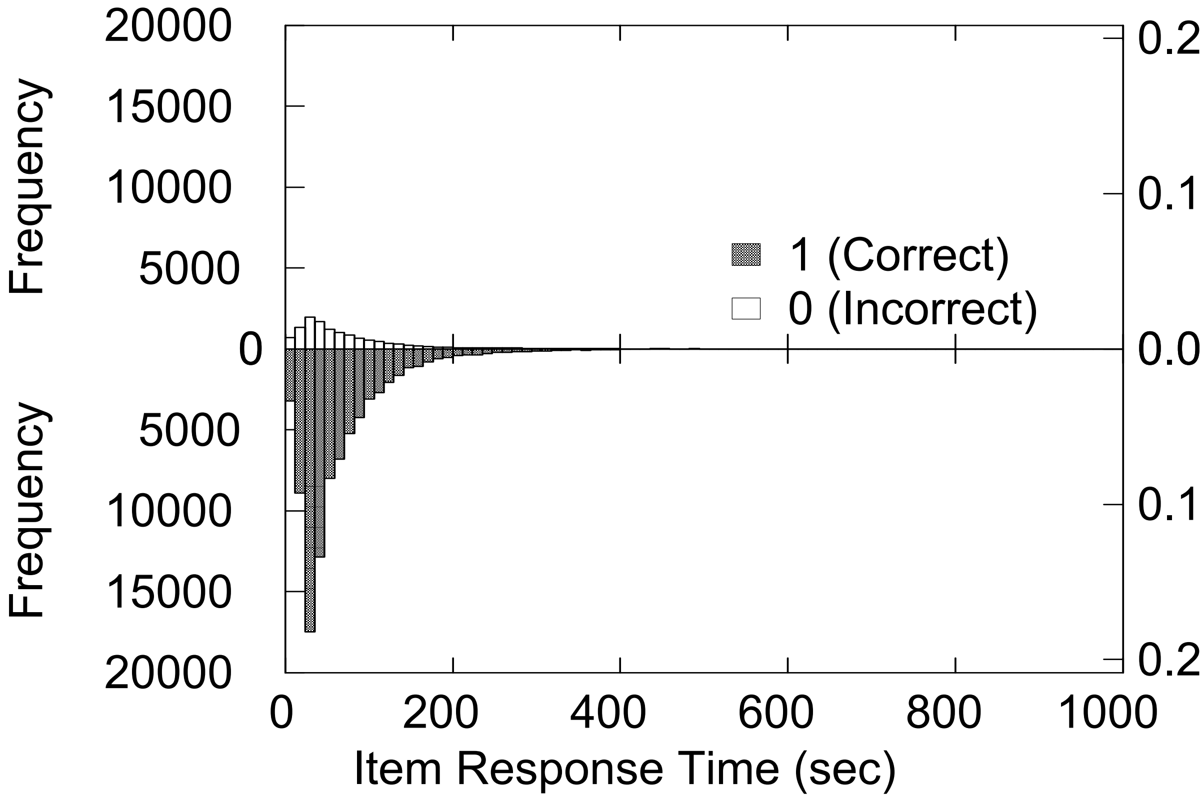


Figure 19. Time Truncated Dataset (N=103751), Plot of Mean Item Residuals and Mean Item 3PL Response Probabilities:
Fall 2005 Administration of Online Computer Skills Assessment

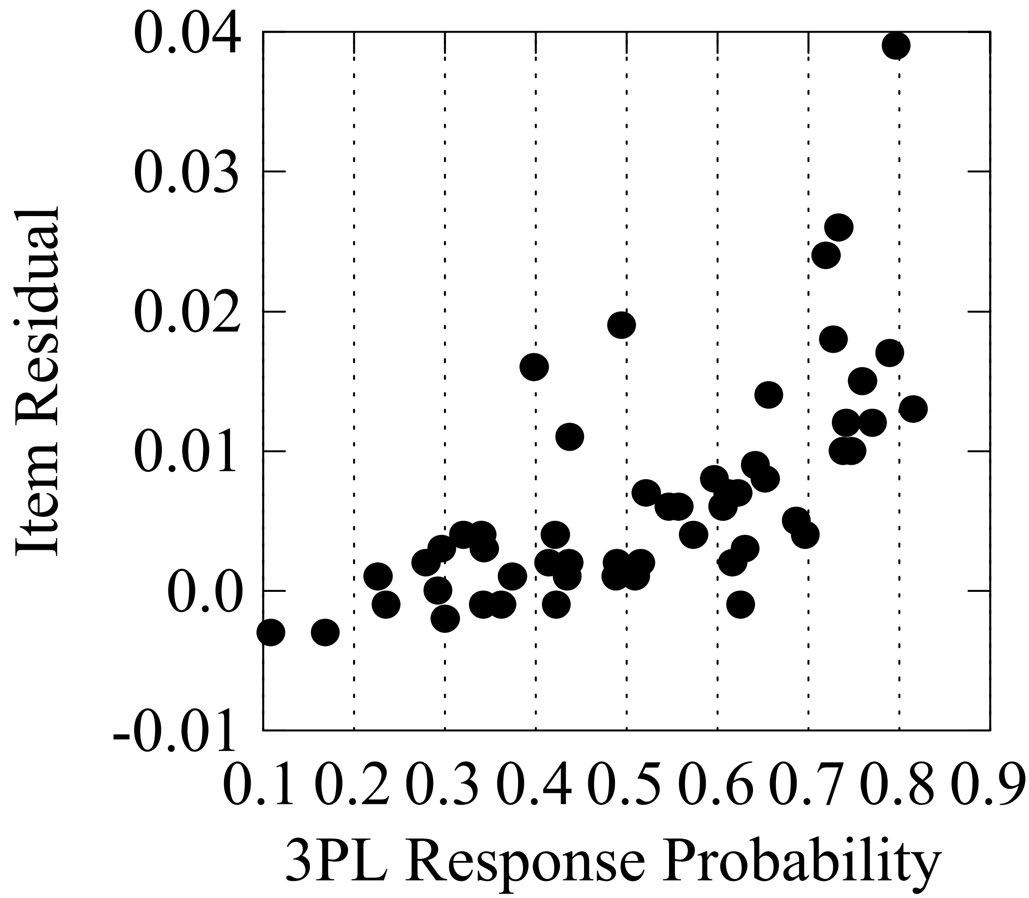
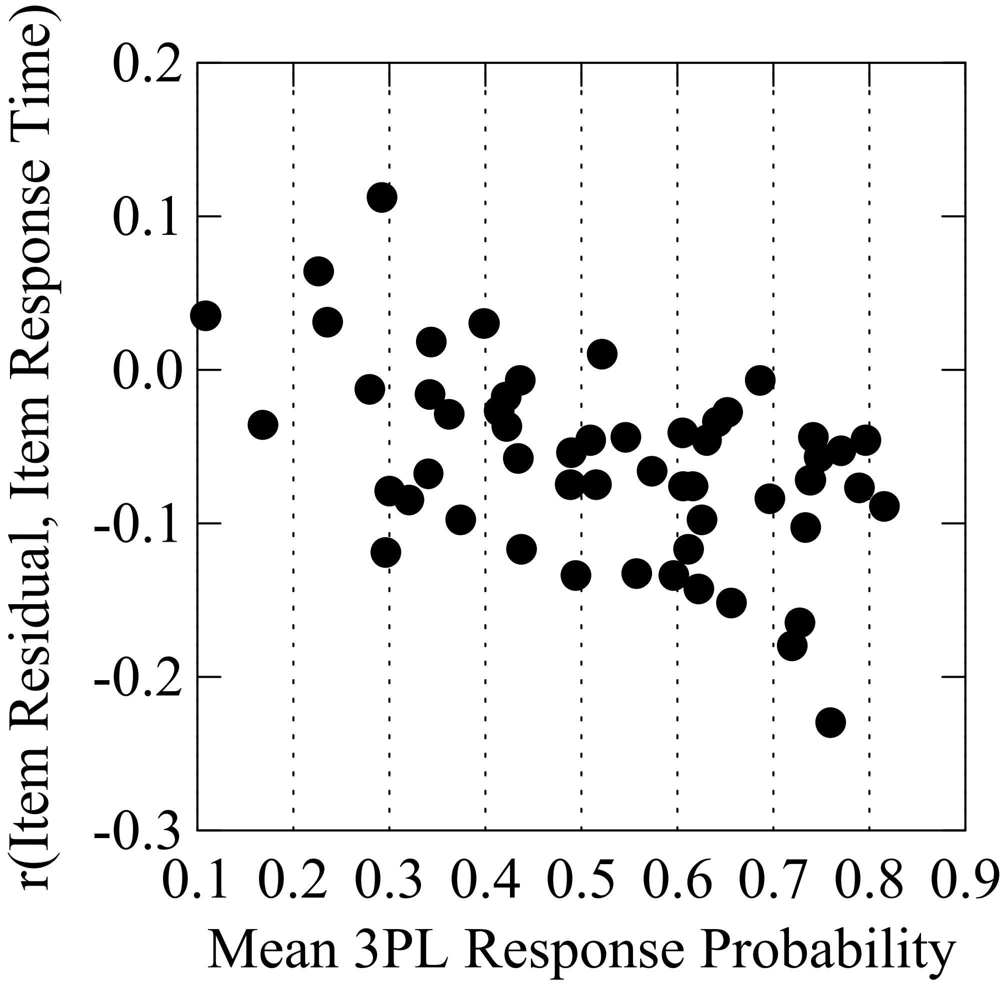


Figure 20. Time Truncated Dataset (N=103751), Plot of Mean Item Semipartial Correlations ($r_{e_{ij}t_{ij}}$) and Mean 3PL Response Probabilities: Fall 2005 Administration of Online Computer Skills Assessment



Appendix B. Sample Software Programs

Table B1. Sample BILOG-MG Program Used in Simulation

```
20_a11_a21_r1_01 v4 - IRT ANALYSIS OF A 20 ITEM TEST
Treatment Condition: 20_a11_a21_r1 Rep: 01 ITEM PARMS CALCULATED.
>GLOBAL DFName = 'T20_a11_a21_r1_01.xdt',
       NPArm = 3,
       SAVe;
>SAVE PARm =  'T20_a11_a21_r1_01.par',
     SCOre =  'T20_a11_a21_r1_01.sco';
>LENGTH NITems = (20);
>INPUT NTOtal = 20,
      NALt = 2,
      NIDchar = 5;
>ITEMS INUmber = (1(1)20), INAmes = (SIM001(1)SIM20);
>TEST1 TNAme = 2011101;
(5A1,7X,20A1)
>CALIB CYCles = 125,
      NEWton = 75,
      CRIt = 0.001,
      PLOt = 1.0000,
      ACCel = 1.0000,
      TPRior;
>SCORE RSCtype = 0,
      METhod=1,
      MOMents=1,
      INFo = 2;
```

Table B2. Sample SAS Program Generating BILOG-MG Scripts

```
/* program: bilogscript.sas
   researcher: john klaric*/
data tp; input x; datalines; 0; run;
%macro do_it(TL,a1,a2,r,rep,nitem);
%let pthnam2=C:\Program Files\bilogmg\Simulations\2009_q;
data bilog; set tp;
%let dfname=T&TL._a1&a1._a2&a2._r&r._&rep.;
file "&pthnam2.\T&TL._a1&a1._a2&a2._r&r._&rep..blm" ;
if _n_=1 then put
"&TL._a1&a1._a2&a2._r&r._&rep. v4 - IRT ANALYSIS OF A &nitem. ITEM TEST" /
"Treatment Condition: &TL._a1&a1._a2&a2._r&r. Rep: &rep. ITEM PARMS
CALCULATED." /
'>GLOBAL DFName = ""&dfname..xdt"", '/
'       NPArm = 3, '/
'       SAVe;'/
'>SAVE PARm =  ""&dfname..par"", ' /
'     SCOre =  ""&dfname..sco""; ' /
">LENGTH NITems = (&nitem.); " /
">INPUT NTOtal = &nitem., " /
'      NALt = 2, ' /
'      NIDchar = 5; ' /
">ITEMS INUmber = (1(1)&nitem.), INAmes = (SIM001(1)SIM&nitem.); "/
">TEST1 TNAme = &TL.&a1.&a2.&r.&rep.; "/
"(5A1,7X,&nitem.A1) "/
'>CALIB CYCles = 125, NEWton = 75, CRIt = 0.001, '/
'      PLOt = 1.0000, '/
'      ACCel = 1.0000, '/
'      TPRior; '/
'>SCORE RSCtype = 0, '/
'      METhod=1, '/
'      MOMents=1, '/
'      INFo = 2; '/
;
run;
%mend do_it;
%do_it(20,1,1,1,01,20);
%do_it(20,1,1,1,02,20);
.
.
.
%do_it(20,1,3,4,10,20);
```

Table B3: Sample SAS Program for Data Analysis

```
/* program: diss4_tl20_a11.sas
   researcher: john klaric

Strategy:  Data reduction and summarization of the 1800 datasets from the 3x3x4x5
design with 10
replications per treatment condition.

Tactics:  9 programs were written -- one for each
test length * a1 condition.  Each of these 9 summarizes 200 of the 1800 datasets.
Before running these, check *.ph1 files for negative, low point biserials for individual
items.
Then, check convergence (EM cycles, newton cycles) with convergence.sas

If given replications still do not converge, comment out their macro invocations
to exclude them from the treatment replication analysis.

Data are reduced in the following way:
After these 1800 datasets (REP, each with 1000 recs, one for each simulee) were built,
180 treatment condition datasets were built (CELL, each with 10 records, one for each
treatment).  From this, a single DESIGN dataset was built containing 180 records (one
for each treatment)

The True ability estimates were generated with mirtgen2 -- 550 items for 1000 simulees,
all with item discriminations of 2.5, from d=-3 to +3, c=0.0 for all items.

Definitions:
Treatment condition   Levels       Values            Factor type
TestLength(tl)        1,2,3        20,30,60          Fixed
a1 (Item discrim)     1,2,3        .5, .75, 1.0      Random
a2 (Direct effect)    1,2,3,4      0, .25, .5, .75   Random
r  (Indirect effect)  1,2,3,4,5    -.2,0,.2,.4,.6.   Random

rep (rep number)                   01-10
*/
options nomprint nosymbolgen;
-%macro replication(tl,a1,a2,r,rep);
/******Step 1.: input theta (ability, pacing) file******/
%let pthnam2=C:\Program Files\bilogmg\Simulations\2009_q\tlen&tl._np\A1&a1._&tl.;
filename in1 "&pthnam2.\T&TL._a1&a1._a2&a2._r&r._&rep..sco" ;
data score;
infile in1;
if _n_=1 then input //
```

```
@6 simulee $5./@39 theta $9.;
else input @6 simulee $5./@39 theta $9.;
run;

data thetaT&tl._a1&a1._a2&a2._r&r._&rep._2;
set score;
ntheta=input(theta,9.7);
mergevar=1;
run;
proc means; var ntheta; run;
/****** Step 2.: Calculating response probabilities:*****
2.1 readin_parms estimated for each item with bilogmg,
2.2 mirtgen readin (01 file [*.xdt])
2.3 probability calculation, residual calculation
2.4 replication dataset build
*/
filename in2 "&pthnam2.\T&tl._a1&a1._a2&a2._r&r._&rep..PAR";
filename in3 "&pthnam2.\T&tl._a1&a1._a2&a2._r&r._&rep..XDT";
filename in4 "&pthnam2.\T&tl._a1&a1._a2&a2._r&r._&rep..TDT";

data parms;
length item $5.;
infile in2;
if _n_=1 then input ////
@1 item $5. @40 slope $7. @50 slopeSE $7. @59 threshold $8. @70 threshSE $7.
@100 asymptote $7. @110 asympSE $7.;
else input @1 item $5. @40 slope $7. @50 slopeSE $7. @59 threshold $8. @70 threshSE
$7. @100 asymptote $7. @110 asympSE $7.;
run;
data num_parms (drop=slope slopeSE threshold asymptote asympSE threshSE);
set parms;
nslope=input(slope,8.4);
nslopeSE=input(slopeSE,8.4);
nthreshold=input(threshold,8.4);
nthresSE=input(threshSE,8.4);
nasymptote=input(asymptote,8.4);
nasympSE=input(asympSE,8.4);
run;
proc print data=num_parms; var item nslopeSE nthresSE nasympSE;
title Standard Errors for T&tl._a1&a1._a2&a2._r&r._&rep.; run;
proc datasets library=work; delete parms; quit;
%macro do_it(itm);
data trans_parms&itm. (keep = a&itm. b&itm. c&itm. aSE&itm. bSE&itm. cSE&itm.
mergevar);
```

```
set num_parms;
if item="SIM&itm." then do;
  a&itm.=nslope;
  aSE&itm.=nslopeSE;
  b&itm.=nthreshold;
  bSE&itm.=nthresSE;
  c&itm.=nasymptote;
  cSE&itm.=nasympSE;

  mergevar=1;
  output trans_parms&itm.;
end;
run;
%mend do_it;
%do_it(01);%do_it(02);%do_it(03);%do_it(04);%do_it(05);%do_it(06);%do_it(07);%do
_it(08);%do_it(09);%do_it(10);%do_it(11);%do_it(12);%do_it(13);%do_it(14);
%do_it(15); %do_it(16);%do_it(17);%do_it(18);%do_it(19);%do_it(20);
/*the following code just transposes the parameter datasets from bilog*/
data comboT&tl._a1&a1._a2&a2._r&r._&rep.;
merge trans_parms01 trans_parms02 trans_parms03 trans_parms04 trans_parms05
trans_parms06 trans_parms07 trans_parms08 trans_parms09 trans_parms10
trans_parms11 trans_parms12 trans_parms13 trans_parms14 trans_parms15
trans_parms16 trans_parms17 trans_parms18 trans_parms19 trans_parms20 ;
by mergevar;
run;
proc datasets library=work; save thetaT&tl._a1&a1._a2&a2._r&r._&rep._2
comboT&tl._a1&a1._a2&a2._r&r._&rep.;
quit;
/* merge theta (ability) with parm datasets into prelim_3plprob dsets
*/
data p_3plT&tl._a1&a1._a2&a2._r&r._&rep. (drop=aSE01-aSE&tl. bSE01-bSE&tl.
cSE01-cSE&tl.);
merge thetaT&tl._a1&a1._a2&a2._r&r._&rep._2
comboT&tl._a1&a1._a2&a2._r&r._&rep.;
by mergevar;
run;
/* import scored 0/1 files, and then merge these with prelim_3plprob dsets*/
data score_T&tl._a1&a1._a2&a2._r&r._&rep.;
infile in3;
input @1 id $5. @9 Total $3. @13 (scr1-scr&tl.) ($1.);
mergevar=1;
run;
data pscr_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
```

```
merge p_3plT&tl._a1&a1._a2&a2._r&r._&rep.
score_T&tl._a1&a1._a2&a2._r&r._&rep.;
by mergevar;
run;
proc datasets library=work; save thetaT&tl._a1&a1._a2&a2._r&r._&rep._2
comboT&tl._a1&a1._a2&a2._r&r._&rep. pscr_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
quit;
/* calculate response probabilities
   checked out against hambleton, swaminathan, and rogers (1991) tables,
   page 28, 30)
   var2 is ability estimated with mirtgen2
*/
data prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.
   (drop=i term1_1-term1_&tl. term2_1-term2_&tl. term3_1-term3_&tl.);
set pscr_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
d=1.7;
array a{*} a01-a&tl.;
array b{*} b01-b&tl.;
array c{*} c01-c&tl.;
array num{*} num1-num&tl.;
array den{*} den1-den&tl.;
array term1_{*} term1_1-term1_&tl.;
array term2_{*} term2_1-term2_&tl.;
array term3_{*} term3_1-term3_&tl.;
array prob{*} prob1-prob&tl.;
array res{*} res1-res&tl.;
array scr{*} scr1-scr&tl.;

do i=1 to &tl.;
num[i]=exp(d*a[i]*(ntheta-b[i]));
den[i]=1+exp(d*a[i]*(ntheta-b[i]));
term3_[i]=num[i]/den[i];
term2_[i]=1-c[i];
term1_[i]=c[i];
prob[i]=term1_[i]+(term2_[i]*term3_[i]);
res[i]=scr[i]-prob[i];
end;
run;
/* rep dataset build for step 2
*/
data diss.T&tl._a1&a1._a2&a2._r&r._&rep. (keep=tl a1 a2 r rep a01-a&tl. b01-b&tl.
c01-c&tl.);
set prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
```

```
if _n_=1 then output diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
run;
proc means data=prob_3plT&tl._a1&a1._a2&a2._r&r._&rep. noprint;
var ntheta prob1-prob&tl. res1-res&tl.;
output out=temp&tl.&a1.&a2.&rep. mean=mntheta mnprob1-mnprob&tl. mnres1-
mnres&tl. std=repsdtheta;
/* note that residual standard deviations are NOT retained*/;
run;
data temp&tl.&a1.&a2.&rep._2;
set temp&tl.&a1.&a2.&rep.;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
run;
data diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. temp&tl.&a1.&a2.&rep._2;
by tl a1 a2 r rep;
run;
/*
*****Step 3.: Yen Q3 statistics and application of Yen correction*****;
3.1 Residual intercorrelations
3.2 Selection of item pair correlations
3.3 Application of Yen corrction (-1/[&tl.-1])
3.4 Checking Yen statistics with univariate statistics
3.5 Continuing replication dataset build
*/

data prob_3plT&tl._a1&a1._a2&a2._r&r._&rep._2 (drop=scr1-scr&tl. a01-a&tl. b01-
b&tl. c01-c&tl.
    simulee);
set prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
run;
proc corr data=prob_3plT&tl._a1&a1._a2&a2._r&r._&rep._2 outp=q3 noprint;
var res1-res&tl.; with res1-res&tl.; run;
/*the following datastep just takes the residual intercorrelations and identifies item pairs
*/
data pairs_q3test;
set q3;
if _TYPE_ ^='CORR' then delete;
indx=substr(_NAME_,4,2);
Item=indx;
run;
%macro do_it(item);
data test&item. (keep=pair q3 itm);
set pairs_q3test;
array res{*} res1-res&tl.;
```

```
if item="&item." then do;
  itm=input(&item.,best.);
  do i = itm+1 to &tl.;
    ci=put(i,2.);
  pair=&item. || ',' || ci;
  q3=res[i];
  output;
  end;
end;
else delete;
run;
/* the following code applies Yen''s correction -- 1993
*/
data test2_&item. (keep=pair q3un itm q3c group);
set test&item.;
q3un=q3;
criterion=-1/(&tl.-1);
q3c=q3-criterion;
if q3c<=.0275 then group=1;
else if q3c>.0275 then group=2;
*testing put "&item." criterion q3un q3c;
run;
%mend do_it;
%do_it(1);
%do_it(2);
%do_it(3);%do_it(4);%do_it(5);%do_it(6);%do_it(7);%do_it(8);%do_it(9);
%do_it(10);%do_it(11);%do_it(12);%do_it(13);%do_it(14);%do_it(15);%do_it(16);%do
_it(17);%do_it(18);
%do_it(19);%do_it(20);
/* the following code appends item datasets containing the independence statistics for
each item
*/

data combo;
set test2_1  test2_2  test2_3  test2_4  test2_5  test2_6  test2_7  test2_8  test2_9  test2_10
   test2_11 test2_12 test2_13 test2_14 test2_15 test2_16 test2_17 test2_18 test2_19
test2_20;
run;
/*
proc univariate data=combo plots; var q3un; title &tl. a1&a1. a2&a2. r&r. &rep. --
q3uncorrected; run;
*/
proc univariate data=combo plots noprint; var q3c; title &tl. a1&a1. a2&a2. r&r. &rep. --
q3corrected; run;
```

```
proc means data=combo noprint; var q3un;
output out=yen&tl.&a1.&a2.&rep. mean=mnq3 std=sdq3; run;
/* rep dataset build for step 3*/
data yen&tl.&a1.&a2.&rep._2;
set yen&tl.&a1.&a2.&rep.;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
run;

data diss.T&tl._a1&a1._a2&a2._r&r._&rep. (drop=_TYPE_);
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. yen&tl.&a1.&a2.&rep._2;
by tl a1 a2 r rep;
run;
proc datasets library=work; save thetaT&tl._a1&a1._a2&a2._r&r._&rep._2
comboT&tl._a1&a1._a2&a2._r&r._&rep. pscr_3plT&tl._a1&a1._a2&a2._r&r._&rep.
prob_3plT&tl._a1&a1._a2&a2._r&r._&rep. yen&tl.&a1.&a2.&rep._2; quit;

/****** Step 4.: Calculating response time statistics:*****
4.1 readin response time data (minutes) estimated with mirtgen (*.tdt)
4.2 converting to seconds and relabeling tau variable
4.3 mean tau calculation, correlation between theta1 and theta2
4.4 semipartial correlations (item residual and item RT)
4.5 replication dataset build

reading in response time data, converting, relabeling,
        merging with response probabilities dataset, optimizing,
        checking that the tau values from the import files are equivalent*/
PROC IMPORT OUT= WORK.RT_T&tl._a1&a1._a2&a2._r&r._&rep.
            DATAFILE="&pthnam2.\T&tl._a1&a1._a2&a2._r&r._&rep..TDT"
            DBMS=DLM REPLACE;
     DELIMITER='2C'x;
     GETNAMES=NO;
     DATAROW=1;
RUN;
data RT_T&tl._a1&a1._a2&a2._r&r._&rep._2 (drop=var1-var22 i);  *hardcode here too;
set RT_T&tl._a1&a1._a2&a2._r&r._&rep.;
array rtsec{*} rtsec1-rtsec&tl.;
array var{*} var1-var22;  *hardcode the final array subscript;
mergevar=1;
Tau_Pacing=var2;
do i=1 to &tl.;
   RTsec[i]=var[i+2]*60;
   end;
run;
data prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
```

```
merge prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.
RT_T&tl._a1&a1._a2&a2._r&r._&rep._2;
by mergevar;
run;
proc datasets library=work;
save prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.
RT_T&tl._a1&a1._a2&a2._r&r._&rep._2 yen&tl.&a1.&a2.&rep._2; quit;
data RT_3plT&tl._a1&a1._a2&a2._r&r._&rep._3 (drop=scr1-scr&tl. a01-a&tl. b01-b&tl.
c01-c&tl.
    prob1-prob&tl. d den1-den&tl. num1-num&tl.);
set prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.;
run;
/* calculating means for tau parameter, correlating pacing with ability*/
proc means data=RT_3plT&tl._a1&a1._a2&a2._r&r._&rep._3 noprint; var tau_pacing;
output out=rtsummary1&tl.&a1.&a2.&rep. mean=repmntau std=repsdtau; run;
data rtsummary1&tl.&a1.&a2.&rep._2 (drop=_TYPE_);
set rtsummary1&tl.&a1.&a2.&rep.;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
run;
proc corr data=RT_3plT&tl._a1&a1._a2&a2._r&r._&rep._3 outp=corrthetas noprint;
var tau_pacing; with ntheta; run;
data corrthetas2;
set corrthetas;
if _TYPE_ ^='CORR' then delete;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
run;
/* the following code calculates semipartial correlation coefficients for individual items*/
%macro do_it(item);
proc corr data=RT_3plT&tl._a1&a1._a2&a2._r&r._&rep._3 outp=corrsemi&item.
noprint;
var res&item.; with rtsec&item.; run;

data csemi&item._2 (drop=_NAME_ res&item.);
set corrsemi&item.;
if _TYPE_ ^='CORR' then delete;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
RT_ResSP&item.=res&item.;
run;
%mend do_it;
%do_it(1);%do_it(2);%do_it(3);%do_it(4);%do_it(5);%do_it(6);%do_it(7);%do_it(8);%d
o_it(9); %do_it(10);%do_it(11);%do_it(12);%do_it(13);%do_it(14);%do_it(15);
%do_it(16);%do_it(17); %do_it(18);%do_it(19);%do_it(20);
/* as in a previous step, merging the files containing item statistics (here, semipartials)
  into an overall dataset*/
```

```
data corrsemi_overall;
merge corrthetas2 csemi1_2 csemi2_2 csemi3_2 csemi4_2 csemi5_2 csemi6_2 csemi7_2
csemi8_2 csemi9_2 csemi10_2 csemi11_2 csemi12_2 csemi13_2 csemi14_2
csemi15_2 csemi16_2 csemi17_2 csemi18_2 csemi19_2 csemi20_2;
by tl a1 a2 r rep;
run;
/* rep dataset build, step 4*/
data diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. corrsemi_overall;
by tl a1 a2 r rep;
run;
proc univariate data=diss.T&tl._a1&a1._a2&a2._r&r._&rep. noprint;
var rt_ressp1-rt_ressp&tl.; output out=look median=mdnSP1_&tl.; title rt_ressp1-&tl.;
run;
data look2;
set look; tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep."; run;
data diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. look2;
by tl a1 a2 r rep;
run;
data diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. corrthetas2;
by tl a1 a2 r rep;
run;
data diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. rtsummary1&tl.&a1.&a2.&rep._2;
by tl a1 a2 r rep;
run;
proc datasets library=work;
save prob_3plT&tl._a1&a1._a2&a2._r&r._&rep.
RT_T&tl._a1&a1._a2&a2._r&r._&rep._2 yen&tl.&a1.&a2.&rep._2 corrsemi_overall;
quit;
/****** Step 5.: Calculating bias, MSE, RMSE statistics:*****
5.1 making the true theta dataset into a temporary dataset
5.2 merging the temporary dset with the build so far
5.3 calculating bias (var2-ttheta), squared bias [(var2-ttheta)**2] for each person
5.4 determining bias and squared bias means (MSE) across all persons in each replication
5.5 calculating RMSE for each replication by taking the square root of MSE
5.6 replication dataset build
*/
data truetheta;
set diss.truetheta1109; mergevar=1; run;

data allT&tl._a1&a1._a2&a2._r&r._&rep.;
```

```
merge prob_3plT&tl._a1&a1._a2&a2._r&r._&rep. truetheta;
by mergevar;
run;

*/
Calculating bias, squared bias statistics for each person      */;
data biasRMSE_T&tl._a1&a1._a2&a2._r&r._&rep.;
set allT&tl._a1&a1._a2&a2._r&r._&rep.;
biascalc_&tl._a1&a1._a2&a2._r&r._&rep.= ntheta-ttheta;
MSEcalc_&tl._a1&a1._a2&a2._r&r._&rep. = (ntheta-ttheta)**2;
run;
proc datasets library=work; delete allT&tl._a1&a1._a2&a2._r&r._&rep.; quit;
/* finding bias, squared bias means, calculating RMSE, rep dataset build, step 5
MSE is mean of sum of square error, and RMSE is sqrt(MSE)*/
proc means data=biasRMSE_T&tl._a1&a1._a2&a2._r&r._&rep. noprint;
var biascalc_&tl._a1&a1._a2&a2._r&r._&rep.
MSEcalc_&tl._a1&a1._a2&a2._r&r._&rep.;
output out=MSE&tl.&a1.&a2.&rep. mean=BIAS MSE std=repsdBIAS repsdMSE; run;
data RMSE&tl.&a1.&a2.&rep. (drop=_TYPE_);
set MSE&tl.&a1.&a2.&rep.;
RMSE = sqrt(MSE);
tl="&tl."; a1="&a1."; a2="&a2."; r="&r."; rep="&rep.";
run;
data diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
merge diss.T&tl._a1&a1._a2&a2._r&r._&rep. RMSE&tl.&a1.&a2.&rep.;
by tl a1 a2 r rep;
run;
%mend replication;
%replication(20,1,1,1,01);
%replication(20,1,1,1,02);
.
.
.
%replication(20,1,3,4,10);
```

Table B4. SAS Program for Building Summary Datasets

```
/* program: diss4_design_20_1.sas
   researcher: john klaric

before running this program that builds the DESIGN dataset,
build the 1800 replication datasets by running the 9 test length * a1
programs, and then building the 180 treatment condition datasets.

Definitions:
Treatment condition   Levels      Values            Factor type
TestLength(tl)        1,2,3       20,30,60          Fixed
a1 (Item discri)      1,2,3       .5, .75, 1.0      Random
a2 (Direct effect)    1,2,3,4     0, .25, .5, .75   Random
r  (Indirect effect)  1,2,3,4,5   -.2,0,.2,.4,.6.   Random
rep (rep number)                  01-10

Step 1:  build each of the 180 CELL (treatment condition) datasets, 10 replications in
each dataset.*/

%macro cell(tl,a1,a2,r,rep);
data tpT&tl._a1&a1._a2&a2._r&r._&rep.;
set diss.T&tl._a1&a1._a2&a2._r&r._&rep.;
run;
proc append out=diss.cellT&tl._a1&a1._a2&a2._r&r.
data=tpT&tl._a1&a1._a2&a2._r&r._&rep.;
run;
%mend cell;
%cell(20,1,1,1,01);
.
.
.
%cell(20,1,4,5,10);

%macro cellstats(tl,a1,a2,r);
data tpT&tl._a1&a1._a2&a2._r&r.;
set diss.cellT&tl._a1&a1._a2&a2._r&r.;
run;
proc means data=tpT&tl._a1&a1._a2&a2._r&r. ; *noprint;
title This provides the mean of salient variables across reps in each treatment condition;
var
mntheta repmntau repsdtheta repsdtau tau_pacing bias repsdbias mse rmse mnq3 sdq3
RT_ResSP1-RT_ResSP&tl.
a01-a&tl. b01-b&tl. c01-c&tl. mnprob1-mnprob&tl. mnres1-mnres&tl.;
```

```
output out=trtmean_&tl.&a1.&a2.&r.
mean=cellmntheta cellmntau cellsdtheta cellsdtau cellmncorrthetas cellbias cellsdbias
cellmse cellrmse cellmnq3 cellmnstdq3
cellmnRTResSP1-cellmnRTResSP&tl.
mna1-mna&tl. mnb1-mnb&tl. mnc1-mnc&tl. cellmnprob1-cellmnprob&tl. cellmnres1-
cellmnres&tl.;
title &tl. &a1. &a2 &r.;
run;
data trtmean2_&tl.&a1.&a2.&r.; set trtmean_&tl.&a1.&a2.&r.; mergevar=1; run;

proc transpose data=trtmean_&tl.&a1.&a2.&r. out=median_&tl.&a1.&a2.&r.; var
cellmnRTResSP1-cellmnRTResSP&tl.; run;

proc univariate data=median_&tl.&a1.&a2.&r.;*noprint;
var col1; output out=look median=cellmdnSP1_&tl.; run;
data look2;
set look;
mergevar=1;
run;
data diss.trtmean2_&tl.&a1.&a2.&r. (drop=mergevar _TYPE_);
merge trtmean2_&tl.&a1.&a2.&r. look2;
tl="&tl."; a1="&a1."; a2="&a2."; r="&r.";
by mergevar;
run;

%mend cellstats;
%cellstats(20,1,1,1);
.
.
.
%cellstats(20,1,4,5);

/*Step 2:  take each of the TRT datasets, and build the DESIGN dataset*/
%macro design(tl,a1,a2,r);
proc append data=diss.trtmean2_&tl.&a1.&a2.&r. out=diss.design20_2009_npp2;
run;
%mend design;
%design(20,1,1,1);
.
.
.
%design(20,1,4,5);
proc print data=diss.design20_2009_NPp2; title Design 2009; run;
```

Table B5: Sample SYSTAT Program for Graphics

```
use a120
ORIGIN =  2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmnTau*R$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL20, A1: 1' LTITLE='a2 Effect' LLABEL=  '0.0'  '0.25'  '0.5'  '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7  DASH=1,11,7,1

use a130
ORIGIN =  2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL30, A1: 1’ LTITLE='a2 Effect' LLABEL=  '0.0'  '0.25'  '0.5'  '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7  DASH=1,11,7,1

use a160
ORIGIN =  2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL60, A1: 1' LTITLE='a2 Effect' LLABEL=  '0.0'  '0.25'  '0.5'  '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7  DASH=1,11,7,1

use a220
ORIGIN =  2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
```

```
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL20, A1: 2' LTITLE='a2 Effect' LLABEL= '0.0' '0.25' '0.5' '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7 DASH=1,11,7,1

use a230
ORIGIN = 2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL30, A1: 2' LTITLE='a2 Effect' LLABEL= '0.0' '0.25' '0.5' '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7 DASH=1,11,7,1

use a260
ORIGIN = 2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL60, A1: 2' LTITLE='a2 Effect' LLABEL= '0.0' '0.25' '0.5' '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7 DASH=1,11,7,1

use a320
ORIGIN = 2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL20, A1: 3' LTITLE='a2 Effect' LLABEL= '0.0' '0.25' '0.5' '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7 DASH=1,11,7,1

use a330
```

```
ORIGIN =  2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL30, A1: 3' LTITLE='a2 Effect' LLABEL=  '0.0'  '0.25'  '0.5'  '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7  DASH=1,11,7,1

use a360
ORIGIN =  2.25IN, -5.50IN
THICK = 2.000
CSIZE = 1.250
SCALE = 100,100
FACET
EYE = -6,-8,6 /RECTANGULAR
LINE cellmntau*r$ / OVERLAY GROUP=A2$ XLABEL='Indirect Effect: Latent
Variable Correlation' XGRID YLABEL='Pacing (Theta2)' YMIN=-0.4 YMAX=0.4,
TITLE='TL60, A1: 3' LTITLE='a2 Effect' LLABEL=  '0.0'  '0.25'  '0.5'  '0.75',
LEGEND=3.15IN,COLOR=10,7,10,7  DASH=1,11,7,1
```

Table B6. BILOG MG Program Used in Real Data Study[1]

```
COMPUTER SKILLS OPERATIONAL DATA FROM FALL 2005
Priors, all forms, truncated dataset
>GLOBAL DFName = 'C:\Program Files\bilogmg\Dissertation\timescr7.DAT',
       NPArm = 3, SAVe;
>SAVE PARm = 'CS05_7.PAR', SCOre = 'CS05_7.SCO';
>LENGTH NITems = (54);
>INPUT NTOtal = 54, NALt = 4, NIDchar = 9,
      KFName = 'C:\Program Files\bilogmg\Dissertation\timescr7.DAT';
>ITEMS INUmber = (1(1)54), INAmes = (CS1(1)CS54);
>TEST TNAme = 'CS TST1';
(9A1, 1X, 54A1)
>CALIB CYCles = 125, NEWton = 75, CRIt = 0.0001, ACCel = 1.0000,
      TPRior, REAdpri, NOAdjust, plot=1.0;
>SCORE info=2, method=1, moments, rsctype=0;
>PRIORS1 ALPha = (6.0000, 1.1000, 1.1000, 1.1000, 6.0000, 6.0000,
               6.0000, 1.1000, 1.1000, 6.0000(0)4, 1.1000(0)4,
               6.0000, 6.0000, 6.0000, 1.1000, 1.1000, 6.0000,
               6.0000, 1.1000, 1.1000, 6.0000(0)4, 1.1000, 1.1000,
               6.0000, 1.1000(0)4, 6.0000, 1.1000, 1.1000, 6.0000,
               6.0000, 1.1000, 6.0000, 1.1000, 6.0000, 1.1000,
               6.0000, 6.0000, 6.0000, 1.1000, 1.1000, 1.1000, 6.0000),
        BETa = (16.0000, 10000.0000, 10000.0000, 10000.0000, 16.0000,
              16.0000, 16.0000, 10000.0000, 10000.0000, 16.0000(0)4,
              10000.0000(0)4, 16.0000, 16.0000, 16.0000, 10000.0000,
              10000.0000, 16.0000, 16.0000, 10000.0000, 10000.0000,
              16.0000(0)4, 10000.0000, 10000.0000, 16.0000,
              10000.0000(0)4, 16.0000, 10000.0000, 10000.0000,
              16.0000, 16.0000, 10000.0000, 16.0000, 10000.0000,
              16.0000, 10000.0000, 16.0000, 16.0000, 16.0000,
              10000.0000, 10000.0000, 10000.0000, 16.0000);
```

[1] Adapted from L. Kramer. (2006). BILOG MG Computer Program for 2005 Computer Skills Test [Computer Program]. Raleigh NC: North Carolina Department of Public Instruction.

Table B7. A Partial Listing of SAS Programs

```
/* program: reading_time.sas   researcher: john klaric   july, 2008*/
options symbolgen mprint;
%macro reading(file);
filename in1
"C:\Documents and Settings\owner\My Documents\Dissertation backup\SAS
Datasets\studentdata_&file..txt";
data tp&file.;
infile in1 lrecl=366 stopover;
/*use of the stopover function in the infile statement identifies bad record lengths*/
input sid $9. @10 x1 $25. @35 sex $1. @36 eth $1.
    @37 x2 $6. @43 grade $1. @44 x3 $57. @101 flavor $1. @102 lengthnc $1.
    @103 x4 $44. (I1-I54) ($1.) (T1-T54) ($3.) @363 TotalT $3.;
run;
/* drop identifiers and eliminate duplicates*/
data tp2&file. (drop=x1 x2 x3 x4);
set tp&file.;
run;
proc sort nodupkey data=tp2&file. out=sorted2&file.; by sid; run;
/* make item response and time variables numeric, summing item response
  times, and removing administrations where the total response time is
  either missing or 0*/
data tp3_1&file. tp3&file. error&file.;
set sorted2&file.;
array item{*} 3. item1-item54;
array I{*} $1. I1-I54;
array time{*} 3. time1-time54;
array t{*} $3. t1-t54;
array miss{*} 3. miss1-miss54;
do k=1 to 54;
  Item[k]=input(I[k],1.);
  Time[K]=input(T[k},3.);
  if item[K]=. then time[K]=.;
  if item[K]=. then miss[K]=1; else miss[K]=0;
  end;
tottime=sum(of time1-time54);
totmiss=sum(of miss1-miss54);
if tottime in (.,0) then output error&file.;
else output tp3&file.; run;
%mend reading;
%reading( a);
%reading( b);
%reading( c);
%reading( d);
```

```
/* program : 3PL_prob.sas
   programmer: john klaric
/* this dataset has scores, but create combo dataset with bilogreading.sas first*/
data tp (keep=item1-item54 flavor eth sex mergevar lengthnc sid sex scr1-scr54 time1-
time54
            tottime totscr totmiss grade fid);
set diss.truncatedtime2;
newid=_n_;
fid=put(newid,z9.);
mergevar=1;
run;
proc sort nodupkey data=tp ; by fid; run;
/* this dataset has bilog parms, theta estimates*/
data parmstheta (drop=aSE01-aSE54 bSE01-bSE54 cSE01-cSE54 eapSE);
set combo;
run;
data scoredparms;
merge parmstheta tp; by fid; run;
data probability;
/* checked out against hambleton, swaminathan, and rogers (1991)
   tables, page 28, 30)*/
set scoredparms;
d=1.7;
array a{*} a01-a54;
array b{*} b01-b54;
array c{*} c01-c54;
array num{*} num1-num54;
array den{*} den1-den54;
array term1_{*} term1_1-term1_54;
array term2_{*} term2_1-term2_54;
array term3_{*} term3_1-term3_54;
array prob{*} prob1-prob54;
array res{*} res1-res54;
array scr{*} scr1-scr54;

do I=1 to 54;
num[I]=exp(d*a[I]*(eap-b[I]));
den[I]=1+exp(d*a[I]*(eap-b[I]));
term3_[I]=num[I]/den[I];
term2_[I]=1-c[I];
term1_[I]=c[I];
prob[I]=term1_[I]+(term2_[I]*term3_[I]);
res[I]=scr[I]-prob[I];
end;
```

```
run;

proc print data=probability;
var eap prob1-prob6 res1-res6;
format prob1-prob6 res1-res6 8.3;
run;
/*--------------------
proc datasets library=work; delete probresmeans; quit;
proc print data=residuals; var TYPE NAME res1; run;

calculation of variance co-var matrix
determination of column mean, min, max for each item
output to table b12
output to excel for tables b9-b11
*/
data tp;
set diss.prob3plresid;
run;
proc corr data=tp outp=residuals cov; var res1-res54; with res1-res54; run;
proc corr data=residuals; var xres1-xres54; with xres1-xres54; run;
data covariance;
set residuals;
if TYPE='COV' then output covariance;
run;
```

```
/* program:  q3corr.sas   researcher: john klaric
/*----------------Q3 statistic-------------------;*/
data tp (drop=scr1-scr54 sid fid I total totmiss totscr lengthnc
        mergevar ncorrect pctcorrect correct eth flavor grade);
set diss.prob3plresid; run;
proc corr data=tp outp=q3 noprint; var res1-res54; with res1-res54; run;
data pairs_q3test; set q3; if TYPE ^=’CORR’ then delete;
indx=substr(_NAME_,4,2); Item=indx; run;
%macro do_it(item);
data test&item. (keep=pair q3 itm); set pairs_q3test;
array res{*} res1-res54;
if item=”&item.” then do;
  itm=input(&item.,best.);
  do I = itm+1 to 54;
    ci=put(I,2.);
  pair=&item. || ‘,’ || ci;
  q3=res[I];
  output;
  end;
end;
else delete; run;
data test2_&item. (keep=pair q3un itm q3c group);
set test&item.;
q3un=q3; criterion=-1/53; q3c=q3-criterion;
if q3c<=.0275 then group=1;
else if q3c>.0275 then group=2;
run;
%mend do_it;
%do_it(1);%do_it(2);%do_it(3);%do_it(4);%do_it(5);%do_it(6);%do_it(7);%do_it(8);
%do_it(9);%do_it(10);%do_it(11);%do_it(12);%do_it(13);%do_it(14);%do_it(15);
%do_it(16);%do_it(17);%do_it(18);%do_it(19);%do_it(20);%do_it(21);%do_it(22);
%do_it(23);%do_it(24);%do_it(25);%do_it(26);%do_it(27);%do_it(28);%do_it(29);
%do_it(30);%do_it(31);%do_it(32);%do_it(33);%do_it(34);%do_it(35);%do_it(36);
%do_it(37);%do_it(38);%do_it(39);%do_it(40);%do_it(41);%do_it(42);%do_it(43);
%do_it(44);%do_it(45);%do_it(46);%do_it(47);%do_it(48);%do_it(49);%do_it(50);
%do_it(51);%do_it(52);%do_it(53);%do_it(54);
data combo;
set test2_1  test2_2  test2_3  test2_4  test2_5  test2_6  test2_7  test2_8  test2_9  test2_10
test2_11 test2_12 test2_13 test2_14 test2_15 test2_16 test2_17 test2_18 test2_19
test2_20 test2_21 test2_22 test2_23 test2_24 test2_25 test2_26 test2_27 test2_28
test2_29 test2_30 test2_31 test2_32 test2_33 test2_34 test2_35 test2_36 test2_37
test2_38 test2_39 test2_40 test2_41 test2_42 test2_43 test2_44 test2_45 test2_46
test2_47 test2_48 test2_49 test2_50 test2_51 test2_52 test2_53 test2_54 ;
run;
```

Appendix C. Form Item Statistics and Residual Variance Co-Variance Matrix

Table C1. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 1: N=17266

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	62.6	0.227	0.290	19	62.0	0.385	0.491	37	73.7	0.541	0.730
2	73.8	0.493	0.665	20	68.9	0.361	0.473	38	60.4	0.334	0.424
3	26.6	0.408	0.549	21	32.3	0.410	0.535	39	54.8	0.470	0.591
4	59.8	0.459	0.582	22	78.3	0.342	0.479	40	79.4	0.349	0.496
5	51.0	0.268	0.336	23	82.0	0.440	0.644	41	56.8	0.452	0.569
6	60.8	0.282	0.359	24	42.7	0.237	0.299	42	42.0	0.281	0.355
7	51.0	0.305	0.383	25	56.3	0.413	0.520	43	73.9	0.556	0.752
8	63.2	0.373	0.477	26	32.9	0.503	0.653	44	65.6	0.412	0.531
9	76.9	0.356	0.493	27	63.3	0.469	0.600	45	80.4	0.463	0.666
10	69.4	0.363	0.477	28	74.9	0.265	0.361	46	51.1	0.456	0.571
11	43.4	0.303	0.382	29	74.7	0.329	0.448	47	28.6	0.535	0.711
12	75.1	0.168	0.228	30	41.1	0.223	0.282	48	33.8	0.280	0.362
13	22.9	0.165	0.229	31	34.2	0.420	0.542	49	64.2	0.277	0.355
14	10.5	0.300	0.506	32	29.1	0.440	0.583	50	60.4	0.428	0.543
15	22.3	0.370	0.517	33	41.7	0.320	0.404	51	15.9	0.362	0.546
16	66.0	0.471	0.609	34	33.4	0.564	0.731	52	40.2	0.473	0.600
17	29.9	0.451	0.594	35	37.3	0.448	0.572	53	44.1	0.631	0.794
18	48.6	0.326	0.408	36	48.9	0.571	0.715	54	51.7	0.289	0.362

Table C2. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 2: N=17270

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	62.9	0.219	0.280	19	61.3	0.386	0.491	37	73.2	0.541	0.728
2	73.6	0.485	0.654	20	69.4	0.366	0.481	38	60.9	0.348	0.442
3	29.0	0.440	0.583	21	34.0	0.430	0.556	39	55.0	0.466	0.586
4	59.8	0.467	0.591	22	78.0	0.342	0.479	40	79.4	0.344	0.489
5	48.6	0.284	0.356	23	82.3	0.434	0.638	41	57.1	0.437	0.551
6	62.5	0.278	0.354	24	42.2	0.257	0.324	42	43.7	0.277	0.348
7	52.1	0.316	0.396	25	56.2	0.432	0.544	43	73.3	0.552	0.743
8	62.7	0.367	0.469	26	30.9	0.507	0.665	44	65.4	0.425	0.548
9	77.1	0.340	0.472	27	62.4	0.464	0.593	45	82.1	0.449	0.658
10	69.8	0.373	0.492	28	74.7	0.264	0.358	46	51.0	0.471	0.590
11	43.8	0.304	0.382	29	75.0	0.357	0.487	47	27.6	0.523	0.700
12	75.1	0.157	0.214	30	41.0	0.222	0.281	48	33.7	0.283	0.367
13	23.4	0.161	0.222	31	34.6	0.435	0.561	49	64.5	0.269	0.346
14	10.7	0.291	0.488	32	28.9	0.447	0.593	50	59.7	0.435	0.551
15	22.5	0.385	0.537	33	41.7	0.336	0.424	51	16.4	0.367	0.549
16	66.3	0.474	0.614	34	32.9	0.576	0.748	52	40.6	0.501	0.634
17	29.3	0.444	0.587	35	37.2	0.454	0.580	53	41.8	0.632	0.798
18	48.9	0.339	0.425	36	49.8	0.572	0.717	54	52.0	0.279	0.349

Table C3. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 3: N=17308

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	61.7	0.220	0.280	19	62.1	0.376	0.479	37	73.3	0.551	0.742
2	73.9	0.481	0.651	20	68.8	0.374	0.490	38	60.1	0.316	0.400
3	28.0	0.420	0.561	21	34.5	0.417	0.538	39	55.2	0.474	0.596
4	59.8	0.462	0.586	22	78.0	0.348	0.486	40	79.8	0.359	0.512
5	48.4	0.260	0.326	23	82.7	0.440	0.650	41	57.4	0.419	0.528
6	62.1	0.273	0.348	24	42.2	0.241	0.304	42	43.9	0.284	0.358
7	51.7	0.308	0.387	25	55.8	0.414	0.521	43	73.8	0.551	0.744
8	63.1	0.355	0.454	26	33.1	0.506	0.656	44	65.7	0.422	0.545
9	76.7	0.351	0.486	27	64.2	0.472	0.606	45	81.1	0.456	0.660
10	69.8	0.364	0.479	28	73.5	0.256	0.346	46	51.6	0.454	0.569
11	43.5	0.302	0.380	29	73.7	0.334	0.451	47	33.3	0.569	0.738
12	74.9	0.167	0.227	30	41.1	0.216	0.273	48	34.3	0.270	0.349
13	23.7	0.172	0.237	31	34.8	0.426	0.549	49	65.2	0.291	0.375
14	10.8	0.306	0.512	32	28.9	0.442	0.587	50	60.6	0.450	0.572
15	22.6	0.370	0.515	33	42.0	0.327	0.413	51	16.5	0.362	0.541
16	66.5	0.473	0.613	34	33.4	0.574	0.744	52	40.0	0.485	0.615
17	29.8	0.448	0.591	35	37.9	0.447	0.570	53	45.4	0.639	0.803
18	49.1	0.325	0.408	36	48.6	0.578	0.724	54	52.6	0.278	0.348

Table C4. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 4: N=17369

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	63.6	0.235	0.302	19	60.8	0.371	0.472	37	74.0	0.560	0.758
2	72.4	0.500	0.669	20	68.6	0.385	0.504	38	61.4	0.341	0.434
3	28.3	0.432	0.576	21	36.4	0.433	0.555	39	54.5	0.472	0.593
4	61.4	0.451	0.574	22	78.0	0.345	0.482	40	79.2	0.376	0.532
5	50.0	0.285	0.358	23	82.1	0.439	0.643	41	55.9	0.432	0.543
6	60.8	0.289	0.367	24	41.3	0.236	0.298	42	44.1	0.275	0.347
7	51.3	0.318	0.398	25	55.6	0.411	0.517	43	73.5	0.564	0.760
8	61.8	0.381	0.485	26	32.5	0.513	0.668	44	65.6	0.430	0.555
9	76.7	0.352	0.487	27	62.4	0.477	0.609	45	80.0	0.469	0.670
10	69.8	0.375	0.494	28	74.7	0.274	0.373	46	51.1	0.466	0.584
11	42.9	0.308	0.388	29	74.2	0.362	0.491	47	29.5	0.547	0.722
12	75.4	0.174	0.238	30	41.1	0.219	0.277	48	33.6	0.269	0.349
13	23.4	0.170	0.235	31	37.2	0.435	0.556	49	65.3	0.291	0.376
14	10.6	0.309	0.519	32	28.9	0.453	0.602	50	61.1	0.452	0.575
15	22.3	0.374	0.522	33	42.0	0.330	0.416	51	17.3	0.375	0.554
16	66.0	0.477	0.616	34	33.2	0.565	0.733	52	39.0	0.495	0.630
17	30.3	0.441	0.581	35	36.6	0.449	0.575	53	43.7	0.630	0.793
18	48.6	0.319	0.400	36	48.9	0.584	0.731	54	52.2	0.299	0.375

Table C5. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 5: N=12937

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	61.8	0.227	0.289	19	62.5	0.378	0.482	37	72.8	0.566	0.760
2	70.1	0.513	0.676	20	69.3	0.376	0.494	38	61.1	0.330	0.420
3	27.0	0.415	0.558	21	32.9	0.430	0.559	39	54.3	0.473	0.594
4	58.9	0.471	0.596	22	77.2	0.361	0.501	40	79.7	0.365	0.519
5	47.2	0.276	0.346	23	82.0	0.436	0.639	41	61.3	0.438	0.557
6	61.3	0.282	0.358	24	42.6	0.244	0.308	42	43.6	0.270	0.340
7	51.1	0.301	0.377	25	55.5	0.425	0.535	43	74.6	0.552	0.750
8	62.1	0.388	0.495	26	30.9	0.505	0.662	44	66.5	0.433	0.560
9	75.8	0.352	0.483	27	63.0	0.468	0.599	45	79.6	0.466	0.663
10	70.0	0.373	0.492	28	74.8	0.300	0.408	46	49.8	0.443	0.556
11	42.7	0.303	0.382	29	74.3	0.357	0.483	47	29.4	0.550	0.727
12	75.1	0.181	0.247	30	41.9	0.224	0.283	48	33.7	0.265	0.343
13	23.3	0.167	0.230	31	39.6	0.437	0.555	49	63.7	0.283	0.362
14	10.6	0.303	0.508	32	28.8	0.443	0.588	50	60.0	0.458	0.580
15	22.3	0.372	0.519	33	41.7	0.318	0.402	51	16.0	0.361	0.544
16	65.6	0.492	0.634	34	35.7	0.590	0.758	52	38.6	0.490	0.624
17	29.3	0.451	0.597	35	36.9	0.445	0.570	53	42.5	0.634	0.800
18	48.5	0.330	0.414	36	49.1	0.588	0.737	54	52.3	0.289	0.362

Table C6. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 6: N=12983

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	62.0	0.230	0.293	19	61.4	0.384	0.488	37	73.7	0.570	0.770
2	70.8	0.502	0.665	20	68.5	0.380	0.496	38	60.9	0.338	0.429
3	27.0	0.419	0.562	21	35.2	0.435	0.559	39	55.1	0.491	0.618
4	59.8	0.464	0.588	22	77.6	0.360	0.501	40	80.8	0.363	0.523
5	47.2	0.274	0.344	23	82.2	0.447	0.655	41	57.0	0.446	0.562
6	62.1	0.295	0.376	24	42.6	0.242	0.305	42	43.5	0.275	0.346
7	51.5	0.320	0.401	25	57.0	0.413	0.521	43	75.3	0.542	0.740
8	62.7	0.382	0.487	26	31.3	0.487	0.637	44	65.6	0.438	0.565
9	76.8	0.331	0.458	27	63.2	0.465	0.595	45	79.5	0.458	0.652
10	69.8	0.386	0.508	28	75.4	0.287	0.393	46	50.7	0.459	0.576
11	43.0	0.302	0.380	29	74.8	0.352	0.478	47	27.9	0.521	0.695
12	75.2	0.179	0.244	30	41.8	0.234	0.296	48	34.7	0.277	0.357
13	23.7	0.156	0.215	31	38.4	0.436	0.556	49	63.3	0.280	0.358
14	10.7	0.297	0.498	32	28.9	0.454	0.602	50	60.5	0.446	0.566
15	22.6	0.374	0.520	33	41.8	0.324	0.410	51	16.9	0.364	0.541
16	66.3	0.473	0.612	34	33.3	0.570	0.739	52	39.5	0.492	0.625
17	29.8	0.445	0.587	35	36.6	0.432	0.552	53	43.5	0.622	0.783
18	48.8	0.342	0.428	36	49.9	0.584	0.732	54	52.6	0.295	0.370

Table C7. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 7: N=4356

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	61.6	0.205	0.261	19	63.2	0.354	0.453	37	68.7	0.564	0.738
2	70.8	0.512	0.678	20	69.3	0.377	0.495	38	63.5	0.362	0.464
3	26.5	0.415	0.559	21	33.0	0.422	0.548	39	53.0	0.484	0.608
4	58.0	0.481	0.608	22	76.9	0.370	0.512	40	81.2	0.352	0.511
5	47.2	0.282	0.354	23	84.1	0.408	0.617	41	57.2	0.443	0.558
6	62.2	0.276	0.352	24	41.7	0.240	0.303	42	41.2	0.272	0.344
7	50.3	0.321	0.403	25	55.0	0.407	0.512	43	73.1	0.537	0.721
8	63.1	0.386	0.493	26	27.0	0.469	0.630	44	64.4	0.412	0.530
9	77.3	0.338	0.470	27	63.9	0.431	0.553	45	80.1	0.463	0.663
10	71.1	0.374	0.496	28	77.6	0.248	0.345	46	47.5	0.413	0.519
11	42.4	0.296	0.374	29	74.4	0.344	0.466	47	24.3	0.520	0.712
12	75.5	0.171	0.233	30	40.8	0.219	0.277	48	32.8	0.267	0.348
13	22.3	0.177	0.246	31	34.0	0.416	0.538	49	63.9	0.266	0.341
14	10.4	0.300	0.506	32	28.8	0.451	0.598	50	60.8	0.418	0.531
15	22.2	0.367	0.512	33	42.3	0.372	0.470	51	16.4	0.354	0.529
16	64.9	0.459	0.591	34	33.2	0.560	0.726	52	39.1	0.494	0.627
17	29.3	0.443	0.586	35	38.2	0.463	0.590	53	39.9	0.622	0.788
18	49.2	0.324	0.406	36	50.5	0.562	0.704	54	51.7	0.289	0.362

Table C8. Classical Item Statistics: Using Time-Truncated Data from Fall 2005 Administration of North Carolina's Online Computer Skills Assessment, Form 8: N=4262

		Item-Total Correlation				Item-Total Correlation				Item-Total Correlation	
Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial	Item	*p*-value	Pearson	Serial
1	63.5	0.224	0.287	19	63.2	0.373	0.477	37	72.7	0.571	0.765
2	71.3	0.497	0.660	20	67.8	0.372	0.484	38	60.6	0.364	0.462
3	27.3	0.428	0.573	21	33.0	0.424	0.551	39	54.1	0.491	0.616
4	60.7	0.469	0.596	22	77.1	0.342	0.474	40	81.0	0.354	0.512
5	47.7	0.266	0.334	23	83.4	0.436	0.651	41	59.5	0.478	0.606
6	61.2	0.283	0.360	24	43.3	0.226	0.284	42	47.7	0.282	0.354
7	52.5	0.336	0.421	25	56.7	0.428	0.539	43	72.0	0.561	0.749
8	62.7	0.392	0.500	26	27.6	0.484	0.647	44	66.4	0.429	0.556
9	76.9	0.345	0.478	27	64.8	0.456	0.586	45	79.6	0.443	0.630
10	68.6	0.378	0.495	28	75.6	0.262	0.359	46	49.9	0.466	0.583
11	42.7	0.297	0.375	29	73.8	0.371	0.501	47	27.1	0.538	0.721
12	75.6	0.147	0.201	30	43.0	0.206	0.259	48	32.9	0.257	0.333
13	23.7	0.150	0.207	31	34.9	0.429	0.553	49	62.2	0.292	0.372
14	10.2	0.303	0.515	32	30.5	0.452	0.594	50	62.3	0.485	0.619
15	21.5	0.366	0.514	33	42.0	0.345	0.435	51	17.0	0.355	0.528
16	66.0	0.478	0.618	34	33.7	0.580	0.750	52	39.3	0.495	0.629
17	29.8	0.449	0.592	35	38.6	0.460	0.586	53	42.1	0.626	0.790
18	49.0	0.324	0.406	36	49.8	0.582	0.729	54	51.3	0.301	0.378

Table C9. Residual matrix ($e_{ij}=u_{ij}-P(\theta_{Ij})$), Items1-15: North Carolina Online Computer Skills Assessment, Fall 2005

Item	**I1**	**I2**	**I3**	**I4**	**I5**	**I6**	**I7**	**I8**	**I9**	**I10**	**I11**	**I12**	**I13**	**I14**	**I15**
1	0.220	-0.001	-0.001	-0.004	-0.001	0.002	0.000	-0.002	-0.002	0.000	-0.002	-0.002	0.000	-0.001	-0.003
2		0.130	0.001	-0.002	0.002	-0.003	-0.001	0.005	0.000	-0.004	-0.001	-0.002	0.000	0.000	-0.001
3			0.166	-0.008	-0.002	-0.001	-0.002	0.000	-0.003	-0.001	0.002	-0.002	-0.001	-0.002	-0.002
4				0.182	-0.005	-0.006	-0.006	-0.001	0.001	-0.004	-0.005	-0.003	-0.002	0.002	-0.004
5					0.228	-0.004	-0.001	-0.004	0.005	-0.002	-0.001	-0.003	-0.001	-0.002	-0.001
6						0.214	-0.002	-0.004	-0.003	0.004	0.002	0.001	0.000	-0.002	-0.003
7							0.221	-0.001	-0.002	-0.002	-0.004	0.000	-0.001	-0.002	-0.002
8								0.195	-0.002	-0.004	-0.002	-0.001	0.001	0.001	0.009
9									0.147	-0.003	-0.002	0.000	-0.001	-0.001	-0.001
10										0.177	0.004	0.003	0.000	-0.001	-0.003
11											0.218	0.000	-0.002	-0.002	-0.003
12												0.175	0.000	0.000	-0.001
13													0.172	-0.001	-0.003
14														0.083	-0.001
15															0.151
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															
26															
27															
28															
29															
30															
31															
32															
33															
34															
35															
36															
37															
38															
39															
40															
41															
42															
43															
44															
45															
46															
47															
48															
49															
50															
51															
52															
53															
54															

Note: Matrix elements on the main diagonal are variances; matrix elements in the upper triangle are residual covariances.

Table C10. Residual matrix ($e_{ij}=u_{ij}-P(\theta_{Ij})$), Items 16-30: North Carolina Online Computer Skills Assessment, Fall 2005

Item	I16	I17	I18	I19	I20	I21	I22	I23	I24	I25	I26	I27	I28	I29	I30
1	-0.003	0.000	0.003	0.002	0.001	-0.002	-0.004	0.000	-0.001	-0.003	-0.006	0.001	-0.001	-0.001	0.005
2	0.002	-0.003	-0.004	-0.004	-0.005	0.001	-0.003	-0.005	-0.002	0.000	-0.003	-0.005	-0.002	-0.005	-0.004
3	-0.003	-0.008	-0.002	-0.001	-0.002	0.046	-0.008	-0.001	-0.004	-0.004	-0.009	-0.001	-0.003	-0.002	-0.002
4	-0.004	0.020	-0.007	-0.008	-0.006	-0.007	0.059	-0.008	-0.004	-0.003	-0.011	-0.009	-0.004	-0.006	-0.006
5	0.000	-0.007	-0.004	-0.003	-0.005	-0.004	-0.003	0.002	-0.001	-0.003	-0.002	-0.004	0.000	-0.002	-0.002
6	-0.004	-0.002	0.008	0.002	0.004	-0.002	-0.006	0.003	0.000	-0.005	-0.005	0.005	-0.001	-0.001	0.003
7	-0.003	-0.005	0.000	0.002	0.000	-0.005	-0.004	-0.002	-0.003	-0.003	-0.007	-0.002	-0.001	0.002	0.002
8	0.003	-0.003	-0.002	-0.004	-0.004	-0.001	-0.006	-0.006	-0.002	0.019	-0.004	-0.006	-0.002	-0.003	-0.001
9	0.004	-0.002	-0.004	-0.004	-0.003	-0.001	0.000	-0.003	-0.002	-0.001	-0.002	-0.006	-0.002	-0.003	-0.002
10	-0.005	-0.003	0.002	0.003	0.001	-0.003	-0.006	0.005	-0.002	-0.004	-0.007	0.002	0.001	0.003	0.003
11	-0.003	-0.005	0.001	0.001	-0.001	-0.003	-0.005	0.001	0.001	-0.003	-0.008	0.003	0.000	-0.002	0.001
12	-0.001	-0.003	0.000	-0.001	0.002	-0.002	-0.003	0.001	0.001	-0.001	-0.002	0.000	0.004	0.002	-0.001
13	-0.001	0.001	-0.001	0.001	0.001	-0.001	-0.003	-0.001	-0.002	-0.002	-0.004	-0.002	-0.001	0.001	0.000
14	0.000	0.001	-0.003	-0.002	-0.001	-0.001	0.002	-0.002	-0.001	0.000	-0.005	-0.001	-0.002	-0.001	-0.002
15	0.001	-0.006	-0.005	-0.002	-0.002	-0.001	-0.004	-0.002	-0.003	0.013	-0.007	-0.004	-0.001	-0.002	-0.003
16	0.160	-0.003	-0.004	-0.004	-0.004	-0.002	-0.005	-0.005	-0.003	0.002	-0.004	-0.006	-0.002	-0.004	-0.004
17		0.163	-0.002	-0.003	-0.003	-0.005	0.023	-0.003	-0.004	-0.006	-0.014	-0.003	-0.001	-0.002	-0.004
18			0.219	0.002	0.001	-0.003	-0.005	0.002	0.001	-0.005	-0.009	0.003	-0.001	0.003	0.007
19				0.198	0.003	-0.004	-0.007	0.009	-0.001	-0.005	-0.010	0.004	0.000	0.002	0.003
20					0.179	-0.003	-0.006	0.000	0.000	-0.002	-0.005	0.002	0.000	0.002	0.000
21						0.187	-0.005	-0.002	-0.004	-0.003	-0.007	-0.006	-0.004	-0.004	-0.004
22							0.142	-0.004	-0.005	-0.006	-0.008	-0.007	-0.003	-0.004	-0.005
23								0.103	0.000	-0.006	-0.004	0.005	0.002	0.002	0.000
24									0.229	-0.002	-0.003	0.001	0.000	-0.001	0.000
25										0.198	-0.003	-0.008	-0.002	-0.005	-0.001
26											0.152	-0.010	-0.003	-0.006	-0.006
27												0.175	-0.002	0.000	0.004
28													0.167	0.003	0.000
29														0.161	0.002
30															0.229
31															
32															
33															
34															
35															
36															
37															
38															
39															
40															
41															
42															
43															
44															
45															
46															
47															
48															
49															
50															
51															
52															
53															
54															

Note: Matrix elements on the main diagonal are variances; matrix elements in the upper triangle are residual covariances.

Table C11. Residual matrix ($e_{ij}=u_{ij}-P(\theta_{Ij})$), Items 31-45: North Carolina Online Computer Skills Assessment, Fall 2005

Item	I31	I32	I33	I34	I35	I36	I37	I38	I39	I40	I41	I42	I43	I44	I45
1	0.000	-0.002	0.000	-0.003	-0.006	-0.002	-0.003	-0.004	-0.001	-0.002	0.003	0.003	-0.002	-0.002	-0.001
2	-0.003	-0.002	-0.002	-0.004	-0.004	-0.007	-0.004	-0.002	-0.007	-0.001	-0.005	-0.004	-0.005	-0.007	-0.002
3	-0.005	-0.004	-0.002	-0.009	-0.008	-0.009	-0.004	-0.002	-0.008	-0.002	-0.003	-0.002	-0.003	-0.004	-0.002
4	0.004	-0.007	-0.006	-0.012	0.034	-0.011	-0.007	-0.008	0.036	-0.002	-0.008	-0.005	-0.009	-0.009	-0.004
5	-0.005	-0.003	0.004	-0.005	-0.003	-0.007	-0.001	0.011	-0.008	-0.001	-0.005	-0.003	-0.006	-0.005	0.000
6	-0.004	-0.003	-0.003	-0.004	-0.006	-0.004	-0.004	-0.002	-0.004	-0.002	0.002	0.003	0.000	0.001	-0.003
7	-0.005	-0.004	0.001	0.000	-0.007	-0.001	0.000	-0.001	-0.007	-0.002	-0.002	0.000	-0.004	-0.001	-0.002
8	0.000	-0.004	-0.001	-0.005	-0.004	-0.006	-0.004	-0.004	-0.006	-0.002	-0.005	-0.004	-0.007	-0.008	-0.007
9	-0.003	0.000	-0.002	-0.003	0.001	-0.004	-0.004	0.000	-0.003	0.000	-0.006	-0.004	-0.003	-0.005	0.000
10	-0.003	-0.004	0.003	-0.006	-0.005	-0.007	-0.005	0.000	-0.004	-0.002	0.000	0.001	-0.004	0.000	-0.001
11	-0.002	-0.004	0.003	-0.007	-0.004	-0.007	-0.003	0.000	-0.007	-0.001	0.000	0.000	-0.002	0.001	-0.002
12	-0.003	0.000	-0.002	-0.001	-0.001	-0.002	-0.002	0.000	-0.002	-0.002	-0.001	0.000	0.000	0.002	0.000
13	0.000	-0.003	0.001	-0.003	-0.003	-0.002	-0.001	-0.002	-0.001	-0.001	-0.002	0.002	-0.001	-0.001	-0.001
14	0.007	-0.002	-0.002	-0.006	0.001	-0.004	-0.002	-0.002	0.001	0.000	-0.002	-0.001	-0.001	-0.003	-0.001
15	-0.001	-0.001	-0.004	-0.001	-0.004	-0.006	-0.005	-0.001	-0.006	-0.002	-0.002	-0.003	-0.001	-0.002	-0.003
16	-0.002	0.005	-0.003	-0.005	-0.003	-0.007	-0.006	-0.002	-0.007	-0.001	-0.006	-0.004	-0.005	-0.006	-0.002
17	0.001	-0.005	-0.005	-0.013	0.013	-0.010	-0.006	-0.007	0.032	0.000	-0.003	-0.002	-0.002	-0.004	-0.003
18	-0.005	-0.004	-0.001	-0.005	-0.009	-0.005	-0.003	-0.003	-0.003	-0.002	0.003	0.002	0.000	0.000	-0.002
19	-0.003	-0.004	0.002	-0.007	-0.007	-0.008	-0.005	-0.001	-0.006	-0.002	0.001	0.002	-0.003	-0.001	-0.003
20	-0.002	-0.002	-0.001	-0.002	-0.004	-0.004	-0.005	-0.002	-0.004	-0.004	-0.002	0.001	-0.003	0.003	-0.004
21	0.000	0.001	-0.005	-0.010	-0.008	-0.010	-0.004	-0.004	-0.006	0.000	-0.006	-0.004	-0.003	-0.006	-0.001
22	-0.001	-0.005	-0.006	-0.009	0.033	-0.008	-0.006	-0.007	0.047	-0.003	-0.008	-0.004	-0.007	-0.007	-0.002
23	-0.003	-0.002	0.000	-0.005	-0.003	-0.006	-0.007	0.002	-0.006	-0.003	0.002	0.000	-0.004	0.000	-0.003
24	-0.003	-0.003	-0.003	-0.005	-0.002	-0.003	-0.002	0.001	-0.006	-0.003	0.003	-0.002	-0.002	0.001	-0.002
25	0.000	-0.001	-0.003	-0.004	-0.003	-0.006	-0.005	-0.004	-0.007	-0.003	-0.008	-0.005	-0.004	-0.007	-0.003
26	-0.009	-0.006	-0.011	-0.014	-0.013	-0.013	-0.005	-0.006	-0.015	-0.003	-0.011	-0.007	-0.006	-0.007	-0.002
27	-0.005	-0.006	0.004	-0.008	-0.005	-0.008	-0.007	0.000	-0.004	-0.002	0.004	0.003	-0.002	0.000	-0.003
28	-0.003	-0.002	-0.001	-0.003	-0.002	-0.003	-0.003	-0.001	-0.003	-0.002	0.000	-0.003	-0.002	0.003	-0.001
29	-0.003	-0.004	-0.002	-0.003	-0.005	-0.004	-0.004	-0.002	-0.004	-0.002	0.001	0.001	-0.003	0.002	-0.003
30	-0.004	-0.006	0.003	-0.003	-0.007	-0.001	-0.001	-0.001	-0.004	-0.002	0.005	0.003	-0.001	-0.002	-0.002
31	0.182	-0.002	-0.004	-0.007	0.002	-0.009	-0.006	-0.004	0.000	-0.001	-0.003	-0.003	-0.002	-0.005	-0.002
32		0.167	-0.005	-0.005	-0.003	-0.008	-0.002	-0.003	-0.007	-0.001	-0.005	-0.004	-0.001	-0.002	0.002
33			0.207	-0.007	-0.005	-0.007	-0.004	0.002	-0.008	-0.003	-0.002	0.002	-0.003	-0.001	-0.001
34				0.141	-0.012	0.044	0.005	-0.004	-0.013	-0.003	-0.007	-0.005	-0.003	-0.005	-0.003
35					0.185	-0.012	-0.008	-0.005	0.024	-0.001	-0.006	-0.006	-0.003	-0.005	-0.003
36						0.159	0.012	-0.006	-0.011	-0.004	-0.006	-0.004	-0.005	-0.006	-0.004
37							0.111	-0.003	-0.008	-0.003	-0.006	-0.003	-0.006	-0.005	-0.002
38								0.207	-0.010	-0.001	0.001	-0.002	-0.004	0.000	-0.001
39									0.187	-0.001	-0.006	-0.004	-0.003	-0.005	-0.003
40										0.127	-0.002	-0.002	0.000	-0.003	0.000
41											0.191	0.004	-0.001	0.002	-0.004
42												0.225	0.000	0.001	-0.002
43													0.107	0.000	-0.001
44														0.176	-0.004
45															0.091
46															
47															
48															
49															
50															
51															
52															
53															
54															

Note: Matrix elements on the main diagonal are variances; matrix elements in the upper triangle are residual covariances.

Table C12. Residual matrix ($e_{ij}=u_{ij}-P(\theta_{ij})$), Items 46-54: North Carolina Online Computer Skills Assessment, Fall 2005

Item	I46	I47	I48	I49	I50	I51	I52	I53	I54
1	-0.004	-0.006	0.001	0.001	0.001	0.000	-0.002	-0.007	-0.001
2	-0.004	-0.004	-0.001	-0.002	-0.003	-0.003	-0.002	-0.007	-0.003
3	-0.005	-0.011	-0.001	-0.003	-0.002	-0.005	-0.004	-0.011	-0.003
4	-0.010	-0.014	-0.006	-0.008	-0.010	0.009	-0.010	-0.016	-0.006
5	0.012	-0.005	-0.002	-0.002	0.001	-0.005	-0.002	-0.005	-0.003
6	-0.004	-0.005	-0.001	0.002	0.000	-0.002	-0.005	-0.006	0.004
7	-0.002	-0.007	0.001	0.000	-0.002	-0.004	-0.003	-0.009	-0.002
8	-0.005	-0.006	0.001	-0.004	-0.005	-0.003	-0.002	-0.007	-0.005
9	0.002	-0.004	-0.002	-0.002	-0.003	-0.003	-0.002	-0.005	-0.003
10	0.000	-0.008	-0.001	0.000	0.000	-0.002	-0.006	-0.010	-0.001
11	0.001	-0.008	-0.001	-0.002	-0.001	-0.003	-0.004	-0.008	-0.004
12	0.000	-0.003	-0.005	0.000	-0.001	-0.002	-0.003	-0.003	-0.001
13	-0.001	-0.005	-0.002	0.001	0.000	0.000	-0.002	-0.004	-0.001
14	-0.003	-0.007	-0.001	-0.001	-0.003	0.004	-0.003	-0.006	-0.002
15	-0.002	-0.008	-0.003	-0.004	-0.004	-0.004	-0.004	-0.007	-0.002
16	-0.006	-0.005	-0.002	-0.004	-0.005	-0.003	-0.005	-0.008	-0.003
17	-0.011	-0.015	-0.003	-0.002	-0.005	0.007	-0.009	-0.015	-0.003
18	-0.005	-0.008	-0.001	0.005	0.000	-0.004	-0.006	-0.009	0.003
19	-0.005	-0.009	0.001	0.002	0.003	-0.003	-0.005	-0.009	0.002
20	-0.004	-0.006	0.000	0.000	-0.001	-0.002	-0.007	-0.006	0.001
21	-0.007	-0.008	-0.004	-0.002	-0.004	-0.003	-0.004	-0.011	-0.004
22	-0.008	-0.011	-0.005	-0.006	-0.009	0.011	-0.010	-0.012	-0.005
23	0.001	-0.005	-0.002	0.002	0.001	-0.002	-0.005	-0.005	0.001
24	-0.001	-0.005	0.002	-0.002	-0.002	-0.002	-0.002	-0.004	0.003
25	-0.004	-0.006	-0.001	-0.006	-0.007	-0.004	-0.003	-0.008	-0.006
26	-0.001	0.027	-0.009	-0.007	-0.008	-0.009	-0.007	0.038	-0.008
27	-0.005	-0.010	0.002	0.000	0.003	-0.001	-0.009	-0.011	0.001
28	0.000	-0.004	-0.002	0.003	0.003	-0.002	-0.002	-0.005	-0.002
29	-0.003	-0.005	-0.002	0.001	0.002	-0.003	-0.004	-0.006	0.000
30	0.000	-0.007	0.000	0.002	0.000	-0.003	-0.003	-0.007	0.001
31	-0.005	-0.011	-0.002	-0.002	-0.005	0.000	-0.006	-0.011	-0.003
32	-0.007	-0.009	-0.004	-0.003	-0.006	-0.004	-0.002	-0.009	-0.002
33	0.003	-0.011	0.001	-0.001	0.006	-0.004	-0.006	-0.011	-0.001
34	-0.007	-0.014	-0.006	-0.004	-0.008	-0.008	-0.002	-0.016	-0.006
35	-0.008	-0.015	-0.006	-0.007	-0.007	0.007	-0.009	-0.014	-0.005
36	-0.010	-0.013	-0.003	-0.004	-0.010	-0.006	0.002	-0.017	-0.005
37	-0.004	-0.004	-0.002	-0.004	-0.006	-0.003	0.006	-0.006	-0.003
38	0.007	-0.008	-0.002	-0.003	0.002	-0.004	-0.002	-0.008	-0.001
39	-0.012	-0.014	-0.005	-0.004	-0.008	0.010	-0.012	-0.015	-0.003
40	-0.003	-0.003	-0.002	0.000	0.000	-0.001	-0.002	-0.005	-0.002
41	-0.006	-0.010	0.001	0.000	-0.001	-0.002	-0.004	-0.011	0.003
42	-0.005	-0.006	0.001	0.002	0.002	-0.002	-0.005	-0.006	0.001
43	-0.005	-0.004	-0.001	-0.001	-0.004	0.000	-0.004	-0.006	0.001
44	0.001	-0.005	-0.006	0.001	0.001	-0.002	-0.005	-0.006	0.001
45	-0.001	-0.002	-0.001	-0.001	-0.002	-0.001	-0.002	-0.003	-0.002
46	0.188	-0.005	-0.004	-0.003	0.000	-0.006	-0.005	-0.003	-0.007
47		0.134	-0.009	-0.004	-0.008	-0.009	-0.008	0.044	-0.006
48			0.194	-0.005	-0.003	-0.002	-0.004	-0.006	0.003
49				0.208	0.004	-0.002	-0.002	-0.005	0.003
50					0.185	-0.004	-0.005	-0.008	0.002
51						0.116	-0.006	-0.008	-0.001
52							0.187	-0.007	-0.003
53								0.126	-0.006
54									0.227

Note: Matrix elements on the main diagonal are variances; matrix elements in the upper triangle are residual covariances.

Table C13. Summary of Residual Variance Covariance Matrix

Item	Summary statistics by item: Median Covariance	Minimum Residual Covariance	Maximum Residual Covariance
1	-0.001	-0.007	0.005
2	-0.003	-0.007	0.005
3	-0.002	-0.011	0.046
4	-0.006	-0.016	0.059
5	-0.002	-0.008	0.012
6	-0.002	-0.006	0.008
7	-0.002	-0.009	0.002
8	-0.004	-0.008	0.019
9	-0.002	-0.006	0.005
10	-0.002	-0.010	0.005
11	-0.002	-0.008	0.004
12	-0.001	-0.005	0.004
13	-0.001	-0.005	0.002
14	-0.001	-0.007	0.007
15	-0.003	-0.008	0.013
16	-0.003	-0.008	0.005
17	-0.003	-0.015	0.032
18	-0.002	-0.009	0.008
19	-0.002	-0.010	0.009
20	-0.002	-0.007	0.004
21	-0.004	-0.011	0.046
22	-0.005	-0.012	0.059
23	-0.002	-0.008	0.009
24	-0.002	-0.006	0.003
25	-0.003	-0.008	0.019
26	-0.007	-0.015	0.038
27	-0.002	-0.011	0.005
28	-0.002	-0.005	0.004
29	-0.002	-0.006	0.003
30	-0.001	-0.007	0.007
31	-0.003	-0.011	0.007
32	-0.003	-0.009	0.005
33	-0.002	-0.011	0.006
34	-0.005	-0.016	0.044
35	-0.005	-0.015	0.034
36	-0.006	-0.017	0.044
37	-0.004	-0.008	0.012
38	-0.002	-0.010	0.011
39	-0.005	-0.015	0.047
40	-0.002	-0.005	0.000
41	-0.002	-0.011	0.005
42	-0.002	-0.007	0.004
43	-0.003	-0.009	0.001
44	-0.002	-0.009	0.003
45	-0.002	-0.007	0.002
46	-0.004	-0.012	0.012
47	-0.007	-0.015	0.044
48	-0.002	-0.009	0.003
49	-0.002	-0.008	0.005
50	-0.002	-0.010	0.006
51	-0.003	-0.009	0.011
52	-0.004	-0.012	0.006
53	-0.007	-0.017	0.044
54	-0.002	-0.008	0.004