Journal of Educational Measurement
Fall 2009, Vol. 46, No. 3, pp. 247–272

Conceptual Issues in Response-Time Modeling

Wim J. van der Linden
CTB/McGraw-Hill

Two different traditions of response-time (RT) modeling are reviewed: the tradition of distinct models for RTs and responses, and the tradition of model integration in which RTs are incorporated in response models or the other way around. Several conceptual issues underlying both traditions are made explicit and analyzed for their consequences. We then propose a hierarchical modeling framework consistent with the first tradition but with the integration of their parameter structures as a second level of modeling. Two examples of the framework are presented. Also, a fundamental equation is derived which relates the RTs on test items to the speed of the test taker and the time intensity of the items. The equation serves as the core of the RT model in the framework. Finally, empirical applications of the framework demonstrating its practical value are reviewed.

Test theorists have always been intrigued by the relationship between responses to test items and the time used by a test taker to produce them. Both seem indicative of the same behavior on test items. Nevertheless, their relationship appears to be difficult to conceptualize, let alone represent coherently in a statistical model.

Although the computerization of educational tests has been a major impetus to the current interest in response-time (RT) modeling, it would be wrong to ignore its historical origins. One early development that has left traces in our current thinking about RTs was Woodbury's (1951, 1963) treatment of test scores as the result of a time-dependent stochastic response process. His theory, which is summarized in Lord and Novick (1968, chap. 5), has linear axioms and theorems that are entirely parallel to those of regular classical test theory. But, more important to the scope of this article, it also lent statistical sophistication to the intuitive idea that total time and numbers of items completed are equivalent measures of the test taker's performance (see the example in Lord & Novick, pp. 104–105).

The same idea was present in Gulliksen's (1950, chap. 17) treatment of speed and power tests. He defined a pure speed test as a test with an unlimited number of items that are easy enough to be answered correctly. Such tests can be scored in two different ways, as (a) the total time used to complete a fixed number of items, and (b) the number of items completed in a fixed time interval. On the other hand, a pure power test was defined by him as a test with unlimited time but a fixed number of items of varying difficulty. Such tests can be scored only by counting the number of correct responses.

A fundamental problem exists with respect to the asymmetry between Gulliksen's scoring rules for speed and power tests. At a practical level, the problem becomes manifest when a test taker produces an incorrect answer on a speed test, which is not very likely for high-ability test takers and easy items but certainly possible. How should we treat such responses? And would it be fair to treat their RTs as equivalent

to those for correct responses? Similarly, it seems odd to ignore the time spent on items in power tests. If two test takers have the same number of items correct but one took much less time than the other, why should we label their performances as equal? As discussed later, at a more technical level the problem is due to the neglect of an important random variable required to describe test behavior as a stochastic process.

One of the very first to address the relation between responses and RTs from a perspective now known as item response theory (IRT) was Thurstone (1937). His main intention was to analyze the notions of ability and speed, which he considered as the core of educational testing. His analysis was based on the idea of a response surface for a fixed person and item, which describes the probability of a correct response to the item as a function of its difficulty and the time for the response. (It is not quite clear from the article whether this is the time allowed or actually taken.)

Thurstone's graphical example of a response surface is reproduced in Figure 1. (Observe that this surface is for a fixed person over a range of possible difficulties of the item; it is thus a generalized person response function, not a generalization of the now more popular item response function.) The main features of the surface are (a) a decrease of the probability of success with the difficulty of the item but (b) an increase of the probability with time. Thurstone then defines speed as "the number of tasks completed in unit time" (p. 250) and the ability of an individual subject as "the difficulty . . . at which the probability is 1/2 that he will do the task in infinite time" (p. 251).

These definitions are in a similar vein as Gulliksen's later treatment of speed and power tests and Woodbury's process model of testing. In fact, Thurstone's definition of ability as a limit of the probability of success with time going to infinity was already much more sophisticated than Gulliksen's unlimited time as a necessary condition for a pure power test. Also, Thurstone's approach seems to do better justice to the hybrid nature of test items; he treats them as tasks that always have a speed as well as a power aspect. Finally, Thurstone already defined important concepts as difficulty, speed, ability, and time at the level of the combination of a fixed person and item. If he had also presented a parametric model for the response surface, it could already have played the same role in earlier item analysis and test design as some of the later IRT models.

The response surface in Figure 1 is, however, based on several tacit assumptions that require explicit reflection. First, it is asymmetric in that it represents the probability of a response but a direct observation of RT. This is at odds with the fact that both seem to be indicative of the same cognitive process in the test taker. If this process should be considered as stochastic, both must be treated as random variables and the response surface should be for their joint probability rather than the response probability only. In addition, Thurstone's treatment has a second asymmetry in that it explains the response probabilities by an item parameter (difficulty) and a person parameter (ability) but leaves the RTs unexplained. Should this be taken to imply that RTs are independent of the features of the item? And that systematic differences in RT between test takers are impossible? Third, the shape of the response surface allows for a tradeoff between time and item difficulty: The probability

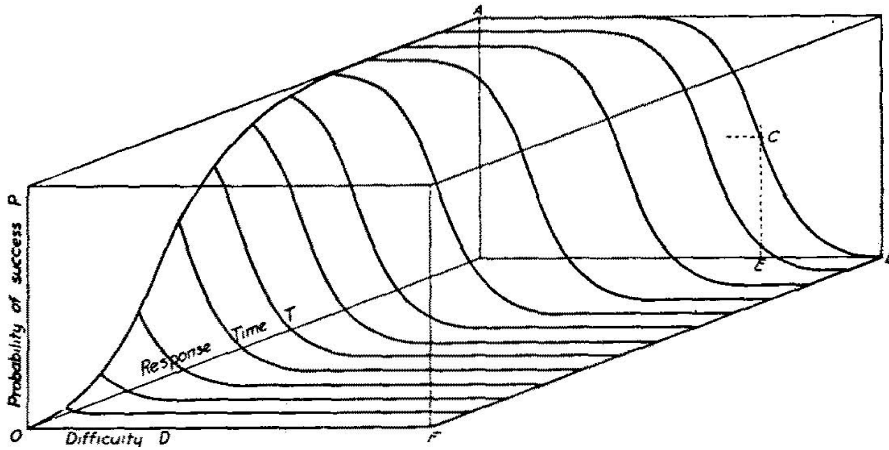

FIGURE 1. *Graphical example of Thurstone's response surface. (Reproduced with permission from L. L. Thurstone (1937). Ability, motivation, and speed. Psychometrika, 2, 249–254.)*

of success on a more difficult item can always be compensated by spending more time on it. Although results from experimental research of reaction times in psychology show ample evidence of speed-accuracy tradeoffs, they are within-person phenomena whereas Figure 1 displays the probability of success as a function of the difficulty of the item. Finally, the response surface also assumes a dependency of the response probabilities on the RT. Although responses and RTs may correlate across test takers, a standard assumption in IRT is conditional independence between the responses to different items by the same test taker ("local independence"). Should the same hold for the responses and RTs? Or should the fact that both seem to represent different sides of the same behavior be taken to imply inherent dependence?

These questions illustrate the fact that to develop appropriate statistical models for RTs, first several conceptual issues have to be resolved. It is the goal of this article to make these issues explicit. We will do so by analyzing a selection of RT models that are typical of the different conceptual choices that have been made in the literature. We then extract a few basic conclusions from the analysis and propose an alternative modeling framework for responses and RTs on test items consistent with these conclusions.

Different Types of Modeling

Two different traditions of RT modeling are reviewed: (a) the tradition of distinct models for RTs and responses, and (b) the tradition of model integration in which RTs are incorporated in response models or the other way around. Our review is not exhaustive; from both traditions only a few typical models are selected. For a more complete review of the history of RT modeling, see Schnipke and Scrams (2002).

The prime example of this tradition of modeling is Rasch's (1960) treatment of oral reading tests, which involves two different types of models—one for the number of misreadings in a text, the other for the speed of reading. Both models are derived from the assumption of reading as a stochastic process; that is, a process in which a reader produces words that take random time. More specifically, Rasch assumed the process to be Poisson.

Generally, a Poisson process for an event arises when the probability θ of its occurrence in an arbitrary small time interval is constant across time (e.g., Casella & Berger, 1990). The assumption implies that at any point of time the probability is independent of the earlier history of the process.

Rasch's model for misreadings. In this model, the focus is on the occurrence of misreadings. For text that is homogenous and readers with a constant probability of making a reading error, the assumption of a Poisson process is plausible. It is a standard statistical result that the number of misreadings a in a text of N words then follows a Poisson distribution with probability function

$$\Pr(a \mid N) = e^{-\lambda}\frac{\lambda^a}{a!}, \tag{1}$$

where $\lambda = N\theta$ is the expected number of misreadings.

Rasch decomposed the probability θ into parameters for the reader and the text. Let j be an arbitrary reader and i an arbitrary text. Rasch's proposal was

$$\theta_{ij} = \frac{\delta_i}{\xi_j}, \tag{2}$$

where δ_i was interpreted by him as the difficulty of text i and $1/\xi_j$ as the ability of the reader j. This simple relation reflects the fact that a more difficult text or a less able reader should have a greater probability of a misreading.

Rasch's model for reading speed. This model for reading speed assumes the same type of Poisson process but this time for the event of completing a unit of text while reading. The process is realized if the text is homogeneous and the test taker reads at a constant speed. It then holds that the number of words N read in a given time T by a reader who reads at speed λ follows the same type of Poisson distribution as in (1).

However, the intended model is not for number of words read in a given time but, reversely, for the time required to finish the reading of a given number of words. The two events are related by a simple probability relation, which states that the probability that reading a words in a given time T exceeds N is equal to that of the time t needed to read N words not exceeding T. Formally,

$$\Pr(a \geq N \mid T) = \Pr(t \leq T \mid N) \tag{3}$$

(Rasch, 1960, p. 38).

Again, it is a standard result in statistics that if the probability on the left is Poisson, the probability on the right follows a gamma distribution with density

$$p(t \mid N) = \lambda e^{-\lambda t} \frac{(\lambda t)^{N-1}}{(N-1)!}, \tag{4}$$

where λ is an "intensity parameter" that characterizes the person's reading speed (p. 38). Rasch also refers to this parameter more directly as a "speed parameter" (p. 41). Formally, it is the expected number of words read in a given time unit (p. 41).

Rasch also suggested decomposing speed parameter λ into separate parameters for the reader and the text:

$$\lambda_{ij} = \frac{\xi_j}{\delta_i}, \tag{5}$$

where δ_i is the difficulty of text i and ξ_j is the ability of person j.

Discussion. The probability law in (3) must have motivated Gulliksen's two scoring rules for speed tests: one of his rules refers to the event in the left-hand side of (3), the other to the event in the right-hand side. The law shows that the two events can be treated as equivalent provided both are the result of the same underlying process.

However, the relationships between the different person and text parameters in the core equations in (2) and (5) are unclear. The notation for these parameters used by Rasch, which we have maintained in our presentation of the models, seems to suggest that they refer to the same empirical variables. In addition, the equations define the parameters as reciprocals of each other—which is in agreement with the reversal of the events in (3) and thus seem to point at the assumption of the same parameters for a single process indeed. Also, Rasch's use of the term "ability" for ξ_j in (2) and (5) suggests an identical interpretation. Nevertheless, the following quote reveals that he might have assumed an analogy only and left the actual relationships between these parameters to further research:

> This analogy between the models for misreadings and reading speed does, however, not indicate that the two pairs of concepts are identical. It seems reasonable that a text which gives rise to many mistakes—e.g., because it contains many unknown words or deals with a little known subject-matter—will also be rather slowly read. But presumably it is not a general rule that a slow reader also makes many mistakes. To which extent the two difficulty parameters for each text and the two ability parameters for each person run parallel is a question to be answered by empirical research, and at present we shall leave it open. (Rasch, 1960, p. 42)

It is not exactly clear what is meant by the phrase "parameters. . .[that] run parallel" in this quote. Later in this article, we will take it to imply a positive correlation between parameters across different persons or items.

Others have followed the tradition begun by Rasch and modeled the responses and RTs distinctly. Jansen (1986, 1997a, 1997b), Jansen and van Duijn (1992),

Oosterloo (1975), and Scheiblechner (1979) used the same models as Rasch in various studies, improved their statistical treatment, and provided such extensions as the incorporation of manifest covariates or structural parameters for the person parameter in the models. Pieters and van der Ven (1982) used the same gamma model to decompose RTs into problem solving and distraction times, whereas Maris (1993) assessed its consistency with different stage models of problem solving from psychology. Distinct RT models from other families of statistical distributions are the Weibull model by Tatsuoka and Tatsuoka (1980) and the lognormal model by van der Linden (2006).

Response Models That Incorporate RT

This alternative type of modeling follows the Thurstonian (1937) tradition and incorporates RTs (or their parameters) in response models. The result is a single model for responses and RTs rather than two distinct models. The models are interesting because they can be viewed as attempts to answer Rasch's question about the relationship between the two types of parameters in the quote above.

Roskam's model. One of the first attempts to build RTs in response models was Roskam's (1987; see also Roskam, 1997). His model is the regular Rasch or one-parameter logistic (1PL) response model with the ability parameter replaced by an "effective ability parameter" defined as "mental speed times processing time." On an exponential scale, the product of mental speed and time is the sum $\theta_j + \ln t_{ij}$. (Roskam uses the traditional notation for the ability parameter also for the speed parameter; only his interpretation changes.) Thus, the model can be written as

$$p_i(\theta_j) = \{1 + \exp[-(\theta_j + \ln t_{ij} - b_i)]\}^{-1}. \tag{6}$$

Observe that the presence of the difference $\ln t_{ij} - b_i$ in the model is in agreement with Thurstone's response surface in Figure 1: An increase in the difficulty of the item can always be compensated by spending more time on it. Also, the model seems to capture a speed-accuracy tradeoff in that an increase in time implies an increase in probability of success on the item. In fact, this tradeoff—more precisely known as an increasing conditional accuracy function (Luce, 1986, sect. 6.5)—was the main motivation for this model (Roskam, 1997, pp. 188–190).

Verhelst, Verstralen, and Jansen (1997) have presented a model identical to (6) but with t_{ij} replaced by a parameter τ_j, which they interpret as a speed parameter for person j:

$$p_i(\theta_j) = \{1 + \exp[-(\theta_j + \tau_j - b_i)]\}^{-\pi_i} \tag{7}$$

The model also has a shape parameter π_i, which is further ignored here. Interestingly, these authors derived their model from a different set of assumptions than Roskam's, namely, that of a combination of a generalized extreme-value distribution for a latent response variable conditional on the time spent on the item and a gamma distribution for the marginal distribution of the time.

Wang & Hanson's model. Wang and Hanson (2005) offer a model that incorporates RTs in the 3PL response model instead of the 1PL model. The result is the response function

$$p_i(\theta_j) = c_i + (1 - c_i)\{1 + \exp[-a_i(\theta_j - \rho_j d_i / t_{ij} - b_i)]\}^{-1}, \tag{8}$$

where a_i and c_i are the usual discrimination and guessing parameters for item i. Except for the more general parameter structure of the 3PL model, the main difference between this and Roskam's model is the replacement of $\ln\ t_{ij}$ by

$$-\rho_j d_i / t_{ij}. \tag{9}$$

In addition to the time t_{ij}, this expression contains parameters ρ_j and d_i referred to as "slowness parameters" for the person and the item by the authors. Their name is motivated by the fact that less time on an item has the same effect on the probability of success as an increase of either of these parameters.

Further, the model shows a structural difference with respect to the speed-accuracy tradeoff in Roskam's model: with increasing time, the probability of success in (8) approaches that of the regular 3PL model whereas for (6) it goes to one. According to the authors, the difference qualifies Roskam's model as a model for speed tests but theirs as one for hybrid tests (Wang & Hanson, 2005, p. 336).

RT Models That Incorporate Responses

The reverse type of modeling incorporates responses (or response parameters) into a model for the distribution of the RT on an item. Models of this type entail different conditional RT distributions given a correct and an incorrect response. This feature seems to be supported by several empirical studies that found substantial differences between average RTs for correct and incorrect responses across test takers. A discussion of these findings follows later in this article.

Gaviria's model. A recent example of this type of modeling is a model by Gaviria (2005). For $u_{ij} = 1$, this author posits the equation

$$\ln\left(\frac{t_{ij} - T_0}{A}\right) = -a_i(\theta_j - b_i) + \varepsilon_{ij}, \tag{10}$$

with A a scaling constant, T_0 the time taken by the person on an infinitely easy item, and $a_i(\theta_j - b_i)$ the usual parameter structure from the 2PL response model. For the residual, a lognormal distribution is chosen:

$$\varepsilon_{ij} \sim LN\left(0, \sigma_i^2\right). \tag{11}$$

The model thus specifies a double lognormal distribution for a rescaled time, $(t_{ij} - T_0)/A$, with mean $-a_i(\theta_j - b_i)$ and item-dependent variance σ_i^2. The scaling constants T_0 and A are to be estimated from testing data.

The representation in (10)–(11) is not Gaviria's; it has been rewritten here to show an analogy with Thissen's model in the next section. The original representation (Gaviria, 2005, eq. 2) yields a tautology for the substitution of $u_{ij} = 0$, which means that the conditional distribution of the RT given an incorrect response is left unspecified. This omission reveals a basic problem for attempts to incorporate responses into RT models: It is generally difficult to specify conditional RT distributions that offer a plausible explanation of the difference between the impact of correct and incorrect responses on RTs.

Thissen's model. Thissen's (1983) well-known model belongs to the same category of RT modeling. But rather than specifying different RT distributions for correct and incorrect responses, it regresses the RT directly on the usual parameter structure for response models. In addition, it introduces parameters for an extra person and item effect on the RT. Formally, the model is

$$\ln T_{ij} = \mu + \tau_j + \beta_i - \rho(a_i\theta_j - b_i) + \varepsilon_{ij}, \tag{12}$$

with

$$\varepsilon_{ij} \sim N(0, \sigma^2). \tag{13}$$

Parameter μ in the model is a general level parameter for the population of persons and domain of test items. Parameters τ_j and β_i are interpreted as the "slowness parameters" for the person and item by Thissen (1983, p. 181), whereas ρ is a slope parameter in the regression of the logRT on the response parameter structure $a_i\theta_j - b_i$. The combination of the logarithmic transformation of the RT with a normal distribution for ε_{ij} yields a lognormal model for the RT.

Ferrando and Lorenzo-Seva (2007) present a version of the model in (12)–(13) for items in personality tests with the regression of the RT on

$$\sqrt{a_i^2(\theta_j - b_i)^2}. \tag{14}$$

instead of $a_i\theta_j - b_i$; otherwise, their model is identical. The choice of (14) is motivated by a distance-difficulty hypotheses, which has gained some support in the field of personality measurement and predicts that the RT on a personality item increases with the distance between the test taker's trait level θ_j and the difficulty of the item b_i.

Both models can be specified to represent a tradeoff between speed and accuracy. For $\rho < 0$, (12) implies a tendency for t_{ij} to increase with an increase of θ_j whereas for (14) the same tendency exists for an increase in the distance between θ_j and b_i. On the other hand, for $\rho > 0$ the relation in (12) reverses. In addition, the models represent an interaction between slowness τ_j and accuracy θ_j on t_{ij} that is more difficult to interpret.

Basic Issues

We now discuss some of the basic issues that have emerged in our review of these different types of modeling. Our conclusions from these issues suggest an alternative type of modeling which is presented later.

Fixed or Random RTs

It is a common experience in reaction-time experiments that the times taken by subjects on replications of simple tasks vary randomly from one observation to another. The same randomness can be assumed to hold for RTs on test items. In fact, we even expect a greater role of randomness: Test items are much more complex and involve more uncertainty than the typical tasks in these psychological experiments. Besides, it seems inconsistent to assume random responses—a basic assumption of IRT—but fixed RTs. Although the assumption of randomness seems obvious, it is generally difficult to verify empirically through replicated administrations of the same item to the same person because of learning and memory effects.

Some of the previous models do treat RTs as random variables; others, for instance those in (6) and (8), treat them as fixed values in a response model. The only way to reconcile the random nature of RTs with the presence of fixed values t_{ij} in a response model is to view the latter as specifications of the conditional distribution of U_{ij} given $T_{ij} = t_{ij}$. However, a full model for T_{ij} and U_{ij} would be for their joint distribution; the part that is missing is thus a model for the marginal distribution of T_{ij}.

Alternatively, the fixed values t_{ij} could be replaced by RT parameters, as in the model in (7). However, as will be argued later, RT parameters make sense in models for a RT distribution but not in models for the distribution of responses.

Conclusion 1: RTs on test items should be treated as realizations of random variables T_{ij}.

Item Completion, Responses, and RTs

Rasch's (1960) models of misreadings and reading speed were derived from an underlying Poisson process. The notation used by Rasch, as well as the reversal of the probabilities of total time and the number of responses in (3), seems to point at a single process. But his earlier quote, in which he entertains the possibility of different empirical behavior of the ability and difficulty parameters in the two models, contradicts the assumption. On the other hand, in the earlier literature about speed tests, the idea that total time and the number of responses are equivalent measures was ingrained deeply. Examples discussed earlier are Gulliksen's claim of equivalence of his two rules for the scoring of speed tests and Woodbury's theory of time-dependent test scores.

Where does this ambiguity come from? The answer lies in the neglect of an extra random variable required to describe the response behavior as a stochastic process. In all, three different types of random variables are necessary: one variable for the response time on the items (T_{ij}) and two variables for response-related events (U_{ij} and D_{ij}). Formally, the latter can be defined as

$$U_{ij} = \begin{cases} 1, & j \text{ answers item } i \text{ correctly,} \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

and

$$D_{ij} = \begin{cases} 1, & j \text{ completes item } i, \\ 0, & \text{otherwise}, \end{cases} \tag{16}$$

respectively. Variables U_{ij} are the regular response variables modeled in IRT; variables D_{ij} are design or indicator variables that show which items the test taker completes.

The distinction between response and design variables was not common in the early history of test theory; it was introduced only when problems with missing data became manifest, such as in large-scale educational assessments and adaptive testing. It is important to distinguish between (15) and (16) because they have different probability distributions. One obvious difference is that the variables U_{ij} can be assumed to be independent for a given test taker j whereas the variables D_{ij} are always dependent (e.g., $D_{ij} = 1$ has a nonzero probability only if j has completed the preceding item).

Using these variables, two different total scores can be defined

$$N_j = \sum_i D_{ij} \tag{17}$$

and

$$N_j = \sum_i U_{ij}. \tag{18}$$

The first total score tells us how far the test taker has proceeded through the test; the second is the regular number-correct score. The two total scores have different probability distributions because their constituent variables have. (Actually, as explained later, it even makes sense to think of them as *independent* variables.)

In Rasch's (1960) description of his models, because of a common notation, the two total scores are easily confounded. But the reversal of probabilities for N and T in (3) that leads to the gamma model for reading speed holds only for total score D_j. On the other hand, the model of misreadings is for number-correct score N_j. Thus, although each model has an underlying Poisson process, the processes are *not* identical: the process for the misreadings generates values for the response variables U_{ij} but that for the speed of reading generates values for the design variables D_{ij}.

One condition exists for which the distinction between response and design variables can be ignored. When the probability distribution of U_{ij} degenerates to $Pr(U_{ij} = 1) = 1$, the two variables become exchangeable. More formally,

$$Pr(U_{ij} = 1) = 1 \longrightarrow D_{ij} = U_{ij}. \tag{19}$$

Under this condition, it holds that $D_j = N_j$. Consequently, we can use the total score N_j (given a fixed time interval) and the total time T_j (given a fixed value of D_j) as equivalent measures of speed.

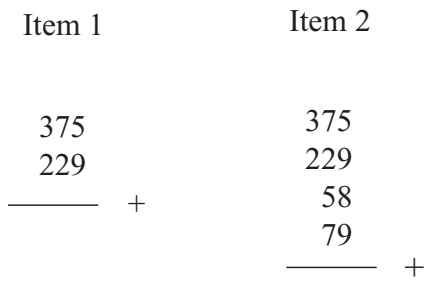

FIGURE 2. *Example of two arithmetic items differing in time intensity.*

The condition in (19) is implied by Gulliksen's definition of a pure speed test as a test with items easy enough to be answered correctly. However, for all response models currently in use for educational testing, the condition can only be met for a test taker of infinitely high ability and/or items of infinitely low difficulty. The equivalence of Gulliksen's two rules for scoring speed tests is thus an ideal that is never met in practice. Similarly, the notion of a pure power test seems superfluous. As opposed to (19), there exists no special condition on the distributions of the three variables that permits us to ignore T_{ij} (or D_{ij}) and focus on U_{ij} only. Therefore, it seems more realistic to treat every real-world test as a hybrid test with both a speed and a power aspect captured by the variables T_{ij} (or D_{ij}) and U_{ij}, respectively.

Conclusion 2: For *any* type of test, RTs, item completions, and responses should be treated as realizations of distinct random variables T_{ij}, D_{ij}, and U_{ij}. The probability distributions of T_{ij} and D_{ij} are different from the distribution of U_{ij} but related through the inverse relation in (3). As a consequence, except for continuity, and provided the other is fixed, the total time T_j spent on the test and the sum of design variables D_j are measures with the same information, but T_j and N_j are not.

RT and Speed

It is not unusual to find the notions of RT and speed treated as equivalent in the psychometric literature on RT models. The idea is also common in reaction-time research in psychology, where speed is invariably measured as the average reaction time on a task; for example, speed-accuracy tradeoffs are usually presented by a plot of the average correct response as a function of the average reaction time in this research (see the schematic of the general tradeoff relation in Luce, 1986, Fig. 6.13). The presence of t_{ij} in (6) was motivated by the idea to build this type of speed-accuracy tradeoff in a RT model (Roskam, 1997, pp. 188–190).

However, a moment's reflection shows that RT and speed are *not* equivalent. A simple counterexample is the case of the two arithmetic items in Figure 2. Suppose one examinee answers item 1 in 9 seconds whereas another takes 13 seconds for item 2. It would be wrong to conclude that the second examinee worked slower than the first: item 2 involves a longer series of operations than item 1 and, in spite of

a longer RT, the second examinee might have worked much faster through it. The example thus shows that it is generally impossible to relate RT to speed unless we have a measure of the amount of labor required by the items.

A fundamental equation. Notions of speed are found not only in test theory but have permeated every area of science. It is quite common, for instance, to speak of the speed of inflation in economics, the speed at which a rumor spreads in sociology, or the speed of recovery in medicine. The definitions of all of these notions share the format of a *rate of change of some measure with respect to time*.

The prototypical definition of speed is that of speed of motion in physics. Let $d(t)$ be the distance traveled from a given point of reference as a function of time and t_1 and t_2 the two end points of a time interval. The textbook definition of the average speed in the interval is

$$\text{average speed} = \frac{d(t_2) - d(t_1)}{t_2 - t_1}. \tag{20}$$

Any other definition of speed has the same format but with a different measure in the numerator.

The appropriate notion of speed on test items is that of speed of labor. Hence, its definition is that in (20) with the numerator replaced by a measure of the amount of labor required by the items. Let β_i^* denote the (unknown) amount of labor required to solve item i. As the clock is reset at the beginning of every item, the point of reference in (20) becomes $t_1 = 0$ and response time t_{ij} can be taken as its denominator. The (average) speed τ_j^* of test taker j on item i, is then defined as

$$\tau_j^* = \frac{\beta_i^*}{t_{ij}}. \tag{21}$$

Alternatively, we can write

$$t_{ij} = \frac{\beta_i^*}{\tau_j^*}, \tag{22}$$

which shows that the definition of speed involves the decomposition of the RT into two unknown parameters: one parameter for the speed of the person and the other for the labor intensity of the item. As parameter β_i^* represents an effect on time, we will mainly refer to it as the time-intensity parameter for item i.

RTs are bounded from below and it is always possible to spend more time on an item. Hence, their distributions tend to be positively skewed. A standard transformation to get a more symmetric distribution is the logarithmic, which gives

$$\ln t_{ij} = \beta_i - \tau_j, \tag{23}$$

where β_i and τ_j are now parameters on a logarithmic scale. Finally, RT is a random variable but the right-hand side of the equation is fixed. Hence, (23) should be

conceived of as an equation for the expected logtime on the item,

$$\mathcal{E}(\ln t_{ij}) = \beta_i - \tau_j. \tag{24}$$

We will therefore refer to the equation as the *fundamental equation of RT modeling*. Each RT model with a person parameter that is interpreted as speed should be based on this equation and have a time-intensity parameter for the items as well. Indeed, the equation plays exactly this role in some of the models reviewed above. It also helps to clarify some of the interpretations associated with the other models.

For example, the model by Thissen in (12) has the equation as its core (with speed parameter τ_j replaced by its negative as a slowness parameter). The same holds for the model by Ferrando and Lorenzo-Seva in (14). On the other hand, these models also have "slowness" parameters for the items whereas the equation above suggests that an interpretation of these parameters as parameters for the time intensity of the items would be more consistent with the notion of speed.

Exactly the same holds for the interpretation of parameter d_i in the model in (8) by Wang and Hanson. More importantly, this model has a formal problem because of (9), which is not an equation relating RT to item and person parameters but just an expression based on these quantities. It is hard to find a satisfactory interpretation for the effect of this expression on the response probability in (8).

The differences between (22) and the core equation in the Rasch model for reading speed in (5) are subtle. Parameter λ_{ij} in the model is equal to the expected number of words read in a given time unit; see directly below (23). Hence, its reciprocal is the expected time on a unit of text, and Rasch's equation in (5) is thus formally identical to the fundamental equation. However, the two have important interpretative differences. Unlike Rasch's interpretation of λ_{ij} as a speed parameter, it is just the expected time by the reader on the text, i.e., the left-hand side of (22); ξ_j is not an ability parameter but a parameter for the speed of the reader; and δ_i is not the difficulty of a text but the amount of labor it involves, just as "ability" is a term more appropriate for the description of the distribution of reading errors.

Conclusion 3: Time and speed are different concepts but are related through the equation in (22). RT models with speed as a person parameter should also have an item parameter for their time intensity.

Speed and Ability

The idea that speed and ability have a tradeoff relation, which motivated several of the models reviewed earlier, certainly makes sense. In reaction-time research in psychology, a speed-accuracy tradeoff is typically presented as a positive relation between the proportion of correct tasks and the average time on the tasks (e.g., Luce, 1986, Fig. 6.13). However, the time spent on an item is driven by the test taker's speed parameter whereas the ability parameter plays this role with respect to the correctness of the response. The equivalent of the speed-accuracy tradeoff in reaction-time research is therefore a *speed-ability tradeoff* in testing. (Note the steps from two different observations to parameters for their distributions. Unlike the

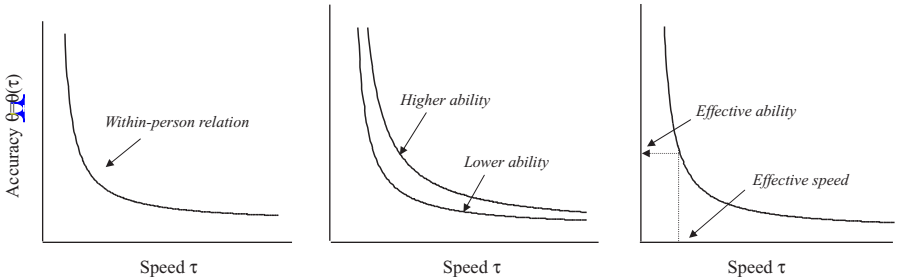

FIGURE 3. *Examples of ability as a monotonically decreasing function of speed τ (left panel), ability functions for a more and a less able test taker (middle panel), and effective speed and ability of a test taker during the administration of a test (right panel).*

observations, these parameters are adjusted for the structural differences between the items.)

A fictitious example of a speed-ability tradeoff is given in Figure 3 (left panel). The shape of the curve is entirely arbitrary; the only thing implied by the tradeoff is a monotonically decreasing relation between speed and ability. In the example, speed is chosen to be an independent variable and ability a dependent variable. This has been done intentionally because it seems plausible that a test taker has some control of speed but has to accept the probability of success that is the result of it. If this view is correct, we may think of ability as a function of speed; that is, adopt a function $\theta = \theta(\tau)$.

It is important to note that this function is for one fixed test taker; the tradeoff between speed and ability is entirely a within-person phenomenon. Its existence can only be demonstrated by forcing a test taker to speed up or slow down and checking the rate of success—not otherwise, particularly not by plotting the speed and ability estimates of different test takers against each other, which is exactly what happens for a typical reaction-time experiment in psychology!

How about the case of two test takers, where one is more able than the other? Intuitively, when they work at the same speed, it seems obvious to assume a higher rate of success for the more able test taker. Reversely, if the two have the same rate of success, the more able test taker can realize this at a higher speed. The case is represented by the two fictitious curves in the middle panel of Figure 3, with the dominating curve for the more able test taker.

When individuals take tests, they choose a certain level of speed based on such factors as their understanding of the test instructions, perception of the time limit, and style of work. We will refer to the result as the test taker's effective speed during the test. As illustrated in the right panel of Figure 3, the result is an effective ability. (Roskam's model discussed earlier was also based on the notion of an effective ability but this was taken to be equal to mental speed times processing time.)

At first sight, the assumption of constancy of speed during a test may seem gratuitous. But it is confirmed by a whole history of successful applications of IRT models to real-world tests: From the speed-ability tradeoff, it follows that response models with a single ability parameter for each test taker can fit only when they operate at

constant speed during the test. Of course, in the real world speed will always fluctuate somewhat during the test. But a recent empirical study (van der Linden, Breithaupt, Chuah, & Zhang, 2007) found such changes to be negligible. Also, it is common to view larger changes in ability or speed as aberrances due to design flaws in the test or misbehavior by the test taker (see the examples below). The standard approach is to assume constancy and then check for possible violations of the assumption due to such aberrances.

Figure 3 implies a ranking of the test takers according to their ability functions—that is, according to their points of intersection with a line through the origin—and not along the vertical ability axis. Thus, even when we account for random error, test scores do not automatically reflect the rank order of the test takers' abilities. They do so only when test takers operate at the same speed; otherwise, the scores are confounded with their decisions on speed. This point becomes an issue of fairness when different test forms are involved and some of them force their test takers to work faster than the others.

The view of ability as a function $\theta = \theta(\tau)$ is different from the usual conception of θ as a single point on an ability scale in IRT. Actually, the view is still not entirely correct. A more appropriate concept would be that of an ability area constrained from above by the curves $\theta = \theta(\tau)$ in Figure 3. These curves then represent the limit of what a test taker can achieve. It is always possible for a test taker to become less than optimally motivated and realize a combination of speed and ability somewhere below the curve but the combinations above it are out of reach. However, when the stakes are high and there are no motivation problems, we can just focus on the upper side of the area and ignore the rest of it.

Conclusion 4: Speed and ability are related through a distinct function $\theta = \theta(\tau)$ for each test taker. The function itself need not be incorporated in models for RTs and responses on achievement test items. But these models do require fixed parameters for the effective speed and ability of the test takers.

Item Difficulty and Time Intensity

As stated in Conclusion 3, RT models with a speed parameter also need a parameter for the time intensity of the items. Some of the earlier models do have a parameter with this interpretation but in combination with the traditional item difficulty parameter from IRT; for example, (8), (12), and (14). Also, Rasch's model for reading speed has a text parameter that was interpreted by him as a difficulty parameter. The intuition behind these models thus points at an assumed impact of the difficulty of the item on its RT distribution. The intuition seems to receive support from several empirical studies that report a positive correlation between RT and item difficulty; see, e.g., Masters (2005), Smith (2000), Swanson, Featherman, Case, Luecht, and Nungester (1999), Swanson, Case, Ripkey, Clauser, and Holtman (2001), and Zenisky and Baldwin (2006). (The nature of such studies is discussed in the next section.)

The two arithmetic items in Figure 2 were chosen with this question in mind. Item 2 involves a longer series of cognitive operations and requires more time than item 1. However, we expect the two items to show hardly any difference in difficulty.

A test taker able to add three-digit numbers (the common part of the two items) will not have much problem adding the extra two-digit numbers in item 2.

This example illustrates our view of the relation between the time intensity and difficulty of an item. The former refers to the amount of processing that has to be done; the latter summarizes how the test taker's ability is challenged by the nature of the operations involved in it. For the items in Figure 2, the difficulty is expected to be determined mainly by the operation with the greatest challenge. For other types of items, a different summary of the challenges may be necessary to define item difficulty.

A more fundamental argument, however, is based on the distinction between manifest and latent parameters. The former are measured directly (e.g., word counts for items) and can therefore be included in different models without any change of meaning. But item difficulty and time intensity are latent parameters that derive their meaning entirely from the fact that they represent the effects of the items on their probability of success and the time spent on it, respectively. Because these are different quantities, the two types of effects are different (although they may correlate across items). In fact, if a "difficulty parameter" would be included in a RT model, it would immediately lose its interpretation as a difficulty parameter and just become a second parameter for the effect of the item on the RT.

Conclusion 5: RT models require item parameters for their time intensity but difficulty parameters belong in response models.

Dependences between RTs and Responses

The issue of possible dependencies between RTs and responses is complicated. On the one hand, descriptive studies of the relationship between RTs and responses show substantial correlations between them. For example, Bergstrom, Gershon, and Lunz (1994) found that correct responses were generally produced faster than incorrect responses. The same results were found by Hornke (2000) and in a rather comprehensive study by Röhling (2006). These results seem to be at odds with the experimental confirmation of the speed-accuracy tradeoff in reaction-time research in psychology, where more time has invariably been shown to lead to a larger proportion of correct responses (see the review of this research in Luce, 1986, chap. 6). The same holds for the earlier models with a positive dependency of the probability of a correct response on RT (e.g., those by Roskam, 1997, and Wang & Hanson, 2005).

On the other hand, empirical studies of the dependencies between RTs and responses in operational testing always involve some kind of data aggregation. For each item-person combination, only one RT and response are observed, and it is impossible to estimate their correlation at this level. In the typical descriptive study, therefore, these observations are correlated across test takers or items. But such data aggregation can be dangerous and easily leads to spurious correlations because of hidden covariates.

An example of a similar spurious correlation is between the responses U_{ij} and $U_{i'j}$ on two different items i and i' for a population of test takers. The correlation vanishes as soon as ability is kept constant, a phenomenon known as "local

TABLE 1
Response and RT Vectors of Two Arbitrary Subjects on a Test of Quantitative and Scientific Proficiency for College Students

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Subject 1															
Response	0	0	1	1	1	1	0	1	0	0	0	0	0	1	1
RT	22	19	40	43	27	27	45	23	14	47	12	5	3	4	16
Subject 2															
Response	0	0	0	1	1	0	1	1	1	1	0	1	1	0	0
RT	26	38	101	57	37	21	116	44	10	117	18	9	25	16	34

independence." Similarly, the assumption of conditional independence between T_{ij} and $T_{i'j}$ on different items i and i' seems natural. In fact, it would be difficult to think of conditionally independent responses but dependent RTs. As long as a test taker's speed on the items is constant, the variation in RTs about their expected values is just random. But when they are aggregated over test takers operating at different speeds, they begin to correlate.

It seems natural to consider the issue of possible dependencies between RTs and responses on a single item along the same lines: For a single test taker, when both speed and ability are constant, we should assume conditional independence between T_{ij} and U_{ij} for every item i. But as soon as we aggregate responses and RTs across different test takers (as was done in the empirical studies above), the correlation between their speed and ability parameters serves as a potential source of variation. If more able test takers also work more quickly, T_{ij} and U_{ij} correlate positively. If they work more slowly, the correlation becomes negative.

The same happens if we correlate T_{ij} and U_{ij} across items and their difficulties and time intensities correlate. An example of spurious correlations due to this hidden source of covariation is given in Table 1, which shows the responses and RTs for two test takers on the first 15 items of a test of quantitative and scientific proficiencies for college students from Wise, Kong, and Pastor (2007). Because the two subjects were arbitrarily selected from a much larger data set, communication between them during the test can be excluded and both their responses and RTs are independent. Nevertheless, the two response vectors show a positive correlation across the items ($r = .20$). The only possible explanation of this is variation in difficulty from one item to the next. The correlation between the RTs is even stronger ($r = .89$), so we expect the variation in time intensity between the items to have a relatively greater impact. More surprisingly, the responses by one test taker even correlate with the RTs by the other ($r = .27$ and .21). These correlations are also spurious; the only meaningful explanation is a positive correlation between the difficulties and time intensities of the items (which we have typically found in our empirical studies; see the review below).

Our analysis has thus revealed three different assumptions of conditional or local independence relevant to the modeling of the relationship between responses and RTs: (a) independence between responses to different items, (b) independence

between RTs on different items, and (c) independence between responses and RTs on the same items.

The dependences are closely related to the more fundamental assumption of constancy of speed and ability during the test discussed earlier. As long as they are constant, speed and ability cannot serve as common covariates. Of course, in real-world test administrations, speed and ability will always fluctuate somewhat during the test and, as a result, minor violations of the conditional independence assumptions will be observed. When the violations are random and minor indeed, there is no problem. But when they become larger and more systematic, they may point at a warming up effect as the result of inadequate instructions to the test, fatigue due to the length of the test, or the impact of a time limit that was set too tight. In such cases, it is better to redesign the test than to replace the model by a more complicated version to allow for conditional dependence.

Conclusion 6: In addition to the usual assumption of conditional independence between responses to different items, it seems reasonable to assume conditional independence between RTs and responses on the same items as well as between RTs on different items.

Modeling Framework

Our six conclusions confirm Rasch's idea of different models for the distributions of responses and RTs on test items. Both models should have distinct person and item parameters. The parameter structure of the response model can be that of a regular IRT model with a person parameter for the effective ability of the test taker and the usual item parameters. The parameter structure of the RT model should be consistent with the fundamental equation derived in this article—a requirement leading to the postulate of a parameter for the time intensity of an item in addition to one for the effective speed of the test taker. Although speed and ability are two person parameters with a tradeoff relation for each test taker, it is not necessary to represent the relationship by the parameter structure of either model as long as the test takers operate at (approximately) constant speed. Conditional independence between responses and RTs seems a plausible assumption. But in order to explain observed correlations between responses and RTs across persons and items, it is necessary to extend the models with a structure that allows for dependencies between their parameters.

A graphical representation of the framework for modeling responses and RTs emerging from these conclusions is given in Figure 4. The framework is hierarchical in that it has two lower-level models for the responses and RTs by a fixed person as well as two higher-level models for the joint distributions of their parameters. The two lower-level models have ability and speed parameters for the test takers as well as difficulty and time-intensity parameters for the items. The two higher-level models are for the distributions of the person parameters in the population of test takers and the item parameters in the domain of test items. The correlations between these parameters are to be inferred from these distributions.

The framework does not prescribe any specific component models. It only defines the general nature of these models as well as the relevant relationships between

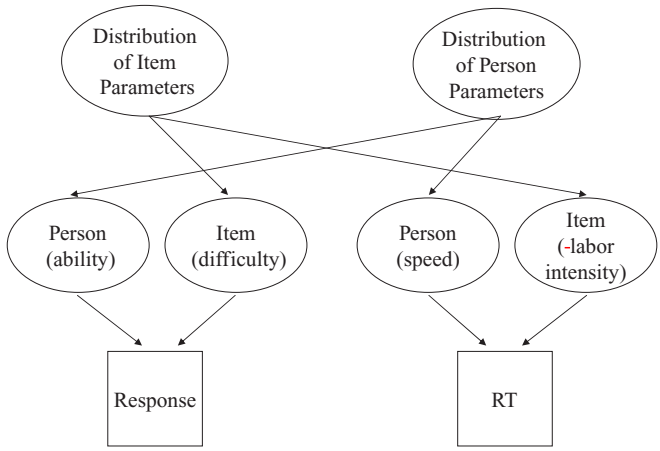

FIGURE 4. *Graphical illustration of a hierarchical framework for the modeling of responses and RTs on test items.*

their parameters. For different types of applications, different component models or different parameterizations may be required. For example, the response model should be chosen to support the response format of the items (e.g., a dichotomous or a graded format). To illustrate the possibilities, we discuss two different choices of component models.

Dichotomous Items and Lognormal RTs

The lower-level models are the regular 3PL model for dichotomously scored items and a lognormal model for the RTs. The higher-level models are multivariate normal distributions for the person and item parameters. This choice of component models was examined in van der Linden (2007).

Lower-level models. The 3PL model postulates a probability of success on item i that can be written as

$$p_i(\theta_j) \equiv c_i + (1 - c_i)\Psi[a_i(\theta_j - b_i)], \tag{25}$$

where $\Psi(\cdot)$ is the logistic function, $\theta_j \in [-\infty, \infty]$ is the ability parameter of test taker j, and $a_i \in [0, \infty]$, $b_i[-\infty, \infty]$, and $c_i \in [0, 1]$ are the usual discrimination, difficulty, and guessing parameters for item i, respectively.

The lognormal RT model follows directly from the fundamental relation in (23); the only extension is the assumption of a normal distribution of the RTs around their expected values $\beta_i - \tau_j$. The result is

$$\ln T_{ij} \equiv \beta_i - \tau_j + \epsilon_i, \quad \epsilon_i \sim N\left(0, \alpha_i^{-2}\right). \tag{26}$$

Equivalently, the model can be written as a lognormal density for the distribution of T_{ij}:

$$f(t_{ij};\tau_j,\alpha_i,\beta_i)=\frac{\alpha_i}{t_{ij}\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{ij}-\{\beta_i-\tau_j\})\right]^2\right\}. \tag{27}$$

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

This lognormal model was proposed earlier just as a flexible model for skewed RT distributions (van der Linden, 2006) but appears to follow directly from the fundamental equation derived earlier in this article.

Higher-level models. Let $\mu=(\mu_\theta,\mu_\tau)$ and $\sigma=(\sigma_\theta,\sigma_\tau)$ be vectors with the means and standard deviations of the ability and speed parameters in the population of test takers and $\rho_{\theta\tau}$ their correlation. The population model is a bivariate normal distribution for the person parameters with density function

$$f(\theta,\tau;\mu,\sigma,\rho_{\theta\tau})=\frac{1}{2\pi\sigma_\theta\sigma_\tau\sqrt{1-\rho_{\theta\tau}^2}}\exp\left[-\frac{1}{2\left(1-\rho_{\theta\tau}^2\right)}\left(z_\theta^2-2\rho_{\theta\tau}z_\theta z_\tau+z_\tau^2\right)\right], \tag{28}$$

with $z_\theta=(\theta-\mu_\theta)/\sigma_\theta$ and $z_\tau=(\tau-\mu_\tau)/\sigma_\tau$.

Similarly, the item-domain model is a multivariate normal distribution for all item parameters in the response and RT models. Let $\xi_i=(a_i,b_i,c_i,\alpha_i,\beta_i)$ be the (column) vector with the item parameters in (25)–(27). It is assumed to have a normal distribution with density function

$$f(\xi_i;\mu,\Sigma)=\frac{|\Sigma^{-1}|^{1/2}}{(2\pi)^{5/2}}\exp\left[-\frac{1}{2}(\xi_i-\mu)^T\Sigma^{-1}(\xi_i-\mu)\right], \tag{29}$$

where $\mu=(\mu_a,\mu_b,\mu_c,\mu_\alpha,\mu_\beta)$ is the (column) vector of the means and Σ is the covariance matrix for the item parameters.

The framework needs to be extended by a few additional restrictions to fix the scale of some of its parameters. Bayesian methods for estimating the parameters and checking the fit of the component models are available. However, the statistical treatment of response and RT models is beyond the scope of this article; for technical details, we refer to Fox, Klein Entink, and van der Linden (2007); Klein Entink, Fox, and van der Linden (2009); and van der Linden (2006, 2007).

Hierarchical Version of Rasch's Models for Oral Reading Tests

The second example is a straightforward extension of the two earlier models for oral reading tests by Rasch. First, we replace the equations for Poisson parameter θ_{ij} in (2) and gamma parameter λ_{ij} in (5) by their logarithmic versions

$$\kappa_{ij}=b_i-\theta_j \tag{30}$$

and

$$\lambda_{ij} = \tau_j - \beta_i, \tag{31}$$

respectively. The only step left is to adopt the normal distribution in (28) for parameters θ and τ in (30) and a bivariate version of the distribution in (29) for parameters b and β in (31).

The statistical treatment of this framework is also omitted here; for technical details and an empirical application, see van der Linden (2008b).

Empirical Applications

In Figure 4, there are no direct arrows between the response and RT; neither are there any from one kind of observation to a parameter for the other. The absence of such arrows reflects the assumptions of conditional independence and constancy of speed and ability. However, as already illustrated using the data in Table 1, the possibility of a nonzero correlation between the observed responses and RTs in a sample of test takers and/or test items arises because of second-level correlations. The size and sign of these observed correlations in a sample of test takers or items depend entirely on the pattern of these higher-level correlations.

In an application of the hierarchical model (25)–(29), the second-level correlations are estimated directly from the observed responses and RTs. For the sake of illustration, we focus on the correlations $\rho_{\theta\tau}$ and $\rho_{b\beta}$. In a recent series of applications, different patterns of correlations were found. For an arithmetic test in the adaptive version of the *Armed Services Vocational Aptitude Battery* (ASVAB), the correlations were estimated to be $\rho_{\theta\tau} = .04$ and $\rho_{b\beta} = .65$ (van der Linden, Scrams, & Schnipke, 1999). In another study, for the computerized *Uniform CPA Exam*, the correlations were found to be $\rho_{\theta\tau} = .30$ and $\rho_{b\beta} = .30$ (van der Linden et al., 2007) whereas they were estimated to be $\rho_{\theta\tau} = -.25$ and $\rho_{b\beta} = -.33$ for the Quantitative section of the *Graduate Management Admission Test* (GMAT) (van der Linden & Guo, 2008). For a test of Dutch as a foreign language, Klein Entink et al. (2009) found $\rho_{\theta\tau} = .25$ and $\rho_{b\beta} = .51$. Two other studies by these authors addressed the person parameters only and showed estimates equal to $\rho_{\theta\tau} = -.65$ for a *Natural World Assessment Test* (NAW-8) and $\rho_{\theta\tau} = .30$ for a neurosis scale in a personality questionnaire.

The impression emerging from these analyses is a strong tendency to a substantial positive correlation between the difficulties and time intensities of the items (with the GMAT test as the only exception), but correlations between speed and ability may be positive or negative. Our explanation of the wider range of correlations between speed and ability is better time-management skills among the more able test takers. When the time limit is tight, they know how to speed up to finish in time and distribute their time correspondingly, implying a positive correlation between speed and ability. But when there is ample time, they slow down to maximally profit from it (negative correlation).

In each of these applications, the fit of the model was checked carefully. Because the items were from an operational test and had already been shown to fit the 3PL

model, we focus on the loglinear RT model. The shape of the loglinear distribution was checked using plots with the cumulative distributions of the posterior predictive *p*-values across all person-item combinations (for examples of these plots, see van der Linden, 2006). The plots tended to show quite a satisfactory fit, except for a tendency for some of the items to have a slightly thinner lower tail than predicted by the model. But even for the worst items, the fit was still satisfactory enough for all practical purposes.

The assumption of constant speed was checked for the data set from the *Uniform CPA Exam* above (van der Linden et al., 2007). This exam has a randomized order of the items across the examinees. Hence, it was possible to study the effect of item order without any confounding with item content. Plots of the residual RTs as a function of the item position showed quite constant speed except for a warming-up effect at the beginning of the test. The effect was practically negligible, though; the maximum average residual among the earlier items in the test was only $+1.3$ seconds relative to an average RTs for the whole test close to one minute.

The same data set was used to check the three different assumptions of conditional independence using formal statistical tests derived for this purpose by van der Linden and Glas (in press). Due to the power of these tests and the large number of test takers, several statistically significant violations of the assumptions were obtained but their sizes were all negligible again and no systematic pattern could be detected. For instance, the average effect of the violations of conditional independence between responses and RTs was estimated to amount to a shift of only .47 seconds for the RT distribution on the items. Similarly, the average residual correlation between the RTs on the pairs of items was .06.

Of course, other types of tests may induce RTs that are more difficult to model. One obvious example is a reading comprehension test with items grouped in sets around common passages. This structure may lead to different strategies in the population of test takers, which the current RT model cannot accommodate (e.g., reading the entire passage carefully and then answering all items vs. going back and forth between the passage and each of the items). On the other hand, although all items in the studies above were dichotomous, we believe the case of polytomous items will only require the substitution of an appropriate response model into the hierarchical framework in (25)–(29). We do not see why the change of response format should have any other effects on the RTs distribution other than on their location and variance. For both effects, the model in (27) does have parameters. But, of course, empirical research will have to prove this claim.

The extension of regular response modeling with models for RTs may help us solve several existing practical testing problems. For instance, so far, the problem of assessing the degree of speededness of a new test form has been difficult. But the presence of RT models with an explicit parameter for the speed of the test takers solves it. For example, it is now possible to fit a RT model and check the residual RTs for any changes in speed during the test (van der Linden et al., 2007). A more effective approach is to entirely prevent the problem of speededness by assembling test forms to satisfy a target level of speed using the time parameters for the items. For adaptive tests, this has been demonstrated to be possible by constraining item selection in real time (van der Linden, 2005, sect. 9.5; 2009). RTs

are also a valuable source of collateral information on IRT parameters. This information was demonstrated to substantially improve small-sample item calibration in van der Linden, Klein Entink, and Fox (2008). The same type of information improves item selection in adaptive testing (van der Linden, 2008a). Another useful application of RTs is to detect aberrant behavior of test takers, for example, due to preknowledge of some of the items or attempts to memorize items during the tests (van der Linden & Guo, 2008). A bivariate version of the lognormal RT model for the detection of collusion between test takers has been proposed in van der Linden (in press).

These applications of RTs are only the first that have been addressed. Many others are on the horizon, for instance, improvement of item analysis or adjustment for differences in speededness between test forms in test equating. In fact, we should rethink all of educational measurement because nearly every aspect of it has a time dimension.

Concluding Remarks

The two different traditions of RT and response modeling reviewed in this article were the tradition of distinct models for the RTs and responses and that of the incorporation of RT parameters into response models (or the reverse). The models in the first tradition do not suggest any relationship between responses and RTs. The second tradition typically conceives of the relationship as a speed-accuracy tradeoff and tries to build this into one of the models.

The third type of modeling in this article maintains the idea of distinct models but adopts relationships between their parameters at a second level of modeling. Our motivation for this hierarchical approach was the realization that speed-accuracy tradeoffs are within-person phenomena that need no representation in response and RT models when the test takers can be assumed to operate at (approximately) constant speed during the test. On the other hand, to account for the correlations between observed responses and RTs among test takers and/or items, the second-level structure with the distributions of their parameters is required.

A fourth type of modeling, on which we have touched only occasionally in this article, exists in mathematical psychology. Its models are for observations in reaction-time experiments with groups of subjects, considered as exchangeable, replicating a standardized task under different conditions. For these models, which have neither item nor person parameters, see Luce (1986). A recent example in the psychometric literature is Rouder, Sun, Speckman, Lu, & Zhou, 2003). For educational testing these parameters are key because they enable us to adjust RTs for item effects when measuring the speed at which test takers operate or for person effects when the interest is in the time required by the cognitive labor the items involve.

Acknowledgments

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this article are those of the author and do not necessarily reflect the policy and position of LSAC. Portions of the research were presented in the author's NCME Career Award winner address, San Francisco, April 8–10, 2006.

References

Bergstrom, B., Gershon, R., & Lunz, M. E. (1994, April). *Computer-adaptive testing: Exploring examinee response time using hierarchical linear modeling*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Casella, G., & Berger, L. (1990). *Statistical inference*. Pacific Grove, CA: Brooks/Cole.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543.

Fox, J.-P., Klein Entink, R. H., & van der Linden, W. J. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*(7), 1–14.

Gaviria, J.-L. (2005). Increase in precision when estimating parameters in computer assisted testing using response times. *Quality & Quantity*, *39*, 45–69.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica*, *21*, 175–189.

Jansen, M. G. H. (1986). A Bayesian version of Rasch's multiplicative Poisson model for the number of errors on achievement tests. *Journal of Educational Statistics*, *11*, 147–160.

Jansen, M. G. H. (1997a). Rasch model for speed tests and some extensions with applications to incomplete designs. *Journal of Educational and Behavioral Statistics*, *22*, 125–140.

Jansen, M. G. H. (1997b). Rasch's model for reading speed with manifest exploratory variables. *Psychometrika*, *62*, 393–409.

Jansen, M. G. H., & van Duijn, M. A. J. (1992). Extensions of Rasch's multiplicative Poisson model. *Psychometrika*, *57*, 405–414.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

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

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

Masters, J. (2005, April). *Comparing item response times and difficulty for calculation items*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Oosterloo, S. J. (1975). *Modellen voor reactie-tijden* (Models for reaction times). Unpublished master's thesis, Faculty of Psychology, University of Groningen, The Netherlands.

Pieters, L. P. M., & van der Ven, A. H. G. S. (1982). Precision, speed, and distraction in time limit-tests. *Applied Psychological Measurement*, *6*, 93–109.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Röhling, P. G. (2006). *Effecten op responstijd bij item respons toetsen* (Response-time effects with item response tests). Unpublished master's thesis, Department of Psychology, University of Amsterdam, The Netherlands.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–171). Amsterdam: North-Holland.

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

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*, 18–38.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, R. W. (2000, April). *An exploratory analysis of item parameters and characteristics that influence item response time*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE Step 1. *Academic Medicine*, *76*, 114–116.

Swanson, D. B., Featherman, C. M., Case, S. M., Luecht, R. M., & Nungester, R. (1999, March). *Relationship of response latency to test design, examinee proficiency and item difficulty in computer-based test administration*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incorporating response-time data in scoring achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236–256). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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*. (pp. 179–203) New York: Academic Press.

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

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: 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 speed and accuracy on test items. *Psychometrika*, *73*, 287–308.

van der Linden, W. J. (2008a). Using response times for item selection in adaptive tests. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J. (2008b). *A hierarchical version of Rasch's Poisson process models for oral reading tests*. Manuscript in preparation.

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

van der Linden, W. J. (in press). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*.

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., & Glas, C. A. W. (in press). Statistical tests of conditional independence between responses and response times on test items. *Psychometrika*, *75*.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2008). *IRT parameter estimation with response times as collateral information*. Manuscript submitted for publication.

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

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

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

Wise, S. L., Kong, X. J., & Pastor, D. A. (2007, April). *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.

Woodbury, M. A. (1951). On the standard length of a test. *Psychometrika*, *16*, 103–106.

Woodbury, M. A. (1963). The stochastic model of mental test theory and an application. *Psychometrika*, *28*, 391–393.

Zenisky, A. L., & Baldwin, P. (2006, April). *Using response time data in test development and validation: Research with beginning computer users*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Author

WIM J. VAN DER LINDEN is Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; wim_van der linden@ctb.com. His primary research interests include test theory, applied statistics, and research methods.