

Modeling response times with latent variables: Principles and applications

*Wim J. van der Linden*¹

Abstract

The introduction of the computer in psychological and educational testing has enabled us to record response times on test items in real time, without any interruption of the response process. This article reviews key principles for probabilistic modeling of these response times and discusses a hierarchical model that follows from the principles. It then shows the potential of the model for improving the current practices of item calibration, adaptive testing, controlling test speededness, and detection of cheating.

Key words: adaptive testing; cheating detection; item calibration; hierarchical modeling; item-response theory (IRT); latent-variable modeling; response-time modeling; speededness; test design

¹ Correspondence concerning this article should be addressed to: Wim J. van der Linden, PhD, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; email: wim_vanderlinden@ctb.com.

A response to a test item is always the result of two equally important factors – the test taker and the item. The very reason why we test a person is an interest in his or her ability. But it would be naive to consider an interest in the ability as a reason to neglect the properties of the item. In fact, we can only infer a test taker's ability level from his responses to the extent that we are able to separate its effect on them for those of the items. Hence the crucial question, how to separate these two kinds of unknown effects?

The question is not new; it has been addressed in a long tradition in test theory, which dates back all the way to Binet's scaling of the items in the first intelligence test (Binet & Simon, 1995), and runs via Thurstone's (1925) pioneering work on scale theory, Lazarsfeld's (1950) idea of latent structure analyses, to the more recent developments in item response theory. Although the tradition is thus much older, we had to wait for Lazarsfeld for the current terminology of latent variables and parameters. All of his work rests on the fundamental distinction between what is manifest and latent: the only thing observed in testing are responses (manifest data), but what we are actually interested in are the unknown abilities (latent variable) and properties of the items (latent parameters).

One of the most lucid discussions of how to disentangle latent effects was given by Rasch (1960). The solution consists of the specification of a formal model for the interaction of the latent effects on the responses. Suppose we have (dichotomously scored) responses U_{ij} to items $i=1,\dots,I$ and persons $j=1,\dots,J$. Let θ_j^* denote the latent ability of person j and b_i^* the latent difficulty of item i . Rasch argued that the probability of a correct response to an item should be a function of the ratio

$$\frac{\theta_j^*}{b_i^*}, \quad (1)$$

or, for $\theta_j^* = \exp(\theta_j)$ and $b_i^* = \exp(b_i)$, the difference

$$\theta_j - b_i. \quad (2)$$

As function, for simplicity, he chose the logistic

$$\Psi(x) = [1 + \exp(-x)]^{-1}. \quad (3)$$

Substitution of (2) into (3) gives the well-known Rasch model for the probability of a success on an item,

$$\Pr\{U_{ij} = 1 | \theta_j\} = [1 + \exp(-(\theta_j - b_i))]^{-1}, \quad (4)$$

as a function of the unknown ability θ_j .

It is key to note that the left-hand side of the equation is the probability of a manifest response whereas its right-hand side contains the latent parameters representing the person and item effects. Separating these effects requires thus nothing but a model with explicit parameters for them that explains their interaction. Therefore, measuring a test

taker's unknown ability just amounts to statistical estimation of his or her parameter from the observed responses.

One of the highlights in Rasch's monograph is his illustration of the universality of the principle of modeling latent effects, using Maxwell's re-analysis of Newton's concepts of mass and force (Rasch, 1960, chap. 7). When Newton introduced his famous laws of motion, it was already known how to measure the acceleration of a body but still unclear how to deal with the notions of mass and force, let alone how to measure them. Clearly, an acceleration a_{ij} is the result of a force F_j by an agent j acting upon a physical body i with mass m_i . Newton's second law equates the manifest acceleration to the ratio of the latent force and mass; that is, as

$$a_{ij} = \frac{F_j}{m_i}. \quad (5)$$

Analogous to ability measurement, the equation allows us to determine the force that acted on a body by adjusting its acceleration for its mass.

Of course, although the parallel between (1) and (5) is striking, and it will even be continued in a fundamental equation of response-time modeling presented below, the assumption of a *ratio* is not necessary. If it were, all of physics would have been built around ratios between pairs of variables – a clear contradiction with its current, highly complex body of knowledge. What counts is formal modeling with separate parameters for each effect and a structure that reflects their interaction. Also, for probabilistic modeling, except for convenience, it is hard to make a principled choice for a specific function as the logistic in (4). In the empirical examples presented later in this paper, we use a modification of the logistic function with two additional item-specific parameters to represent more complicated item effects. The use of explicit parameters for each effect still allows us to separate them.

1. Principles of response-time modeling

Thanks to the advent of the computer in testing, we are now able to record routinely the response times (RTs) by test takers on test items. In anticipation of this development, the recent literature on test theory has already shown several models for the analysis of RTs. Basically, the models belong to two different categories: (i) distinct models for RT distributions without any reference to the responses on the same items; and (ii) models integrating responses and RTs (that is, extensions of response models that incorporate RTs or models for RT distributions that include response features).

A prominent example of a model in the first category is Rasch's gamma model for reading speed, which models RTs independently of the responses to the items. [Actually, Rasch seemed uncertain as to the status of the parameters in this RT model. He referred to them as ability and difficulty parameters, just as in his Poisson model for reading errors; see his comment on these parameters (Rasch, 1960, p. 42).] Gamma distributions were also used for RT modeling in Maris (1993). But they are not the only type of

distributions that have been proposed for RTs; for example, Tatsuoka and Tatsuoka (1980) based their RT model on the Weibull distribution.

Examples of response models that incorporate RTs are the versions of the Rasch model by Roskam (1987, 1997) and Verhelst, Verstralen and Jansen (1997). The major difference between these two models is the addition of the logRT and a separate parameter for the speed of the test taker as an extra term to the parameter structure of the Rasch model in (4), respectively. The best known version of a model for a RT distribution that includes response modeling is Thissen's (1983) model, which regresses the log RT simultaneously on new RT parameters as well as the typical parameter structure used in dichotomous response models.

Before choosing any RT model, we reflect on the principles that should guide our choice. The discussion in the following sections is derived from van der Linden (2009a).

1.1 Single test taker and item

The RT model should be specified for a single test taker responding to a single item. For models at this level, it is always possible to move up and derive results for aggregated RT distributions across items and/or persons; for instance, to explain such distributions across adaptive testing (Hornke, 2000, 2005) or in item banks (Scrams & Schnipke, 1999). On the other hand, models specified directly at a higher level of aggregation do not allow for any result at a lower level. For some of the current models, because of the lack of explicit indices for their parameters, it is even difficult to determine at which level they should apply (e.g., Tatsuoka & Tatsuoka, 1980).

1.2 Random RTs

Each RT observed for a combination of a test taker and item should be treated as the realization of a random variable, even though the event cannot be replicated because of memory and/or learning effects. For simple tasks for which such effects are absent or completely negligible, e.g., quick decisions, judgments, or motor tasks, experimental research on reaction times in psychology has presented overwhelming evidence of randomly varying RTs (e.g., Luce, 1986). Models that regress responses on observed RTs (e.g., Wang & Hanson, 2005) go against this principle.

1.3 Status of parameters

The previous two principles seem to suggest that RT models follow structures typical of item-response theory. Both should allow for person and item effects in the form of appropriately indexed parameters, the major difference being only in the type of distribution they explain (responses vs. RTs). However, it would be inappropriate to assume anything beyond an analogy, for instance, common parameters.

Speaking in favor of this is the earlier identical interpretation of the person and item parameters as ability and difficulty parameters in the Rasch (1960) models for reading speed and reading error. This interpretation ignores the fact that latent parameters are not measured directly and, hence, can derive their meaning only from the manifest variable that is modeled.

For instance, persons parameter in response models can only be interpreted referring to the differences in the probability of success on items between test takers. Likewise, differences between person parameters in RT models reflect differences in time distributions between persons. These points are missed if we refer to both using the same predicate of "ability". For reasons that will become clear below, we will refer to differences in RT distributions between persons as differences in speed.

Similarly, we should not reify item parameters; that is, avoid assigning any existence to them beyond the effects of the items on the response and RT distributions they represent. An item is more difficult because it does have a higher error rate (not the other way around). But a higher error rate is not the same thing as a tendency to a longer response time. The latter depends, for instance, on the amount of material that has to be read or the number of computation steps that are involved. We will refer to items that require longer processing time as more labor- or time-intensive items.

Of course, under certain conditions, ability and speed parameters may correlate across test takers. Likewise, difficulty and labor intensity parameters may correlate across certain types of items. But such (positive or negative) correlations should not be taken to be evidence of their representing identical effects.

1.4 RT and speed

In contradiction with the tradition of latent variable modeling, which forces us to carefully distinguish between manifest RTs and the latent parameters required to model their distribution, the psychological literature on reaction times seems to have confounded these quantities. A prime example is the typical way in which speed-accuracy tradeoffs are reported, as proportions of successes plotted against average response times (e.g., Luce, 1986, chap. 6). Confounding time with speed, however, would be the same error as asking our colleagues how long it takes them to drive to work, and then conclude that the one with the shortest time drives fastest. The missing factor, of course, is the distance driven. Similarly, equating raw RTs on an item with speed overlooks the amount of labor required to solve it.

In fact, response time, speed, and labor intensity relate in the same fundamental way as the probability of success, ability, and difficulty in (1) and acceleration, force, and mass in (5). RTs are measured directly, but a test taker's speed and the labor intensity of an item are not measurable in any direct way. They are related, however, through the definition of speed as the amount of labor per time unit: Let t_{ij} denote the amount of time spent on item i by test taker j , β_i^* the amount of labor required by the item, and τ_j^* the (average) speed at which the test taker works. The standard definition of speed of

labor is the amount of labor over time; that is, $\tau_j^* = \beta_i^*/t_{ij}$. Moving the response time to the left-hand side of the equation, we get

$$t_{ij} = \frac{\beta_i^*}{\tau_j^*}, \quad (6)$$

which, again, models a manifest quantity as a ratio of two latent effects. This fundamental equation is the core of the lognormal model for RTs discussed in the next section.

1.5 Speed-accuracy tradeoff

Another issue that needs further clarification is the relationship between speed and ability. A main motive of the earlier modeling of RTs on test items was an attempt to account for a speed-accuracy tradeoff between the responses and RTs on test items. Psychology had demonstrated such tradeoffs for numerous tasks, and the analogy between experimental subjects working on them and test takers responding to items seemed too convincing to ignore. Obviously, the resulting models belong to the category that we have indicated earlier as integrate models for responses and RTs; examples are those by Roskam (1987, 1997), Thissen (1982), and Verhelst, Verstralen, and Jansen (1997).

A speed-accuracy tradeoff, however, is a within-person relationship that manifests itself as a decrease of accuracy when a subject speeds up. Speed-accuracy experiments in psychology are designed to induce changes in speed in subjects and observe the effects on their accuracy. However, psychological and educational tests are usually built on the assumption of constant speed and ability during testing. Observed violations of the assumption, such as a slow start or hurrying toward the end of the test, are typically viewed as flaws in the design of the test (e.g., unclear instructions; time limit too tight) that need correction.

The assumption of constant ability underlies the presence of one fixed ability parameter per person in the response model. A long history of fitting response models with such parameters confirms the reasonableness of the assumption. Further, the presence of a speed-accuracy tradeoff implies constant speed as a main condition for constant ability. Thus, in somewhat of an ironic twist, the speed-accuracy tradeoff actually forces us to ignore the relationship between speed and ability when modeling responses and RTs on test items!

It would be incorrect to accept a negative correlation between responses and RTs across test takers as evidence of a speed-accuracy tradeoff. Figure 1 shows how the same type of tradeoff could lead to different correlations across test takers. Each of the curves represents a hypothetical tradeoff within a test taker. During the test, most likely on the basis of a combination of the instructions to the test and personality factors, test takers decide on their level of speed and, in doing so, fix their level of accuracy (ability). Depending on these individual decisions, the results can be a positive, negative or zero correlation between speed and ability.

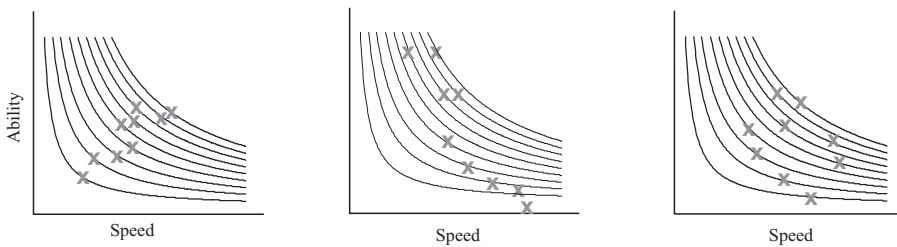


Figure 1:

Positive, negative, zero correlation between speed and accuracy for the same test takers and underlying speed-accuracy tradeoffs

1.6 Conditional independent responses and RTs

A final issue that needs clarification is the dependence between responses and RTs. Descriptive studies of testing data suggest negative correlations between them (incorrect answers take more time) (e.g., Hornke, 2000, 2005). Although our intuition appears to support this finding ("low ability students muddle along for a longer time and then give a wrong answer"), we should be aware of the fact that, for each combination of test taker and item, we have only one response and RT, and to draw any conclusion we have to aggregate them across items and/or persons. Such aggregation is known to lead to artefacts. For instance, when aggregating across test takers with different levels of ability, responses on different test items correlate, but given a level of ability they tend to be independent ("local independence"). The same holds for RTs aggregated across test takers operating at different speed, whereas correlation between speed and ability across test takers leads to correlated responses and RTs on the same item.

In addition to all earlier requirements, such as separate subject and item parameters, random responses and RTs, etc., the modeling framework in the next section assumes conditional independence between responses and/or RTs but also explains correlations observed at a higher level of aggregation.

2. Hierarchical model of responses and RTs

The previous desiderata imply distinct models for the response and RT distributions for a single test taker on a single item. Both models require separate person and item parameters, which can be interpreted only in terms of effects on response probabilities and RT distributions. Unless the goal is modeling changes in persons over time, the person parameters in both models are fixed constants; in particular, it is inappropriate to treat the person parameters in the response and RT model as functions of each other to reflect a speed-accuracy tradeoff. Finally, although it makes sense to assume conditional independence between responses and/or RTs given the person parameters, their models should be part of a hierarchical structure that explains the dependences between them

across items and/or test takers. The following two-level modeling framework meets these principles.

2.1 First-level models

The model for the RT distribution by a test taker on an item follows from the basic equation in (6) in two simple steps: First, it is well known that RT distributions tend to be skewed, the skew being the result of the presence of a natural lower bound but the absence of any upper bound on RTs. A standard way of removing the skew is a logarithmic transformation of (6),

$$\ln t_{ij} = \beta_i - \tau_j, \tag{7}$$

with $\beta_i = \ln \beta_i^*$ and $\tau_j = \ln \tau_j^*$. Second, to allow for the random nature of RTs, we add a random component to the previous result:

$$\ln t_{ij} = \beta_i - \tau_j + \varepsilon_i, \text{ with } \varepsilon_i \sim N(0, \alpha_i^{-2}). \tag{8}$$

The final result is identical to the following lognormal model for the distribution of the RT:

$$p(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[\alpha_i (\ln t_{ij} - (\beta_i - \tau_j)) \right]^2 \right\}, \tag{9}$$

where $\tau_j \in (-\infty, \infty)$ is the speed at which j operates during the test, and $\beta_i \in (-\infty, \infty)$ and $\alpha_i \in (0, \infty)$ are parameters for the labor-intensity and discriminating power of item i , respectively. Because β_i represents the effect of the item on the mean logtime, we also refer to it as its time intensity. Observe that α_i is the inverse standard deviation of the RT distribution. The choice of the inverse allows us to interpret α_i as a discrimination parameter, analogous to the one in the response model below.

The introduction of the RT model has no consequences whatsoever for the response model. Both are distinct, and the responses and RTs are assumed to be conditionally independent. In the applications below, we used the 3-parameter logistic (3PL) model, one of the mainstream models in educational testing. The model can be written as

$$p_i(\theta_j) \equiv \Pr(U_{ji} = 1 | \theta_j) \equiv c_i + (1 - c_i) \frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \tag{10}$$

where θ_j is the parameter for the ability of test taker j , and a_i , b_i , and c_i are parameters for the discriminating power, difficulty, and height of the lower asymptote of the response probability of item i , respectively.

Notice that (10) defines the probability of a correct response; unlike (9), it is not the full probability function for a random variable with a Bernoulli distribution. Although their two parameter structures may suggest so, the two models are therefore not analogous.

For instance, $\beta_i - \tau_j$ is the mean of the log RT on the item, but $\theta_j - b_i$ is not the mean response, which is $p_i(\theta_j)$. Also, the fact that $\ln t_{ij}$ – a quantity measured with a natural unit; e.g. seconds – is part of the argument of (9) gives all of its parameters a scale with a fixed unit; we only need one additional constraint to fix its arbitrary zero (e.g., $\mu_\tau = 0$). On the other hand, the response model in (10) has both an arbitrary unit and origin. For more details on this issue, see van der Linden (2010).

2.2 Second-level models

The previous two models are for the marginal distributions of the response and RT by a test taker on an item. Although these distributions are assumed to be independent, as already noted, responses and RTs may show substantial correlation across test takers and/or items. To allow for such dependencies, two separate second-level models are required, one for the joint distribution of the person parameters in (9) and (10), and another for the distribution of their item parameters.

Suppose that the data are aggregated across test takers in a population \mathcal{P} . We assume that their parameters have a multivariate normal distribution. That is,

$$(\theta, \tau) \sim \text{MVN}(\mu_{\theta\tau}, \Sigma_{\theta\tau}), \quad (11)$$

with $\mu_{\theta\tau}$ the vector of the population means of θ and τ and $\Sigma_{\theta\tau}$ their covariance matrix.

The joint distribution of all item parameters is modeled analogously as

$$(a, b, c, \alpha, \beta) \sim \text{MVN}(\mu_{abc\alpha\beta}, \Sigma_{abc\alpha\beta}), \quad (12)$$

with $\mu_{abc\alpha\beta}$ the vector of the means of all item parameters and $\Sigma_{abc\alpha\beta}$ their covariance matrix.

The full two-level model is represented in Figure 2. Observe that, in spite of the fact that the two models are for the same item and person, the two lower-level models have no direct arrows between them. This symbolizes the fact that the responses and RTs are independent, the response parameters have no impact on the RTs (and the other way around), and the parameters in the two models are entirely distinct.

2.3 Statistical treatment

The lognormal model in (9) was proposed in van der Linden (2006), the extension involving the complete hierarchical framework in van der Linden (2007). Klein Entink, Fox, and van der Linden (2009) extended the framework with covariates for the person parameters. The choice of the specific component models in (9)–(12) is not necessary, though; they can be replaced by any other model with the same basic distinction between

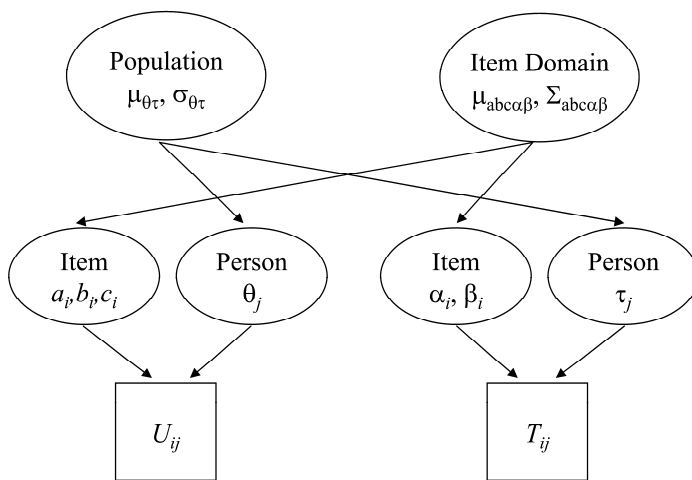


Figure 2:

Hierarchical framework for modeling responses and RTs, with distinct models for the response and RT on an item as well as the dependencies between their parameters across person and items. (Reproduced with permission from W. J. van der Linden (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287-308.)

item and person parameters. An alternative RT model is the Box-Cox normal model in Klein Entink, van der Linden, and Fox (2009). Also, when the responses are polytomous, or the assumption of unidimensional ability is untenable, a response model for the appropriate response format or dimensionality should be substituted for the 3PL model in (10).

The full hierarchical framework lends itself excellently to a Bayesian approach with Gibbs sampling from the posterior distribution of all parameters. For various versions of Gibbs sampling, see the previous references. Bayesian posterior checks for the fit of the RT model were given in van der Linden (2006), whereas model checks based on the Bayes factor and the deviance information criterion (DIC) were derived in Klein Entink, Fox, and van der Linden (2009). Lagrange multiplier tests for the assumptions of conditional independence between responses and/or RTs are presented in Glas and van der Linden (2010) and van der Linden and Glas (2010). For a software package written in *R*, see Fox, Klein Entink, and van der Linden (2007).

Generally, in empirical applications of the modeling framework with the above models, we have found good fit to the response and RT data, with varying patterns of correlations between the item and person parameters in datasets for different types of tests. Typically, the item difficulties and time intensities showed modest to strong positive correlation, but the correlation between speed and ability ranged from positive to negative; for a review, see van der Linden (2009a).

3. Applications of RT modeling

We review applications of the above type of joint response and RT modeling to four different problems in psychological and educational measurement: item calibration, adaptive testing, test design and speededness, and the detection of cheating.

In addition, the same type of modeling has been used to explain the cognitive structure of responses and RTs for certain content domains. For instance, Klein Entink, Kuhn, and Fox (2009) extended the framework with covariates for the item parameters in the response and RT model to explain the difficulties and time intensities of the items in a figural matrix test. Similar modeling with covariates for the person parameters was used by Goldhammer (2010) to test different hypotheses about speed and ability in reasoning tasks.

3.1 Test-item calibration

Before they are used operationally, test items ought to be pretested. This should be done in order to check their behavior, to calibrate them using IRT, and test the fit of the response model to the data. The use of the full model in (9)–(12) extends the traditional stage of item calibration with the estimation of the item parameters in the RT model in (9) as well as the second-level parameters involved. The RTs are recorded automatically with the responses, so from a practical perspective no extra efforts are required. As mentioned above, software for the calibration of the items with respect to all parameters is available in the *R* package in Fox, Klein Entink, and van der Linden (2007).

Two basic options for the calibration are: (i) separate estimation of the parameters in the response and RT model from their respective datasets and (ii) joint estimation of all parameters from the total set. Because of the second-level links between the response and RT parameters in (11)–(12), an important advantage of joint estimation is improved accuracy of estimation. For example, we are able to estimate the difficulties of the items more accurately when this is done jointly with their time intensities.

The improvement is a nice illustration of a statistical principle known as "using collateral information" or "borrowing strength from the estimation of other parameters." The logic of the principle is illustrated using Bayesian estimation of the item parameters $\xi_i = (a_i, b_i, c_i)$ in the response model in combination with the item parameters $\eta_i = (\alpha_i, \beta_i)$ in the RT model. For clarity of presentation, and without loss of generality, we ignore the estimation of all other parameters.

If the ξ_i parameters are estimated separately from the RT parameters, the estimation is based on the posterior distribution of ξ_i given the response vector, u_i . The distribution has density function

$$f(\xi_i | u_i) \propto f(u_i; \xi_i) f(\xi_i), \quad (13)$$

where $f(u_i; \xi_i)$ is the probability of the response data given the item parameters and $f(\xi_i)$ is the prior distribution of ξ_i . In the current context, it makes sense to avoid specifying a subjective prior distribution and use an empirical Bayes approach with the (trivariate normal) marginal distribution of ξ_i in (12) as prior distribution.

Now suppose we follow the second option and estimate the item parameters ξ_i both from the responses and RTs. We then have to replace (13) by the posterior distribution of ξ_i given u_i and t_i , which can be shown to have density function

$$f(\xi_i | u_i, t_i) \propto f(u_i; \xi_i) f(\xi_i | t_i). \tag{14}$$

The only difference between the two posterior distributions is the replacement of the prior distribution of $f(\xi_i)$ by $f(\xi_i | t_i)$; that is, the posterior predictive distribution of ξ_i given t_i , which has the density

$$f(\xi_i | t_i) = \int f(\xi_i | \eta_i) f(\eta_i | t_i) d\eta_i. \tag{15}$$

Figure 3 illustrates how the posterior distribution in (14) is derived: The first factor $f(u_i; \xi_i)$ in (14) is the probability of the response vector captured by the response model in the left-hand side of the figure. The second factor, $f(\xi_i | t_i)$, is built up in the right-hand side of the figure. The first step is the calculation of the posterior distribution of item parameter η_i in the RT model given the vector with the RTs for the item, t_i . The posterior distribution is then combined with the conditional probability of the item parameters ξ_i in the response model given η_i . The result gives us the posterior prediction distribution of ξ_i given t_i . The calculation steps are summarized in (15).

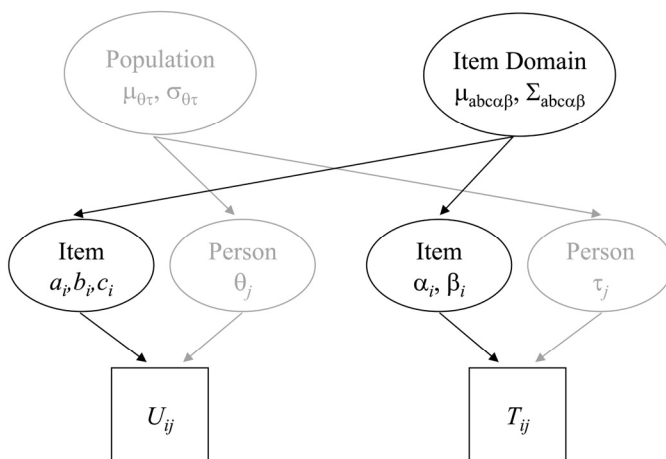


Figure 3:

Link between item parameters in the response and RT models that allows us to borrow the information in the RTs when estimating the item parameters in the response model (and the other way around).

Suppose again that the first option is based on the marginal distribution of ξ_i in (12) as prior distribution for the item parameters. The replacement of $f(\xi_i)$ by $f(\xi_i | t_i)$ in the second option results in the following gains:

1. The conditional distribution of ξ_i given t is generally much narrower than the marginal distribution of ξ_i , which means more (empirical) information about ξ_i and therefore more accurate estimation.
2. The conditional distribution of ξ_i given t depends on the actual vector of RTs for the item. It thus replaces a common prior $f(\xi_i)$ for all item parameters by a separate prior $f(\xi_i | t)$ for each individual parameter. The use of these RT-based, item-dependent priors avoids the typical bias toward the location of a common prior in traditional item calibration.

A recent empirical study in van der Linden, Klein Entink and Fox (2010) sheds some light on the potential gains involved in joint parameter estimation. The study only involved manipulation of the correlation between the person parameters θ and τ in (11). Specifically, it did not yet include the covariance matrix of the item parameters in (12). The observed gains were thus entirely due to the presence of more informative posteriors for the θ parameters (see the next section) when sampling from the posterior distributions of ξ_i in the MCMC approach that was used.

Some of the results are presented in Figure 4. The upper panel shows the mean-square error (MSE) of the estimates of the difficulty parameters b_i from a sample of $N = 300$ test takers. The lower panel shows the reduction in MSE due to the use of the RTs. The average reduction for a correlation of $\rho_{\theta\tau} = .75$ was approximately 20%. The dotted curve in the lower panel is Fisher's information about b_i in the sample of test takers. It reveals that the reduction is larger than the average exactly where it is most needed, toward the two ends of the scale. This feature is due to the use of the more informative posterior updates based on the responses and RTs. Because of the extra information, the posterior distribution moved faster toward the true item difficulty parameters when they were at one of the two ends of the scale than the distributions with the updates based on the responses only.

Again, addition of the direct correlation between the item parameters in the response and RT model to the study would have led to additional – and most likely: much stronger – reduction of the MSEs of the item parameters.

3.2 Adaptive testing

The same principle of borrowing information can be used for the estimation of ability parameter θ . An attractive application is computerized adaptive testing, in which the ability estimates are updated in real time to select items optimal at the updates. RTs can be used to boost the empirical information in the ability estimates, which is particularly

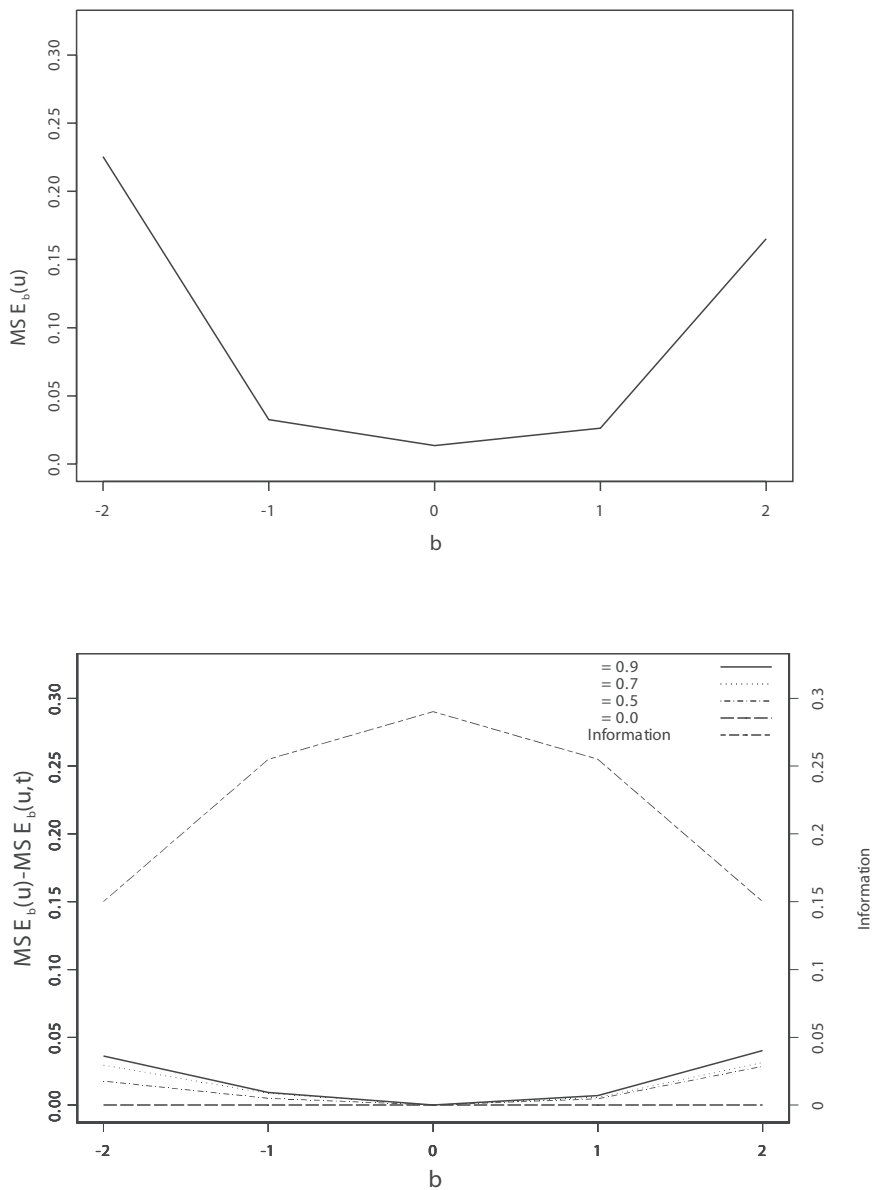


Figure 4:

MSE of the difficulty parameters b_i estimated without RTs (upper panel) and reduction in the MSE as a result of the use of the RTs (lower panel) for different levels of correlation between θ and τ (sample size $N=300$). (Reproduced with permission from W. J. van der Linden, R. H., Klein Entink, & J.-P. Fox (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, 34, 327-347.)

helpful in the beginning of the test when traditional estimates have not yet converged and the adaptive algorithm runs a substantial risk of selecting items too far away from the true ability of the test taker.

In this adaptive environment, the update of a test taker's ability estimate both from the recorded responses and RTs is relatively straightforward. Analogous to (14), the posterior predictive density of θ_j given the test taker's responses u_j and RTs t_j has density function

$$f(\theta_j | u_j, t_j) \propto f(u_j; \theta_j) f(\theta_j | t_j). \tag{16}$$

The density is calculated using the steps in Figure 5 analogously to those for the posterior distribution of the item parameters in Figure 4.

The use of logtimes instead of the RTs measured in regular units (e.g., second) yields a normal version of the model in (9), which has the normal distribution of \mathcal{P} as a conjugate distribution. For this case, the posterior predictive density $f(\theta_j | t_j)$ in (16) is also normal, and we can derive closed-form expressions for its mean and variance. For the selection of the k th item in the adaptive test, i.e., after a vector t_{k-1} with the RTs on the first $k-1$ items, they can be written as

$$\mu_{\theta | t_{k-1}} = \frac{\sigma_{\theta\tau} \sum_{i=1}^{k-1} \alpha_i^2 (\beta_i - \ln t_i)}{1 + \sigma_{\tau}^2 \sum_{i=1}^{k-1} \alpha_i^2} \tag{17}$$

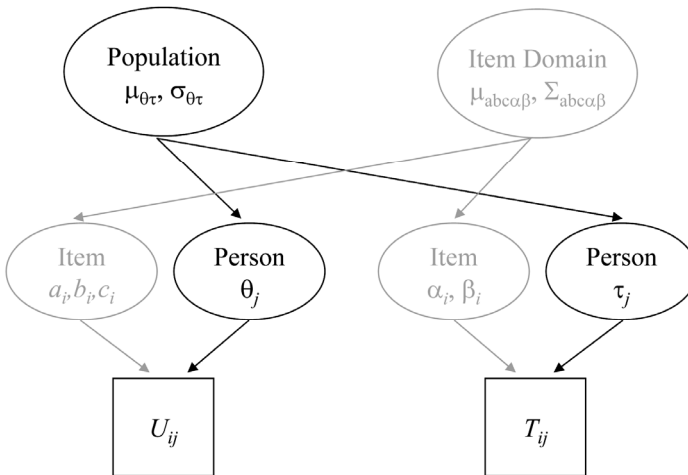


Figure 5:

Link between person parameters in the response and RT models that allows us to borrow the information in the RTs when estimating the ability parameters (and the other way around)

and

$$\sigma^2_{\theta|k-1} = 1 - \frac{\sigma^2_{\theta\tau}}{\sigma^2_\tau} + \frac{\sigma^2_{\theta\tau}}{1 + \sigma^2_\tau \sum_{i=1}^{k-1} \alpha_i^2} \tag{18}$$

(van der Linden, 2008). The only modification required is thus the replacement of the typical standard normal density used as a common prior for all test takers in adaptive testing by the normal density with the individual mean and variances in (17)–(18) after each new item.

Figure 6 gives an empirical example of the gains in the MSE of the ability estimates for a typical adaptive test after the administration of 10 and 20 items, taken from van der Linden (2008). It shows that a 10-item adaptive test with the use of RTs and a correlation

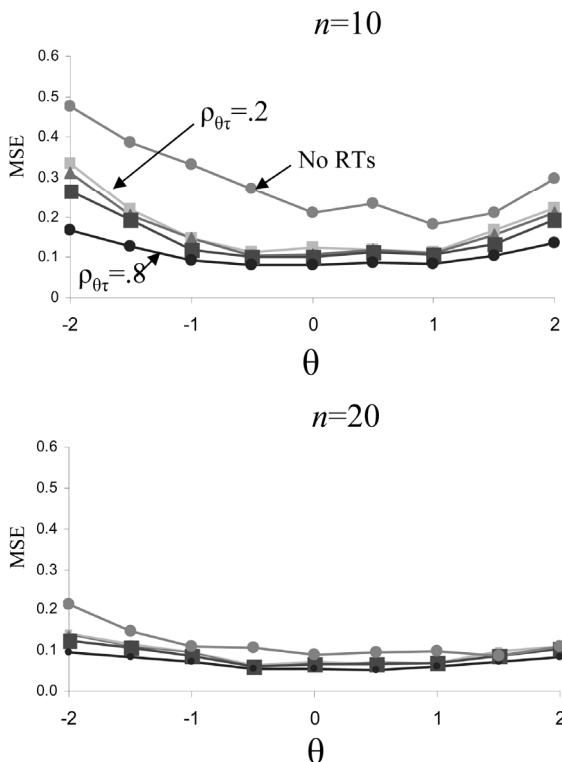


Figure 6:

MSE of the ability estimates after adaptive testing for *n*=10 (upper panel) and *n*=20 (lower panel) without and with the use of RTs ($\rho_{\theta\tau}=.2, .4, .6,$ and $.8$) (Reproduced with permission from W. J. van der Linden (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33, 5-20.)

between ability and speed equal to .6 yields similarly accurate estimates of θ as a 20-item version without the use of RTs.

3.3 Test design and speededness

A test is speeded to the degree that test takers run the risk of not completing all of its items. Before introducing a new test form, it is important to address its degree of speededness. Traditionally, the degree has been determined by pretesting a trial version of the form and determining how many test takers guess randomly at the end of the test because they run out of time. A more recent approach involves the use of mixture Rasch modeling to detect the transition from serious response behavior to random guessing (Boughton & Yamamoto, 2006; Yamamoto & Everson, 1997). The results are then used to re-assemble the form and/or adjust the time limit for the test.

The use of the RT model in (9) allows us to diagnose the speededness of a test form more directly, both at the level of the RTs on the individual items and the total time on the test. In addition, it can be used to select time limits in agreement with the intended level of speededness, or even to assemble new forms to meet an existing limit with the intended degree.

Examples of checking test forms for speededness at the level of the individual RTs are given in van der Linden, Breithaupt, Chuah, and Zhang (2007). The key quantity used in their diagnosis is the estimated residual logRT on the items, which are defined as

$$\alpha_i(\ln t_i - (\beta_i - \hat{\tau}_j)), \quad (19)$$

where the item parameters α_i and β_i are assumed to be known as a result of earlier item calibration and the estimate of the test taker's speed $\hat{\tau}_j$ is calculated from his or her RT vector. A maximum-likelihood estimate of τ_j can conveniently be calculated as

$$\hat{\tau}_j = \frac{\sum_{i=1}^n \alpha_i^2 (\beta_i - \ln \tau_j)}{\sum_{i=1}^n \alpha_i^2} \quad (20)$$

(van der Linden, 2006, Eq. 34). Except for estimation error, the residuals in (19) are asymptotically standard normally distributed. It is thus easy to check for unusual patterns across items.

An example of the analysis is given in Figure 7, which shows the mean residual RT as a function of the position of the items in one of the subtests of the computerized CPA Examination in van der Linden et al. (2007). No speededness problem due to a tight time limit was found for any of the subtests. The only noticeable trend was slightly more positive residuals early on in the test, with compensation during later items. This trend can be explained as a warming-up effect. However, the mean residual RT on the first

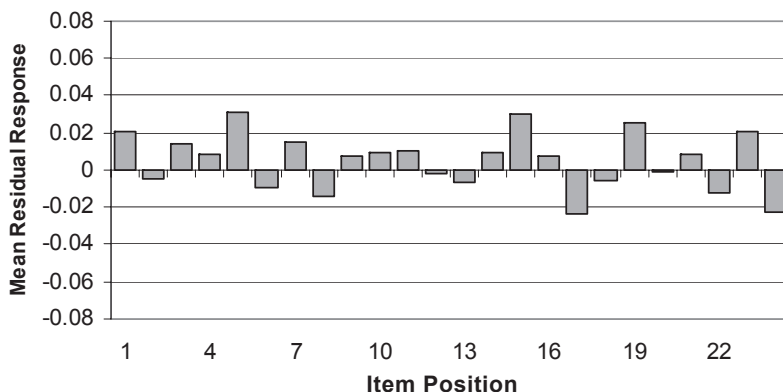


Figure 7:

Mean residual RT as a function of the position of the item in the test. (Reproduced with permission from W. J. van der Linden, K. Breithaupt, S. C. Chuah, & Y. Zhang (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, 44, 117-130.)

items was never more than 1.3 seconds per item, a size negligible relative to the total time of 50–70 seconds spent by the average test taker on them.

A key quantity to evaluate the degree of speededness of a test is an estimate of the actual probability of the test taker not completing the test in time. We use T_i to denote the RT by a test taker recorded for item $i=1, \dots, n$. The total time on the test is thus $T_{tot} = \sum_{i=1}^n T_i$. For a limit t_{lim} , the risk of not completing the test in time is the probability

$$\pi = \Pr \left\{ \sum_{i=1}^n T_i > t_{lim} \mid \tau_j, \alpha, \beta \right\}, \tag{21}$$

where α and β are the vectors with the discrimination and time-intensity parameters of the items.

The distribution of the total time T_{tot} is the n -fold convolution integral of the times T_i on the individual items. The exact distribution is unknown, and the integral is numerically intractable. However, the distribution can be approximated quite accurately by a standard lognormal with μ and σ^2 calculated from the item parameters α_i and β_i as

$$\mu = -\tau + \ln \left(\sum_{i=1}^n q_i \right) - \ln \left(\frac{\sum_{i=1}^n r_i}{\left[\sum_{i=1}^n q_i \right]^2} + 1 \right) / 2 \tag{22}$$

and

$$\sigma^2 = \ln \left(\frac{\sum_{i=1}^n r_i}{\left[\sum_{i=1}^n q_i \right]^2} + 1 \right), \tag{23}$$

where $q_i = \exp(\beta_i + \alpha_i^{-2}/2)$ and $r_i = \exp(2\beta_i + \alpha_i^{-2})[\exp(\alpha_i^{-2}) - 1]$ are new item parameters derived from α_i and β_i (van der Linden, 2011b).

For a new test form with item parameters α_i and β_i estimated during a pretest, it is thus possible to project the risk of running out of time for a critical speed level in (21) in advance, and adjust the actual time limit based on the projection. Likewise, for a given population of test takers, we can integrate (21) over τ and project the proportion of them running out of time. For a normal population, the proportion is even a function of the time limit available in closed form. Figure 8 shows the function for a normal population with the average speed centered at $\tau = 0$ and an empirical standard deviation

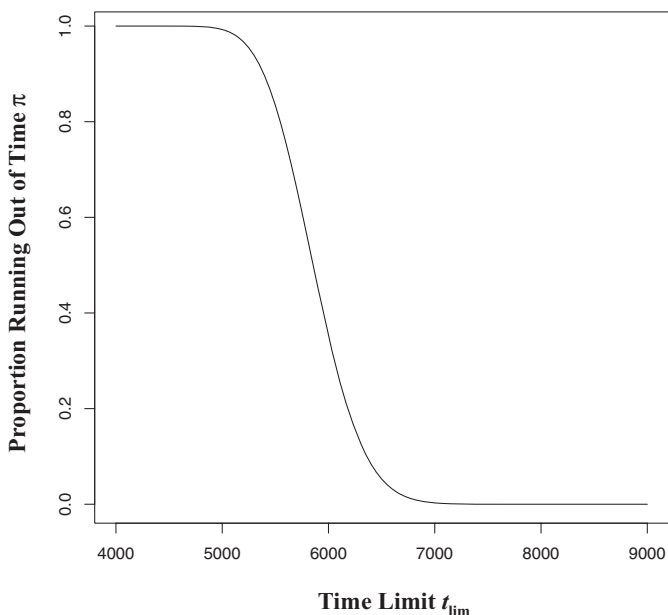


Figure 8:

Proportion of test takers from a population with speed distributions $\tau \sim N(0,242)$ running out of time as a function of the time limit on the test. (Reproduced with permission from W. J. van der Linden (2011). Setting time limits on tests. *Applied Psychological Measurement*, 35, 183-199.)

of $\sigma_r = .24$ for one of the test forms studied in van der Linden (2011b). The function enables us to select a time limit with the required degree of speededness for the population in advance. For example, if we do not want to let more than 5% of the test takers out of time, the time limit for this test should be set at approximately 6,600 seconds (or 110 minutes).

One further step is to assemble new test forms to have a predetermined risk. Note that the sums in (22) and (23) run over the item parameters q_i and r_i . In order to assemble a test with prespecified total-time distribution, we only have to control the sums of these parameters. This can easily be done, along with control of the other content specifications of the test, in an application of 0-1 linear programming to test assembly (van der Linden, 2005).

For examples of the assembly of test forms with prespecified total-time distributions for an entire range of different test-design problems, see van der Linden (2011a). One of the examples is reproduced in Figure 9. It shows a total-time distribution on a new test form nearly identical to the distribution on a reference form, which was created by matching the sums of the q_i and r_i parameters in (22) and (23) for the former to the latter. Although the two distributions are displayed for only one level of speed, they match similarly at all levels. It is interesting to know that the new form was built according to

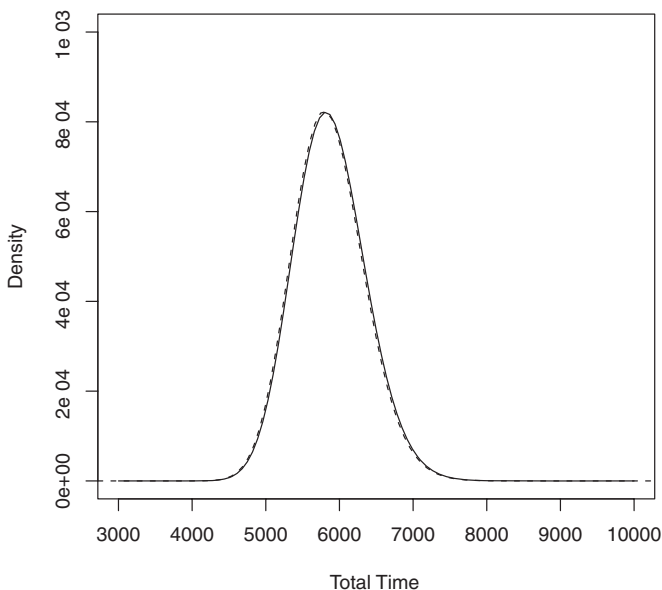


Figure 9:

Total-time distribution for a test taker working at average speed ($\tau=0$) on the reference form and the new form assembled to have matching first two cumulants. (Reproduced with permission from W. J. van der Linden (2011). Test design and speededness. *Journal of Educational Measurement*, 48, 43-59.)

the full set of specifications for an existing testing program; in addition to the two matching sums, the new form satisfied several hundreds of other constraints to deal with these specifications.

A similar method of 0-1 linear programming has been used for real-time control of speededness in adaptive testing (van der Linden, 2009b). Although speededness problems are a potential threat to the validity of fixed test forms, without any control, they are bound to hit any adaptive testing program with a fixed time limit in the form of serious differential speededness: Because each test taker gets a different selection of items, which easily differ substantially in their time intensity, unless the selection of the item is controlled with respect to their time intensity, adaptive testing will give some test takers ample time but force others to struggle with the time limit.

3.4 Detection of cheating

As shown in the previous section, irregularities in response behavior due to speededness or a warming up are easily detected analyzing RT patterns across the items in a test. An important aspect of the risk associated with computer-based testing, especially when used for such high-stakes purposes as selection, admission, and certification, are irregularities in the forms of attempts by test takers to cheat on them; for instance, stealing items by memorizing them with the intent to sell them to future candidates or exploiting knowledge about items that have already been compromised.

The traditional methods for the detection of cheating in use in the testing industry are all response based. The use of RTs instead of responses is expected to increase the power of such methods for three different reasons: First, RTs are continuous random variables instead of binary. They therefore offer much more information about the size of possible irregularities. Second, in adaptive testing, the response probabilities on the items converge to a value close to .5 toward the end of the test. As a result, irregularly looking response patterns for regularly operating test takers are not unlikely. Statistically, the effect is less power to detect various types of irregular behavior. Methods based on RTs do not suffer from the effect because items are not chosen based on their RTs rather based on item difficulty and discrimination. Finally, in a typical educational test, the time intensities of the items differ easily by a factor greater than five. As they do not know these intensities, it is difficult for test takers to hide their cheating by faking regular RT patterns.

Two different types of the use of RTs for the detection of cheating are reviewed. The first is to detect attempts by individual test takers to steal items by memorizing them or identify test takers with preknowledge of some of the items. The second deals with test takers cooperating illegally during test taking, for instance, using electronic devices to consult with each other about the correctness of their answers.

The first type of detection is based on a more sophisticated Bayesian version of the residual RTs as in (19). More specifically, the method calculates the residuals under the posterior predictive distribution of the RT on a suspicious item given the RTs on the

other items in the test. If logtimes are used, this distribution is normal with expressions for the mean and variance in closed form that follow directly from the known item parameters; for details, see van der Linden and Guo (2008).

Figure 10 shows one of the results from a case study for the Quantitative section of the Graduate Management Admission Test (GMAT) using the method. The figure consists of two superimposed plots for the same test taker for the 27 items in this section of the adaptive test, one with the observed logRTs in minutes and the other with their residual logRTs predicted from all other items in the test.

The following should be observed: First, it is important to note the large variation in the observed RTs; their (natural) logarithms ran from below one to close to five. Second, the observed logRTs and the residuals differed hugely. The differences imply that if we just checked the former for irregularities, we would be bound to make serious mistakes. Third, the residuals are standardized and exactly $N(0,1)$ distributed; we thus expect nearly all of them in the band between approximately -2 and $+2$ about zero. The pattern in Figure 10 shows one exception: a suspiciously short RT on Item 14, nearly five standard deviations below its expected value. Although the item was rather difficult relatively to the test taker's estimated ability, the response on it was nevertheless correct. Such combinations of information suggest preknowledge of the item and justify further investigation both of the status of the item and the behavior of the test taker.

The second type of detection uses suspiciously high correlation between the RTs of pairs test takers as evidence of illegitimate cooperation during the test. However, positive correlation between RTs is not only the result of cooperation but is also created by the differences in time intensities across test items: even while working fully independently,

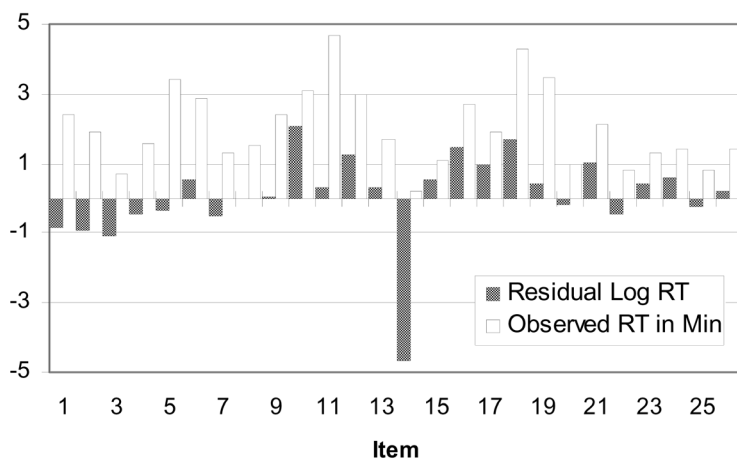


Figure 10:

Example of a RT pattern with possible preknowledge of one item. (Reproduced with permission from W. J. van der Linden & F. Guo (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73, 365-384.)

when progressing through the test, the RTs for any pair of test takers tend to go up on more time-intensive and down on less time-intensive items.

Again, the only way to adjust for these (latent) item effects is by using a model with explicit parameters for these quantities. In this case, the model we need is for the joint distribution of the RTs by two test takers on the same item. It is easy to generalize the lognormal model in (9) to a bivariate model with the same item and person parameters but an additional parameter for the correlation between the RTs by the two test takers. The presence of the time-intensity parameters for the items automatically adjusts the correlation between the RTs for spuriousness due to differences in time intensity between them. Hence, the correlation parameter directly reflects the size of suspicious agreement in RTs between test takers. For details about this bivariate model and its use for the detection cheating, see van der Linden (2009c).

4. Concluding remarks

At first sight, it may seem more complex than necessary to model RTs using latent variables and parameters. After all, we are now able to measure them to any desirable degree of accuracy during testing, so why bother about any modeling? However, just as responses, RTs are not only the result of the speed at which the test taker produces a solution but equally of the properties of the item, especially the amount of labor involved in solving them. The equation presented in (6) shows how to separate the latent effects of speed and amount of labor on the – accurately measured – RTs.

Knowing the interaction between latent item and person effects is not only theoretically important but also practically relevant. The applications reviewed in this paper were to substantiate the latter. Without separate parameters for the two different kinds of effects, it is impossible to estimate a test taker's speed from different items in adaptive testing and use the estimate to improve the ability estimation, assess the effectiveness of a time limit relative to the labor-intensities of the items, select items for fixed test forms or adaptive tests to meet a given time limit, or separate regular RTs from RTs that are suspiciously low or high. In fact, the last example in the preceding section illustrated the importance of having a model with separate latent parameters when we are interested just in the correlation between the RTs of a pair of test takers. Without such parameters, we would easily confound spurious correlation between the RTs of independently working test takers with evidence of illegitimate cooperation between them.

5. References

- Binet, A., & Simon, Th. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologie*, *11*, 191-336.
- Boughton, K., & Yamamoto, K. (2006). A hybrid model for test speededness. In M. von Davier and C. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models; Extensions and applications* (pp. 147-156). New York: Springer.

- 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.
- Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603-626.
- Goldhammer, F. (2010, May). *Application of response-time modeling: Speed in reasoning tasks and its distinctness to reasoning ability*. Paper presented at the annual meeting of the National Council on Educational Measurement, Denver, CO.
- Hornke, L. F. (2000). Response times in computerized adaptive testing. *Psicológica*, *21*, 175-189.
- Hornke, L. F. (2005). Response time in computer-aided testing: A "Verbal Memory" test for routes and maps. *Psychological Science*, *47*, 280-293.
- 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.
- Klein Entink, R. H., Kuhn, J. T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*, 54-75.
- Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621-640.
- Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. (Eds.), *The American soldier: Studies in social psychology in World War II* (Vol. 4; Chap. 10). Princeton, NJ: Princeton University Press.
- 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.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- 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.
- Schnipke, D. L., & Scrams, D. J. (1999). *Representing response time information in item banks* (LSAC Computerized Testing Report No. 97-09). Newtown, PA: Law School Admission Council.
- Tatsuoka, K. K., & Tatsuoka, M. M. (1980). A model for incorporating response-time data in the scoring of achievement tests. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 236-256). Minnesota, MN: 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. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, *16*, 433-451.
- van der Linden, W. J. (2005). *Linear models for optimal test assembly*. 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*, *72*, 287-308.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive tests. *Journal of Educational and Behavioral Statistics*, *33*, 5-20.
- van der Linden, W. J. (2009a). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272.
- van der Linden, W. J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25-41.
- van der Linden, W. J. (2009c). A bivariate lognormal model response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*, 378-394.
- van der Linden, W. J. (2010). Linking response-time parameters onto a common scale. *Journal of Educational Measurement*, *47*, 92-114.
- van der Linden, W. J. (2011a). Test design and speededness. *Journal of Educational Measurement*, *48*, 43-59.
- van der Linden, W. J. (2011b). Setting time limits on tests. *Applied Psychological Measurement*, *35*, 183-199.
- 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. (2010). Statistical tests of conditional independence between responses and response times on test items. *Psychometrika*, *75*, 120-139.
- 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. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327-347.
- Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323-339.
- Yamamoto, K., & Everson, H. T. (1995). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & J. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). New York: Waxmann.