Final Version

Joint Hierarchical Modeling of Responses and Response Times

Wim J. van der Linden & Jean-Paul Fox

Correspondence should be sent to Wim J. van der Linden, Pacific Metrics Corporation, 1 Lower Ragsdale, Building 1, Suite 150, Monterey, CA 93940. Email: wvanderlinden@pacificmetrics.com

Introduction

In spite of its apparent simplicity, the event of a test taker responding to a test item is hard to disentangle, especially if the interest is both in the observed response and the time used to produce it. For one thing, the test taker's ability to solve an item is not the only factor that plays a role; the speed at which s/he operates should also be accounted for, which immediately raises the question of how the two are related. For some tests there is more at stake for their test takers than for others, so motivation may have an impact on both as well. Similar questions arise with respect to the difficulty of the item and the amount of labor demanded by it. Besides, the nature of all these relationships is even more difficult to disentangle if the conditions change during testing; for example, due to increased fatigue toward the end of the test.

It is easy to confound or overlook the effects of some of these factors. In order to illustrate this point, Table 1 shows an extended version of the potentially confusing empirical example presented in an earlier chapter in this volume (van der Linden, vol. 1, chap. 16), which along with the first ten response times (RTs) by an arbitrary pair of test takers on a 65-item cognitive ability test now also shows their responses (Column 2–5). In addition, the table illustrates the RTs and responses on an arbitrary pair of items by the first ten test takers in the dataset (Column 7–10). Although the two test takers operated independently, their RTs appear to correlate $r = .89$. The same holds for their responses ($r = .20$). In fact, the data seem even more mysterious in that the responses by the first test taker appear to correlate with the RTs by the second ($r = .21$)! Similar patterns were observed for the pair of items in the dataset; both their response and RT vectors correlated positively ($r = .54$ and $.19$, respectively), while the responses on the first item also correlated with the RTs on the second, but this time negatively ($r = -.20$). Why do these correlations not vary just randomly about zero? How could the responses by one test taker have been impacted by the RTs of the other? And why all of a sudden the negative correlation between the responses and RTs on the two items?

[Table 1 about here]

The only way to disentangle such apparently conflicting results is by careful modeling of the probability experiment of a test taker responding to an item. An

important preliminary question, however, is if we should actually conceive of it as a single experiment or consider two distinct experiments—one for the response and the other for the RT, each driven by its own parameters. Exactly the same question bothered George Rasch when he introduced his two separate models for reading errors and reading speed (Rasch, 1960; Jansen, vol. 1, chap. 15). Both of them had a parameter ξ_p for the test taker labeled as "ability" and an item parameter δ_i labeled as "difficulty". Rasch's question was whether these parameters were on the same reading scale or, in spite of their common names and notation, actually represented two distinct abilities and difficulties. His tentative solution was a combination of the two options. As for the item parameters, he thought it "*reasonable that a text that 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 for the ability parameters he didn't expect it to be "*a general rule that a slow reader also makes many mistakes*" (Rasch, 1960, p. 42). For a further discussion of this issue, which touches on the very nature of the distinction between speed and power tests, see van der Linden (2009, pp. 250-251, 255-257; vol. 3, chap. 12).

The position we take is the one of two independent experiments each with its own test taker and item parameters, provided we can treat these parameters as fixed. The position is motivated by the long history of large-scale educational and psychological testing programs ignoring the RTs on their items, simply because the technical means necessary record them had not yet arrived. The fit of the their response models to the numerous collections of response data they did gather during these days would have been impossible if the responses were actually driven by separate RT parameters as well. But, equally important, it should be noted that the independence is assumed to hold at the level of *fixed* parameters only, analogous to the assumption of local independence made throughout IRT. Under this condition, specifying a relationship between any response and RT parameters would not make much sense. For instance, Rasch's statement that slow readers not necessarily make many mistakes is without any observational meaning for readers each operating at constant speed and ability. The only thing our two models have to do at this level is adequately represent the probability distributions of the response and the RT for each combination of test taker and item. Imposing any relationship on their speed and ability parameters would unnecessarily constrain the models and therefore

deteriorate their fit.

Continuing our second argument, in order to observe a possible relationships between ability and speed parameters, they thus have to vary over their range of possible values. They can do so in two entirely different ways: as parameters for the same test taker that change during the experiment or as parameters for different test takers that take different values. The former leads to the observation of a within-person relationship; the latter to a between-person relationship. Rasch's earlier quote on the speed of reading and the number of mistakes reminds us of the speed-accuracy tradeoff (SAT) established through extensive psychological research (Luce, 1986), which is an example of the former. It tells us that if someone decides to works faster s/he would do so at the expense of loss of accuracy (or in our current terminology: a lower effective ability). It is impossible to observe within-person relationships when test takers work at constant speed and ability. On the other hand, if the same quote is taken to refer to a between-person relationship, it would amount to an observation for a population of test takers reading at different speeds, with the slower readers tending to make fewer errors. The two different types of relationships do not imply each other at all, a point we will take up again when we introduce Figure 1 below.

The RT literature has suffered from a serious confounding of its levels of modeling—at least this is how we interpret its attempts to introduce relationships between response and RT parameters in single-level models, motivate these relationships by references to within-person or between-person phenomena, or more generally treat responses and RTs as dependent or with models that are cross-parameterized. The reason for this confounding may vary well go back to be the fact that for most of educational and psychological testing, due to memory and/or learning effects, it is generally impossible to replicate the administration of an item to a test taker. Consequently, those who would want to study relationships between responses and RTs simply by "looking at the data" typically aggregate them across different items, test takers, and/or testing conditions. But such aggregations just are the operational equivalent of confounding the different levels of modeling for the underlying experiments that need to be considered.

The hierarchical framework of modeling responses and RTs reviewed in this chapter is an attempt to disentangle all relevant levels and model each of them appropriately.

Although we make an obvious choice of models, the formal structure of the overall framework is the more important part of it. If necessary, each of our choices can be replaced by an alternative better suited to the specific application that has to be addressed. The Bayesian approach to parameter estimation and model evaluation presented below nicely complements the "plug-and-play approach" to the joint modeling of responses and RTs advocated in this chapter.

Levels of Modeling

The different levels of modeling in the framework are all empirical. A statistical level of prior distributions for the parameters at the highest level in the framework will be added later.

(1) *Fixed test takers and fixed items*. At this level, both the person and item parameters have fixed values; the only randomness exist in the responses and the RTs. Typically, multiple test takers and items are modeled jointly in this way. But, as just noticed, it is unnecessary—and potentially even dangerous—to model any functional or statistical relationship between their response and RT parameters.

(2) *Fixed items but changing test takers*. This within-person level applies when the conditions under which the test items are taken change or the test takers changes their behavior for some more autonomous reason during the test. The assumption of fixed person parameters no longer holds and a relationship between speed and ability may now manifest itself. We expect these relations to have the shape of a "psychological law," that is, a constraint on the possible combinations of values for the speed and ability parameter established by psychological research. A prime example is the SAT referred to earlier. Figure 1 below illustrates the monotonically decreasing relationship between the speed and ability parameters implied by the constraint (otherwise the shape of the curves is arbitrary).

(3) *Fixed items and a random sample from a population of fixed test takers*. The sampling creates a new level of randomness—that of a joint distribution for the speed and ability parameters in the two lower-level models. This case is typical of several large-scale educational assessments (Mazzeo, vol. 43, chap. 14). The assumption of a population distribution of the ability parameters is also made in maximum mar-

ginal likelihood estimation of the item parameters in IRT (Glas, vol. 2, chap. 11). It is hard to think of any psychological law that would define the distribution, however. As illustrated in Figure 1, possible laws at the within-person level certainly do not imply anything at the between-person level. The critical factor that moderates the two is the speed at which each individual test taker decides to operate within the range of possibilities left by his SAT—a decision likely to depend on such factors as their motivation, how they have interpreted the instructions, the time limit, strategic considerations, etc.

At the current level of modeling, the responses and RTs for the given population of test takers may be further explored as a function of possible individual-level covariates of the person parameters (e.g., general intelligence, motivation, and physical fitness). Analogous to the idea of explanatory item response theory (De Boeck & Wilson, vol. 1, chap. 33), these covariates can be incorporated as predictors in the modeling framework. Depending on their real-valued or categorical nature, the presence of the covariates introduces a regression or ANOVA structure in the population model.

(4) *Fixed items and a stratified random sample from a population of fixed test takers with a hierarchical structure.* This case arises when the test takers are nested in groups and we observe a stratified random sample of them. If group membership does have an impact on the test takers' speed or abilities, the modeling framework should adopt an ANOVA structure with nested factors. But the hierarchical structure could also force us to add a higher level to the regression model above, namely when some of its parameters appear to vary as a function of group-level covariates.

(5) *Similar levels for the items.* Each of the preceding alternative levels of modeling was specified for the test takers and their parameters only. The same classification makes sense for the items though. For instance, knowledge of the relationships between the items in a domain for a given subject area helps us to explain possible spurious correlations between responses and RTs due to their aggregation across items. Possible further dependencies between the response and RT parameters for the items could be explored by introducing item-level covariates for them (e.g., word counts, readability indices, and computational load). In the context of rule based-item generation, with

families of items each generated by identical settings of the algorithm with some minor random variation added to it, it even makes sense to introduce a hierarchical structure for their parameters (Glas, van der Linden & Geerlings, vol. 1, chap. 26; Klein Entink, Kuhn, Hornke & Fox, 2009). Generally, the benefits of treating item parameters as random instead of fixed have been underestimated (De Boeck, 2008).

[Figure 1 about here]

The framework reviewed in the next sections encompasses each of these possible levels both for the test takers and the items, with the exception of the within-person and within-item levels. Inclusion of the former would make sense for the analysis of experimental data with systematic manipulation of the test takers' speed and ability, but less so when the focus is on educational and psychological testing as in this chapter. For the latter, we refer to its treatment in the chapter by Glas, van der Linden, and Geerlings (vol. 1, chap. 26). Also, the extension of the framework to higher levels for the item domain is omitted. Our current focus will thus primarily be on the levels of fixed and randomly sampled test takers and items, with a possible individual-level regression structure to explain observed dependencies in their joint distributions. The option of test takers sampled from higher-level groups will be reviewed only briefly.

Presentation of the Model

At its lowest level, the framework has two distinct types of models for the responses and RTs. The next level consists of two models that specify the separate joint distributions of the person parameters for the population of test takers and item parameters for the domain of items, with possible individual-level explanatory variables for them. The possible extension of these distributions with a higher-level structure for the test-taker population and item domain will then be outlined. Our review of the framework draws heavily on earlier publications by Klein Entink, Fox, and van der Linden (2009), Klein Entink et al. (2009) and van der Linden (2007) (the reader should be aware of the different parameterizations for the first-level models and their impact on the higher-level parameters in the first two of these publications though).

It is instructive to follow the different types of separation and interaction between

all parameters as we introduce the candidate models. At the lowest level, the response and RT parameters are not allowed to interact. But at the second level, the perspective is rotated by 90°. Now the test taker and item parameters are treated entirely separately; only the response and RT parameters are allowed to correlate. For a graphical illustration in the form of a causal diagram, see Figure 2. The same second-level pattern of separation and interaction continues when we add higher-level covariates of the parameters to the framework.

[Figure 2 about here]

First-Level Models

Let U_{pi} and T_{pi} denote the random response and RT by test taker $p = 1, ..., P$ on item $i = 1, ..., I$.

As response model, the three-parameter logistic (3PL) model is adopted. That is, each response variable is assumed to be distributed as

$$U_{pi} \sim f(u_{pi}; \theta_p, a_i, b_i, c_i), \tag{1}$$

where $f(\cdot)$ denotes a Bernoulli probability function with success parameter

$$p_p(\theta_j) \equiv c_i + (1 - c_i)\Psi(a_i(\theta_p - b_i)), \tag{2}$$

$\theta_p \in \mathbb{R}$ is the ability parameter for test taker p, $a_i \in \mathbb{R}^+$, $b_i \in \mathbb{R}$, and $c_i \in [0, 1]$ are the discrimination, difficulty, and guessing parameters for item i, respectively, and $\Psi(\cdot)$ denotes the logistic distribution function. As the 3PL model is reviewed more completely in an earlier chapter (van der Linden, vol. 1, chap. 2), it does not need any further introduction here.

The response-time model specifies the distribution of T_{pi} as a lognormal with density function

$$T_{pi} \sim f(t_{pi}; \tau_p, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{pi}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\ln t_{pi} - (\beta_i - \tau_p))]^2\right\}, \tag{3}$$

where $\tau_p \in \mathbb{R}$ can be interpreted as a speed parameter for test taker p and $\beta_i \in \mathbb{R}$ and $\alpha_i \in \mathbb{R}^+$ as the time intensity and the discriminating power of item i, respectively. For a

full treatment of the lognormal model, including its straightforward derivation from the definition of speed by a test taker on an item, see van der Linden (vol. 1, chap. 16).

Although the two discrimination parameters in (2) and (3) have an analogous impact on the response and RT distribution (van der Linden, 2006, Figure 1; vol. 1, chap. 16) and the speed and time intensity parameters do have similar antagonistic effects on the RTs as the ability and item difficulty parameters on the responses, the two models have important formal differences as well. For instance, the lognormal model in (3) directly specifies a density function for the distribution of the RT. Its counterpart is the Bernoulli distribution in (1), of which the 3PL model explains the success parameter. Also, (2) has a guessing parameter that serves as a lower asymptote to this success parameter, whereas the RT distribution in (3) is not constrained any further than by its natural lower bound at zero.

Second-Level Models

Let $\boldsymbol{\xi}_p = (\theta_p, \tau_p)$ denote the vector with all parameters for test taker p in the response and RT model in (2)–(3) and $\boldsymbol{\psi}_i = (a_i, b_i, c_i, \alpha_i, \beta_i)$ the vector with all parameters for item i in the two models. The second-level models specify multivariate distributions of $\boldsymbol{\xi}_p$ and $\boldsymbol{\psi}_i$ for the population of test takers and the domain of items represented by the test, respectively.

An obvious choice of model for the distribution of $\boldsymbol{\xi}_p$ in a population $\mathcal{P}$ of test takers is a bivariate normal density function

$$f(\boldsymbol{\xi}_p; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) = \frac{\left|\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}\right|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}(\boldsymbol{\xi}_p - \boldsymbol{\mu}_{\mathcal{P}})^T \Sigma_{\mathcal{P}}^{-1} (\boldsymbol{\xi}_j - \boldsymbol{\mu}_{\mathcal{P}})\right], \tag{4}$$

with mean vector $\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau)$ and (2×2)-covariance matrix

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta\tau} \\ \sigma_{\theta\tau} & \sigma_\tau^2 \end{pmatrix}. \tag{5}$$

Similarly, the item-domain model specifies the distribution of parameter vector $\boldsymbol{\psi}_i$ over a domain of items $\mathcal{I}$ as a multivariate normal density function

$$f(\boldsymbol{\psi}_i; \boldsymbol{\mu}_{\mathcal{I}}, \boldsymbol{\Sigma}_{\mathcal{I}}) = \frac{\left|\boldsymbol{\Sigma}_{\mathcal{I}}^{-1}\right|^{1/2}}{(2\pi)^{5/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\psi}_i - \boldsymbol{\mu}_{\mathcal{I}})^T \Sigma_{\mathcal{I}}^{-1} (\boldsymbol{\psi}_i - \boldsymbol{\mu}_{\mathcal{I}})\right], \tag{6}$$

with mean vector $\boldsymbol{\mu}_{\mathcal{I}}$ and (5×5)-covariance matrix $\boldsymbol{\Sigma}_{\mathcal{I}}$ defined analogously to (5).

The assumption of normality make sense for an unrestricted population and item domain. However, the discrimination parameters in both first-level models and the guessing parameter in the response model have a restricted range. If the restriction appears to hamper the application, (4)–(6) could be specified for transformed versions of the parameters. Obvious choices are log transformations for the discrimination parameters and logits for the guessing parameters; for the resulting lognormal and logit normal distributions, see Casabianca and Junker (vol. 2, chap. 3).

Let $\mathbf{x}_p$ denote a $(Q+1)$-dimensional vector with the scores of test taker p on Q joint covariates of his ability and speed parameter and "1" as its first component, and assume a linear regression model could be fitted to their relationships. It then holds that

$$\theta_p = \mathbf{x}_p' \boldsymbol{\gamma}_\theta + \epsilon_{\theta_p}, \tag{7}$$

and

$$\tau_p = \mathbf{x}_p' \boldsymbol{\gamma}_\tau + \epsilon_{\tau_p}, \tag{8}$$

where $\boldsymbol{\gamma}_\theta$ and $\boldsymbol{\gamma}_\tau$ are the vectors with the regression weights and ϵ_{θ_p} and ϵ_{τ_p} the residual terms in the regression of θ and τ on $\mathbf{x}$ for test taker p, respectively. Because of random sampling of the test takers, $\epsilon_{\theta_p} \sim N(0, \sigma^2_{\epsilon_\theta})$ and $\epsilon_{\tau_p} \sim N(0, \sigma^2_{\epsilon_\tau})$, where both are taken to be independent.

Likewise, it may be possible to predict the difficulty and time intensity parameters of the items in the domain as

$$b_i = \mathbf{y}_i' \boldsymbol{\gamma}_b + \epsilon_{b_i} \tag{9}$$

and

$$\beta_i = \mathbf{y}_i' \boldsymbol{\gamma}_\beta + \epsilon_{\beta_i} \tag{10}$$

where $\boldsymbol{\gamma}_b$ and $\boldsymbol{\gamma}_\beta$ are now the weights in the regression b and β on common predictors $\mathbf{y}$ and $\epsilon_{b_i} \sim N(0, \sigma^2_{\epsilon_b})$ and $\epsilon_{\beta_i} \sim N(0, \sigma^2_{\epsilon_\beta})$ are their independent residual terms, respectively. Similar prediction of the other item parameters may be possible (although empirical research has not resulted in many serious predictors of them so far).

In the case of categorical covariates, dummy variables can be used to indicate their values. As the modification is straightforward, further details are omitted here. All covariates in (7)–(10) were taken to jointly impact the test takers' ability and speed or the difficulties and time intensities of the item, albeit with possible different strengths of effects. The changes necessary to deal with separate covariates are obvious.

The regression structure explains the structure of the mean vectors and covariance matrices of $\mathcal{P}$ and $\mathcal{I}$. For instance, for $\mathcal{P}$ it now holds that

$$\mu_\theta = \mathcal{E}_P(\boldsymbol{\gamma}'_\theta \mathbf{x} + \epsilon_\theta) = \boldsymbol{\gamma}'_\theta \mathcal{E}_P(\mathbf{x}), \tag{11}$$

$$\mu_\tau = \mathcal{E}_P(\boldsymbol{\gamma}_\tau \mathbf{x} + \epsilon_\tau) = \boldsymbol{\gamma}'_\tau \mathcal{E}_P(\mathbf{x}), \tag{12}$$

and

$$\sigma_{\theta\tau} = \text{Cov}(\boldsymbol{\gamma}'_\theta \mathbf{x} + \epsilon_\theta, \boldsymbol{\gamma}'_\beta \mathbf{x} + \epsilon_\beta) = \sum_u \sum_v \gamma_{\theta u} \gamma_{\tau_v} \text{Cov}(x_u, x_v). \tag{13}$$

Higher-Level Models

For a population of test takers nested in groups, the introduction of (joint) group-level covariates may be meaningful. Let index $g = 1, ..., G$ denote the groups in a single-level structure of $\mathcal{P}$ and $\mathbf{y}_g$ the $(R + 1)$-dimensional vector of their scores on R of these covariates (again with a "1" in its first position). Assuming linear regression, we can treat the regression parameters in (7)–(10) as

$$\gamma_{\theta g} = \mathbf{y}'_g \delta_\theta + \epsilon_{\gamma_{\theta g}} \tag{14}$$

and

$$\gamma_{\tau g} = \mathbf{y}'_g \delta_\tau + \epsilon_{\gamma_{\tau g}}, \tag{15}$$

with all parameters and residual terms defined analogously to (7) and (8). The option is not further explored here; it entirely parallels the case of multilevel response modeling addressed in Fox and Glas (vol. 2, chap. 24), where further details can be found.

Identifiability

The response and RT models in (2) and (3) are not yet identified. But, to obtain identification it suffices to set

$$\mu_\theta = 0;\ \sigma_\theta^2 = 1;\ \mu_\tau = 0. \tag{16}$$

The first two restrictions are the ones generally used in item calibration studies for the 3PL model along with the assumption of a normal population distribution of θ. The third prevents the possible tradeoff between β_i and τ_j in (3). Observe that $\mu_\tau = 0$ implies $\mu_\beta = \mathcal{E}(\ln(T))$, where the expectation is taken across all test takers and items in $\mathcal{P}$ and $\mathcal{I}$. Once the first-level models are identified, the same automatically holds for the proposed higher-level models.

It may come as a surprise that σ_τ^2 does not need to be fixed as well. But, as already indicated, although the proposed response and RT models may look to have completely analogous parameterizations at first sight, they are different: The 3PL model has latent parameter θ_p as its argument, whereas the argument of the RT model in (3) is a manifest variable t_{pi}. Consequently, discrimination parameters α_i in the RT model are automatically fixed by the unit of time in which t_{ij} is measured (van der Linden, vol. 1, chap. 16).

Alternative Plug-in Models

The key idea in this chapter is joint hierarchical modeling framework of the different types of interactions between the response and RT parameters at different levels. The individual models proposed here do make sense because of earlier successful applications. But for the case of items with a polytomous response format, the 3PL model can easily be replaced by any of the popular polytomous models, for instance, the (generalized) partial credit (Masters, vol. 1, chap. 7; Muraki & Muraki, vol. 1, chap. 8) or graded response model (Samejima, vol. 1, chap. 6). Options for models with still different response formats can be found in several of the other earlier chapters in this volume.

It is unclear how the response format of test items would impact their RT distributions. Generally, the amount of time spent on a test item seems to be more

influenced by the nature of the problem formulated in it than how the test takers are assumed to submit their answers. Nevertheless, a more flexible alternative to the lognormal model is a normal Box-Cox type of model. However, a necessary condition for its compatibility with the current modeling framework is a common Box-Cox transformation of the RTs across all items; the more general case of item-specific transformations leads to item-specific scales for the RT parameters and therefore less plausible specification of the item-domain model as a multivariate normal (Klein Entink, van der Linden, & Fox, 2009). A rich family of models for RT distributions is the linear transformation model proposed by Wang, Chang and Douglas (2013). The family includes both parametric RT densities, including those based on the Box-Cox transformation, and the option of continuous nonparametric densities that are otherwise arbitrary.

Another example of a pair of response and RT models that fits the framework in this chapter quite naturally are Rasch's models for reading error and speed (van der Linden, 2009).

As already indicated, transformation of some of the first-level parameters may lead to improved fit of the second-level models or their possible regression specifications. This option may be a simple alternative to the choice of an entirely different family of models. Besides, another way to make the current choice of second-level models more robust is omitting any of the first-level parameters in whose behavior we are not interested at the higher levels of the framework. For instance, often the interest is primarily in the correlations between the test takers' abilities and speed and/or the difficulties and time intensities of the items. These parameters have obvious interpretations and knowledge of their dependencies would definitely help testing agencies to improve their programs (for instance, nonzero correlations between the difficulty and time intensity of the items tells them they would need to control their adaptive testing programs for possible differential test speededness). Statistically, leaving out any unnecessary first-level parameters at the higher levels reduces the number of parameters that have to be estimated and increases the likelihood of model fit, without introducing any bias in the estimates of the remaining parameters.

Dependency Structure of the Data

Table 1 illustrated the rather complex dependencies that seem to arise between the responses and RTs in an arbitrary dataset. However, responses and the RTs are always nested within test takers and items. Besides, for fixed test forms, the same set of items is administered to the same set of test takers. The result of both features is a multivariate dataset with a nested, cross-classified structure. To explain observed correlations between responses and RTs for such structures, different levels of modeling are required to correctly represent their underlying dependencies.

At the lowest level, both item and test-taker parameters need to be introduced. The item parameters are required to explain the dependencies between the responses and RTs across test takers. Likewise, test-taker parameters are necessary to account for the dependencies between the responses and RTs across items. However, given both sets of parameters, we assume the responses and RTs to be independently distributed—an assumption typically referred to as the local independence assumption in IRT. In the current context of multivariate data, the assumption applies in three different ways: between the responses of each test taker, between the RTs of each test taker, and between the responses and RTs for the same test takers (Glas & van der Linden, 2009; van der Linden & Glas, 2009).

At the next higher level, both the test taker and item parameters may show different degrees of dependency; hence, our choice of free parameters for their possible covariances. The covariances between the test-taker parameters explain the observed within-person dependencies between the responses and RT across the items. The same holds for the covariances between the item parameters and the observed within-item dependencies between the responses and RTs across the test takers.

When this cross-classified, nested structure of the data is ignored, it is easy to confuse causal and spurious correlations. For instance, the observed correlation of .21 between the responses and RTs of the two test takers in Table 1 should not be interpreted to have any causal meaning. It was spurious because the covariance between the difficulties and labor intensities of the items in the dataset was ignored. Similarly, the correlation of -.20 between the responses and RTs on the two different items can be explained by a negative covariance between the test takers' speed and ability in the data set (just as in

our empirical example below).

When the observed correlations are the result of a mixture of such higher-order effects, it becomes impossible to explain them intuitively. The statistical framework reviewed in this chapter is then required to shift our attention from the observed correlations to estimates of the more fundamental covariances between item and test-taker parameters.

Parameter Estimation

The proposed procedure is Bayesian estimation of all parameters through Gibbs sampling of their joint posterior distribution. The procedure involves the division of all unknown parameters into blocks, with iterative sampling of the conditional posterior distributions of the parameters in each block given the preceding draws for the parameters in all other blocks (Fox, 2010, chap. 4; Junker, Patz, & Vanhoudnos, vol. 2, chap. 15). It is convenient to choose blocks that follow the individual models in the framework. This choice minimizes the number of changes necessary if we would replace one or more of them, and thus supports the idea of a "plug-and-play" approach proposed earlier in this chapter.

In addition to the number of levels and the specific models adopted at each of them, the nature of the Gibbs sampler also depends on possible additional choices one has made, such as reparameterization of some of the models or use of data augmentation. Depending on these choices, the likelihoods associated with some of the blocks may combine with conjugate prior distributions for their parameters, with the benefit of efficient iterative updates of the conditional posterior distributions through simple adjustments of the values of their parameter. However, in the absence of conjugate prior distributions, an obvious alternative is the use of Metropolis-Hastings (MH) steps with a well-chosen proposal density (Junker, Patz, & Vanhoudnos, vol. 2, chap. 15). Although tuning of its proposal density used to be bit of a waste of time, adaptive versions of MH sampling (Atchadé & Rosenthal, 2005; Rosenthal, 2007) have automated the job; for examples of applications based on Robbins-Monro stochastic approximation, see Cai (2010) and van der Linden and Ren (2015).

For the basic two-level framework with the standard parameterization in (1)–(6), a

convenient choice of Gibbs sampler is one with a mixture of MH steps and steps with direct sampling from conjugate posteriors in closed form. In each of these steps, the prior distribution for the first-level item and test-taker parameters is a conditional version of one of the joint second-level distributions in (4) or (6). As these distributions are specified to be multivariate normal, each of their conditional distribution is also normal with a mean and variance that follow directly from their $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$; see Casabianca and Junker (vol. 2, chap. 3, theorem 2.1) for their closed-form expressions. The prior distributions for the second-level parameters need to be specified separately, though.

Further, it is efficient to use the version of the lognormal RT model that postulates a normal distribution for the logtimes (van der Linden, vol. 1, chap. 16, eq. 8). Its choice implies a combination of a normal likelihood with a conjugate prior that yields a normal posterior for two of the three parameters in the RT model. Introductions to this normal-normal case are offered in nearly every Bayesian text (e.g., Gelman et al, 2013; Gill, 2015)

More specifically, the proposed Gibbs sampler has the following steps:

(1) Item parameters (α_i, b_i, c_i) in the 3PL model are sampled using the MH algorithm with as prior distribution their second-level conditional distribution given the item parameters (α_i, β_i) in the lognormal model.

(2) Item discrimination parameters α_i in the lognormal model are sampled using a similar MH step with as prior distribution their second-level conditional distribution given all other item parameters in the response and RT model.

(3) Time-intensity parameters β_i have a normal likelihood for the observed logtimes in combination with a normal prior distribution given all other item parameters. They are sampled directly from normal posterior distributions with means and variances that have the typical Bayesian form of the combination of data and prior parameters for the normal-normal case. Because of the identifiability restriction $\mu_\tau = 0$ in (16), the mean of the prior distribution of β_i can conveniently be set equal to $\mu_\beta = \overline{\ln t}$ (average logtime in the dataset).

(4) Ability parameters θ_p are sampled using the MH algorithm with as prior distribution their second-level conditional distribution given τ_p. In fact, for this simple bivariate case, using the restrictions $\mu_\theta = 0$ and $\sigma_\theta = 1$ in (16), the means and common

variance of these prior distributions reduce all the way to $\mu_{\theta|\tau_p} = \sigma_{\theta\tau}\tau_p/\sigma_\tau^2$ and $\sigma^2_{\theta|\tau_j} = 1 - \sigma^2_{\theta\tau}/\sigma^2_\tau$.

(5) Speed parameters τ_p have a normal likelihood and normal prior distributions given θ_p with means and a common variance that, because of the same restrictions, simplify to $\mu_{\tau|\theta_p} = \sigma_{\theta\tau}\theta_p/\sigma^2_\theta$ and $\sigma^2_{\tau|\theta_p} = \sigma^2_\tau - \sigma^2_{\theta\tau}$. Just as the time-intensity parameters, the speed parameters can thus be sampled from normal posterior distributions with means and variance that take the typical Bayesian form of the combination of data and prior parameters for the normal-normal case.

(6) Choosing a member of the multivariate normal-inverse Wishart family as prior distribution, population and item domain parameters $(\mu_\mathcal{P}, \Sigma_\mathcal{P})$ and $(\mu_\mathcal{I}, \Sigma_\mathcal{I})$ can be sampled from the same family with the typical Bayesian combination of data and prior parameters for multivariate normal data as parameters.

Alternatively, substituting the normal-ogive model with the slope-intercept parameterization for the 3PL model in (2), the adoption of data augmentation allows for the replacement of the MH steps above by more efficient sampling from normal and beta posterior distributions. Details are provided in Johnson and Albert (1999), Fox (2010), Klein Entink, Fox and van der Linden (2009), and van der Linden (2007). However, the price to be paid for the reparameterization exists in the form of loss of the current interpretation of several of the parameters in the modeling framework. For instance, the intercept parameters cannot be interpreted to represent the difficulties of the items and, consequently, covariance matrix $\boldsymbol{\Sigma}_\mathcal{I}$ in (6) no longer provides us with the correlation between item difficulties and time intensities.

If a regression structure is added to the population model, its former mean vector $\boldsymbol{\mu}_\mathcal{P} = (\mu_\theta, \mu_\tau)$ is further specified as in (11) and (12). Likewise, the specification of $\boldsymbol{\Sigma}_\mathcal{P}$ then follows from (13) and the residual variances are those in (7) and (8). For this case, all regression parameters and residual variances can be sampled using the multivariate normal-inverse Wishart specification detailed in Fox (2010, chap. 3) and Klein Entink, Fox, and van der Linden (2009). The same references should be consulted for versions of the framework with a regression structure for the item-domain model or extensions to higher-level structures.

When the items have already been calibrated under the response and/or RT model of

choice, an extremely efficient option is the use of MH steps for the item parameters with an independence sampler in the form of resampling of vectors of independent posterior draws saved for them during the calibration (for details, see vol. 1, chap. 3 and 16). The same option exists, for instance, when new item or person parameters have to be estimated but the higher-level structure can be assumed not to have changed.

Model Fit

In this section, new tools for evaluating the fit of the RT and response models are introduced following the procedure of Marianti et al. (2014). The basic idea is use of the loglikelihood statistic by Levine and Rubin (1979) in a Bayesian context to quantify the extremeness of responses and RTs under the model. In order to account for the uncertainty in each of the model parameters, a posterior probability of extremeness is computed by integrating the statistics across the prior distributions of all unknown model parameters. The approach corresponds to the prior predictive approach to hypothesis testing advocated by Box (1980). For a more complete review of possible fit analyses, see the chapters on the unidimensional logistic models (van der Linden, vol. 1, chap. 2) and the lognormal RT model (van der Linden, vol. 1, chap. 16).

Person Fit of the Response Model

Drasgow, Levine, and Williams (1985) proposed a standardized version of Levine-Rubin statistic. This standardized version has been shown to have statistical power to detect aberrant response patterns in educational testing (Karabatsos, 2003).

For two-parameter (2PL) response model, the original loglikelihood person-fit statistic is defined as

$$\begin{aligned} l_0\left(\mathbf{U}_p;\boldsymbol{\theta}_p,\mathbf{a},\mathbf{b}\right) &\equiv -\ln f\left(\mathbf{U}_p;\theta_p,\mathbf{a},\mathbf{b}\right) \\ &= -\sum_{i=1}^{I} u_{pi}\ln P(u_{pi}) + (1-u_{pi})\ln\left(1-P(u_{pi})\right), \end{aligned}$$

where $P(u_{pi}) \equiv \Pr\{U_{pi} = 1 \mid \theta_p, a_i, b_i\}$. The statistic can be standardized using the

following expressions for its mean and variance

$$\begin{aligned} E\left(l_0\left(\mathbf{U}_p;\theta_p,\mathbf{a},\mathbf{b}\right)\right) &= -\sum_{i=1}^{I} P(u_{pi})\ln P(u_{pi}) + (1-P(u_{pi}))\ln\left(1-P(u_{pi})\right) \\ Var\left(l_0\left(\mathbf{U}_p;\theta_p,\mathbf{a},\mathbf{b}\right)\right) &= \sum_{i=1}^{I} P(u_{pi})\left(1-P(u_{pi})\right)\ln\left(\frac{P(u_{pi})}{1-P(u_{pi})}\right)^2, \end{aligned}$$

respectively. The result has an approximate standard normal distribution.

The person-fit test can be adapted for use under the 3PL model by introducing a dichotomous classification variable S_{pi} that classifies a correct response to be either a random guess with probability c_i $(S_{pi}=0)$ or a response according to the two-parameter item response model $(S_{pi}=1)$. Test statistic l_0 is then defined conditionally on $S_{pi}=1$ and evaluates the extremeness of non-guessed responses. The approach thus ignores the guessed responses, which are assumed to be random with the guessing parameters as probability of success.

A Bayesian version of the loglikelihood fit statistics is defined to adjust for the uncertainty in the model parameters. The result is a test that quantifies the extremeness of each response pattern as the probability of the statistic being greater than the threshold value C associated with a (frequentist) significance level of $\alpha = .05$. The (marginal) posterior probability of the statistic being greater than the threshold is

$$\begin{aligned} \Pr\{l_0\left(\mathbf{U}_p\right) > C\} &= \int\cdots\int \Pr\{l_0\left(\mathbf{U}_p;\boldsymbol{\theta}_p,\mathbf{a},\mathbf{b}\right) > C\} p\left(\boldsymbol{\theta}_p\right) p\left(\mathbf{a},\mathbf{b}\right) d\boldsymbol{\theta}_p d\mathbf{a} d\mathbf{b} \\ &= \int\cdots\int \Phi\left(l_0\left(\mathbf{U}_p;\boldsymbol{\theta}_p,\mathbf{a},\mathbf{b}\right) > C\right) p\left(\boldsymbol{\theta}_p\right) p\left(\mathbf{a},\mathbf{b}\right) d\boldsymbol{\theta}_p d\mathbf{a} d\mathbf{b} \\ &= p_l. \end{aligned}$$

Note that the likelihood statistic is integrated directly over the prior distributions of all parameters. The test can thus be interpreted as a prior predictive test as well.

Besides, it is also possible to compute the posterior probability of an aberrant response pattern under the model for the given significance level. Let F_p^u denote a random variable that takes the value one when an observed response pattern $\mathbf{U}_p = \mathbf{u}_p$ is marked as extreme and zero otherwise,

$$F_p^u = \begin{cases} 1, & \text{if} \quad I(l_0\left(\mathbf{U}_p;\boldsymbol{\theta}_p,\mathbf{a},\mathbf{b}\right) > C), \\ 0, & \text{if} \quad I(l_0\left(\mathbf{U}_p;\boldsymbol{\theta}_p,\mathbf{a},\mathbf{b}\right) \leq C), \end{cases}$$

where $I(.)$ is the indicator function. The posterior probability of $F_p^u = 1$ is computed by integrating over the model parameters,

$$\Pr\{F_p^u = 1\} = \int \ldots \int I\left(l_0\left(\mathbf{U}_p; \boldsymbol{\theta}_p, \mathbf{a}, \mathbf{b}\right) > C\right) p\left(\boldsymbol{\theta}_p\right) p\left(\mathbf{a}, \mathbf{b}\right) d\boldsymbol{\theta}_p d\mathbf{a} d\mathbf{b}.$$

A response pattern $\mathbf{u}_p$ can be identified as extreme when F_p^u equals one, for instance, with at least $.95$ posterior probability.

Person Fit of RT Model

Analogously, the loglikelihood of the RTs can be used to define a person-fit statistic to identify aberrant RT patterns. From the RT model in (3), the loglikelihood statistic for the RT pattern of test taker p is

$$\log f\left(\mathbf{T}_p; \tau_p, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \sum_{i=1}^{I} \ln f\left(T_{pi}; \tau_p, \alpha_i, \beta_i\right).$$

We use the sum of the squared standardized residuals in the exponent of the resulting expression as person-fit statistic for the RT pattern:

$$l^t\left(\mathbf{T}_p; \tau_p, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = -\sum_{i=1}^{I} \alpha_i \left(\ln T_{pi} - \left(\beta_i - \tau_p\right)\right)^2. \tag{17}$$

Given the person and item parameters, the person-fit statistic is chi-squared distributed with I degrees of freedom. A Bayesian significance test can be defined based on

$$\Pr\{l^t\left(\mathbf{T}_p; \boldsymbol{\tau}_p, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) > C\} = \Pr\{\chi_I^2 > C\}.$$

Another classification variable can be defined to quantify the posterior probability of an extreme RT pattern given a threshold value C. Let F_p^t denote the random variable which equals one when the RT pattern is flagged as extreme and zero otherwise; that is,

$$F_p^t = \begin{cases} 1, & \text{if} \quad \Pr\{l^t\left(\mathbf{T}_p; \boldsymbol{\tau}_p, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\} > C, \\ 0, & \text{if} \quad \Pr\{l^t\left(\mathbf{T}_p; \boldsymbol{\tau}_p, \boldsymbol{\alpha}, \boldsymbol{\beta}\right)\} \leq C. \end{cases}$$

Again, an observed RT pattern $\mathbf{t}_p$ is reported as extreme when the F_p^t equals one with a least $.95$ posterior probability.

Finally, classification variables F_p^u and F_p^t can be used jointly to flag test takers for

their observed response and RT patterns. In order to do so, the joint posterior probability

$$\begin{aligned}\Pr\left(F_p^u = 1, F_p^t = 1 \mid \mathbf{t}_p, \mathbf{u}_p\right) &= \int\int \Pr\left(F_p^u = 1, F_p^t = 1 \mid \xi_p, \boldsymbol{\psi}, \mathbf{t}_p, \mathbf{u}_p\right) f\left(\xi_p, \boldsymbol{\psi}, \right) d\xi_p \, d\boldsymbol{\psi} \\ &= \int \ldots \int \Pr\left(F_p^u = 1 \mid \theta_p, \mathbf{a}, \mathbf{b}, \mathbf{u}_p\right) \Pr\left(F_p^t = 1 \mid \tau_p, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{t}_p\right) \\ &\quad \cdot f\left(\theta_p, \tau_p\right) f\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{a}, \mathbf{b}\right) d\theta_p \, d\tau_p \, d\boldsymbol{\alpha} \, d\boldsymbol{\beta} \, d\mathbf{a} \, d\mathbf{b}\end{aligned}$$

should be greater than $.95$. Observe how the local independence assumption between responses and RTs given speed and ability is used to represent the joint probability of aberrant patterns as a product of the probability of an aberrant response and of an aberrant RT pattern.

Item Fit

The Bayesian person-fit tests can easily be modified to evaluate the fit of responses and RTs to an item. The only necessary change is the definition of the loglikelihood statistic across the test takers for each item; everything else is similar. Just as the person-fit tests, the result is a more conservative test that is less likely to flag response and RT patterns on the items as aberrant because due to its accounting for the uncertainty in the model parameters.

Empirical Example

Two version of the joint model were estimated using a dataset consisting of the responses and RTs of $P = 454$ test takers on the $I = 60$ items of a certification exam. Both versions included the lognormal RT model in (3). For a reason that will become clear below, one version included the two-parameter (2PL) response model, the other the three-parameter (3PL) model in (2). The two response models were identified by restricting the population means of the ability and speed person parameters to be equal to zero and setting the product of all discrimination parameters equal to one. The parameters were estimated using the Gibbs sampler with the MH steps outlined above. As software program, a modified version of the `cirt` package (Fox, Klein Entink & van der Linden, 2007) was used to accommodate the parameterization in (3)–(6). The total of number of iterations was 5,000, with the first 1,000 iterations used for burn in. Visual inspection of

the chains showed good convergence.

Table 2 gives the mean and standard deviations of the EAP estimates of the item parameters for the two versions of the joint model. The results appeared to be quite close. The mean time-intensity parameter of 3.87 corresponds to $\exp(3.87) = 47.94$ seconds for test takers who operated at the average speed of $\tau = 0$. Basically, the only impact of the addition of the guessing parameters to the response model was a slight shift in the estimates of the difficulty parameters (from -.61 to -.47).

The correlation matrices between the speed and ability parameters for the two version of the model were equal to

$$\Sigma_{\mathcal{P}} = \begin{pmatrix} 1.00 & -.12 \\ -.13 & 1.00 \end{pmatrix}.$$

The two correlations (above and below the diagonal for the 3PL and 2PL model, respectively) were mildly negative differing only in their second decimal. Negative correlations between speed and ability are not uncommon; in fact, a similar correlation was already met in our explanation of the negative correlation between the responses and RTs for the two items in Table 1. A plausible explanation of them might be better time-management skills of the more able test takers. Consequently, when they observe a less tight time limit, they might decide to slow down exploiting the opportunity to increase their scores, whereas less able test takers may be less inclined to do so (van der Linden, 2009).

[Table 2 about here]

The two correlation matrices for the item parameters are shown in Figure 3. A comparison between them shows a minor impact of the presence of the guessing parameter on the estimates of the other model parameters only. It is especially instructive to note that the lack of impact by the guessing parameter on the correlation between the parameters in the RT model—a result that confirms our earlier observation of the possibility of a plug-and-play approach to the joint modeling of responses and RTs. Otherwise, the two patterns of correlation are typical of the results for other datasets observed by the authors. Generally, the correlation between the item difficulty and time-intensity parameters in these datasets tended to be positive and rather high while the correlations between all other item parameters were much smaller with a less predictable

pattern. Intuitively, it does make sense for more difficult items to take more time. The usually relatively high correlation between these two item features is also the reason for the differential speededness observed in adaptive testing, where more able test takers tend to get more difficult items and then run the risk of experiencing considerably more time pressure (van der Linden, vol. 3, chap. 12)

[Table 3 about here]

Figure 3 shows a plot of the estimated person-fit statistics against their Bayesian levels of significance. The upper plot represents the curves for the RT patterns, the lower plot for the response patterns under the 2PL model. The nominal significance level of $\alpha = .05$ led to 7.4% of the RT patterns flagged as aberrant for the original loglikelihood statistic, of which 7.0% were flagged with a posterior probability of at least .95. For the response patterns, the percentage flagged as extreme went down from 2.9% to 2.7% for the adopted posterior probability of at least .95. Finally, although the nominal significance level remained at .05, only 0.9% of test takers were flagged as aberrant for their joint RT and response pattern with the same posterior certainty.

[Figure 3 about here]

We also checked the residual RTs for the test taker-item combinations. For the lognormal RT model, the residuals are normally distributed given the item and test-taker parameters. Following Fox (2010, p. 247), for a posterior probability of .95, 7.4% of the residuals were found to differ more than two standard deviations from their expected value of zero.

For the choice of the 3PL model, the percentage of RT patterns flagged as aberrant remained the same but the percentage of response patterns significant at $\alpha = -.5$ decreased from 1.8% to $.2\%$ for the adopted minimum posterior probability of .95. Consequently, none of the test takers was flagged as extreme with respect to their joint RT and response pattern. The results suggest that one of main reasons of the previous flagging of model misfit for some of the test takers might have been the lack of accounting for their guessing by the 2PL model.

The item-fit statistics did not yield any significant results both for the responses and RTs . It was concluded that the 60 vectors of item responses and RTs observed in our

dataset behaved according to the model.

Discussion

Our newly gained easy access to the RTs by test takers on items may force us to reconsider nearly every theoretical question and practical application addressed in the history of testing so far. Instead of pure speculation on the impact of the test taker's speed and the time features of the items on the observed behavior by the test takers, we are now able to base our conclusions on empirical data analyzed under statistical models checked for their fit.

As for practical applications, several of those related to the issue of test speededness and the selection of time limits are reviewed in van der Linden (vol. 3, chap. 12). Other possible applications include support to the diagnosis of differential item functioning, improved analysis of item misfit, more refined assessment of the impact of test speededness on parameter linking and observed-score equating, improved standard setting procedures, and enhanced protection of item and test security.

In fact, the joint modeling of RTs and responses in this chapter offers the attractive statistical option to import the empirical collateral information collected in the RTs during testing to improve the power of all our current response-based inferences (and conversely). As an example, it allows for the use of the RTs collected during adaptive testing to improve the interim ability estimates—and hence item selection—in adaptive testing. The statistical foundation for doing so is a Bayes approach with an empirical update of the initial prior distribution after each newly observed RTs. The approach is novel in that, unlike a regular empirical Bayes approach, it updates both the test taker's likelihood and the prior distribution (van der Linden, 2008). A similar new approach to item calibration is possible (Klein Entink, van der Linden & Fox, 2010). For a discussion of the advantages of using RTs as collateral information in testing, see Ranger (2013).

References

Atchadé, Y. F., & Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli, 20*, 815-828.

Box, G. E. P. (1980). Sampling and Bayesian inference in scientific modeling and robustness. *Journal of the Royal Statistical Society A, 143*, 383-430.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307–335.

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

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. 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.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gill, J. (2015). *Bayesian methods: A social and behavioral sciences approach* (3rd ed.). Boca Raton, FL: CRC Press.

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.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 14*, 54–75.

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.

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.

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

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*, 426–452.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Ranger, J. (2013). A note on the hierarchical model of responses and response times on test of van der Linden (2007). *Psychometrika, 78*, 358–544.

Rosenthal, J. S. (2007). AMCMC: An R interface for adaptive MCMC. *Computational Statistics & Data Analysis, 51*, 5467-5470.

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. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5-20.

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

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

van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika, 80*, In press.

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*, 144–168

Table 1

Responses and response times by two test takers on the first ten items in *a cognitive ability test (Column 1–5) along with the responses and RTs on two items by the first ten test takers from the same data set (Column 6-10)*

	Response		RT			Response		RT	
Item	Test Taker		Test Taker		Test Taker	Item	Item	Item	Item
	$p = 1$	$p = 2$	$p = 1$	$p = 2$		$i = 1$	$i = 2$	$i = 1$	$i = 2$
1	0	0	22	26	1	0	1	10	14
2	0	0	19	38	2	1	1	27	56
3	1	0	40	101	3	0	1	22	40
4	1	1	43	57	4	0	0	26	101
5	1	1	27	37	5	1	1	29	42
6	0	0	21	27	6	0	1	18	8
7	1	1	45	116	7	1	0	18	37
8	0	1	23	44	8	0	1	12	36
9	1	1	14	10	9	1	0	20	51
10	1	1	47	117	10	0	1	21	22
r	.89		.20			.54		.19	
	.21					-.20			

Table 2

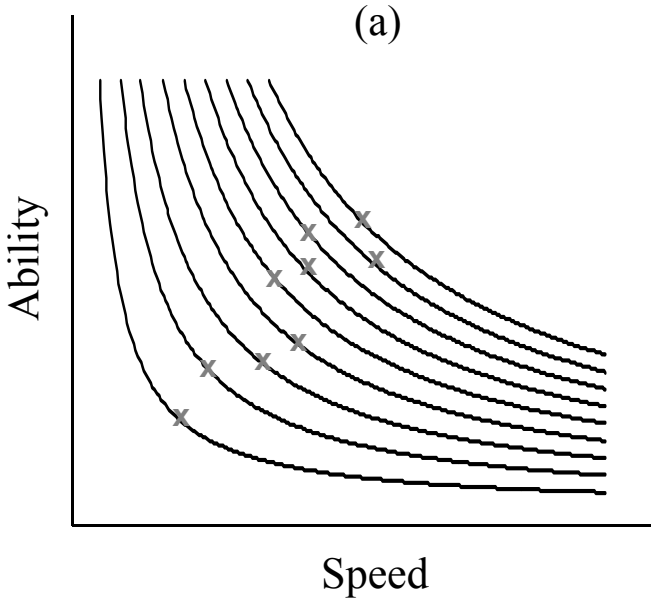
Mean and SDs of the estimated item parameters for the two versions of the joint model

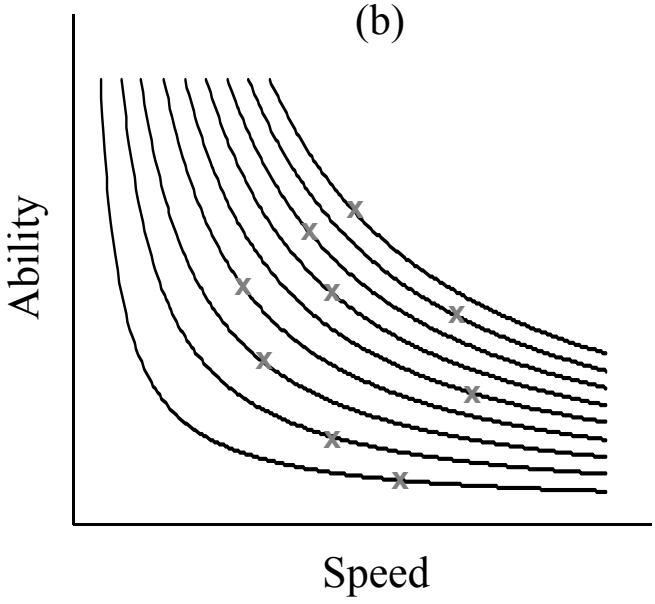
Parameter	Version with 2PL Model		Version with 3PL Model	
	Mean Estimate	SD	Mean Estimate	SD
α_i	1.85	.04	1.86	.04
β_i	3.87	.07	3.86	.07
a_i	1.06	.05	1.07	.06
b_i	-.61	.09	-.47	.10
c_i	-	-	.16	.03

Table 3

Estimated correlation matrices for the item parameters for the two versions of the joint model

	Version with 2PL Model				Version with 3PL Model			
	α_i	β_i	a_i	b_i	α_i	β_i	a_i	b_i
α_i	1.00	.20	-.20	-.09	1.00	.20	-.23	-.07
β_i		1.00	.12	.61		1.00	.28	.61
a_i			1.00	.02			1.00	.15
b_i				1.00				1.00


(a)
Ability
Speed


(b)
Ability
Speed

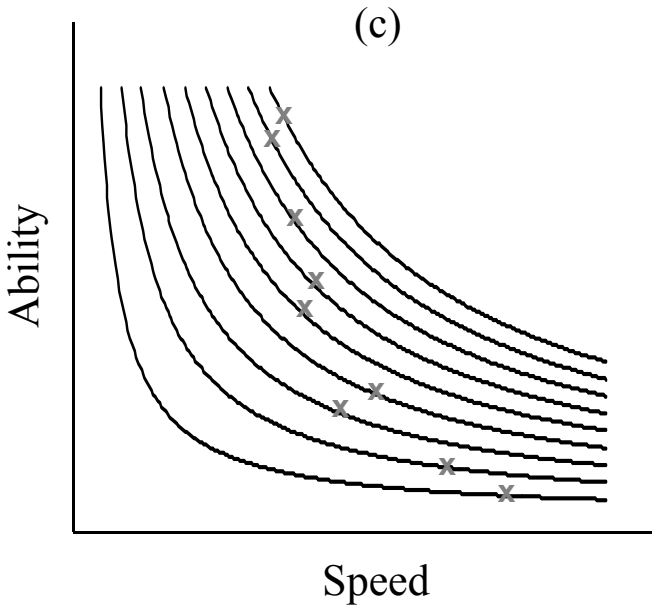
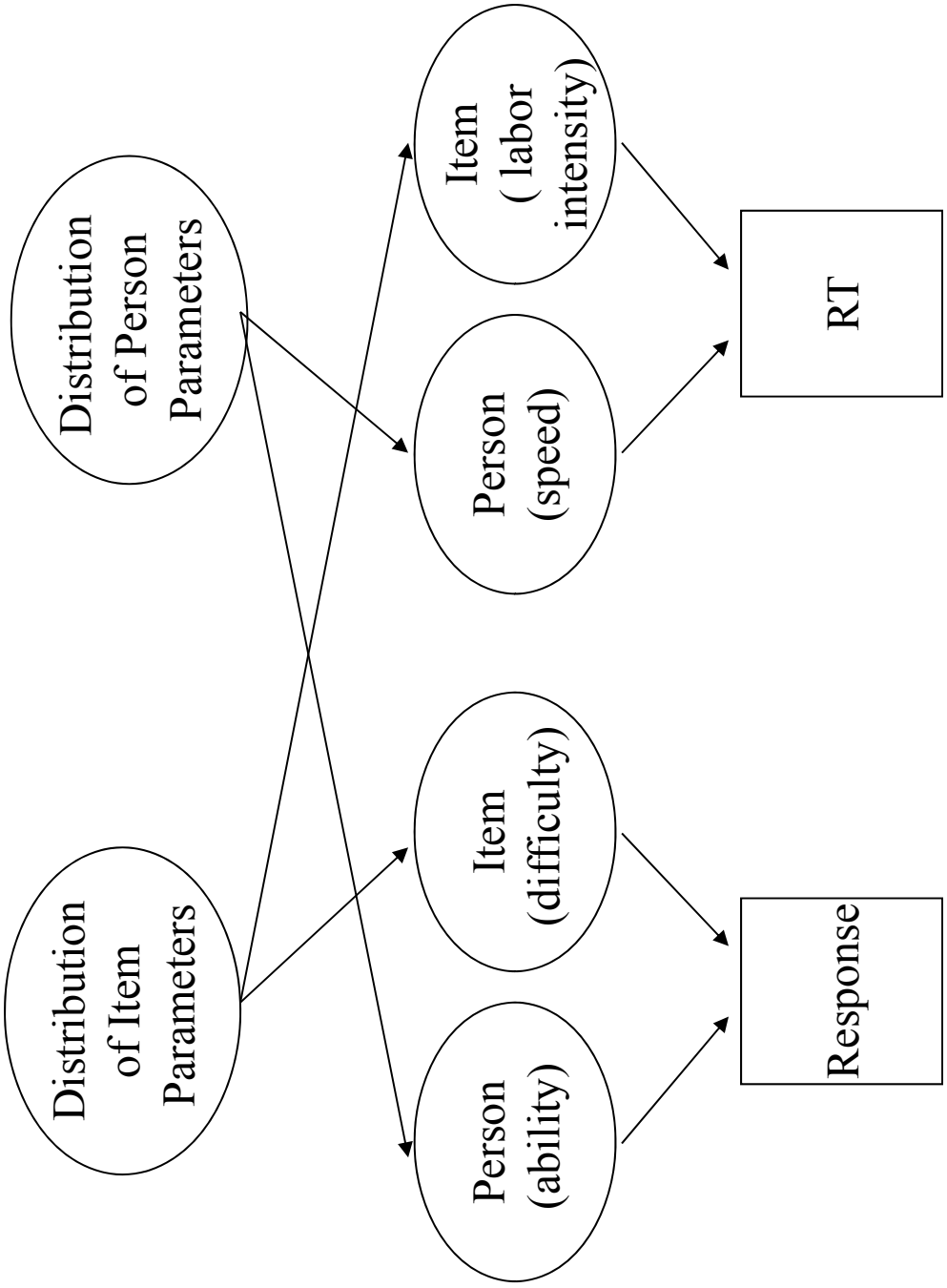
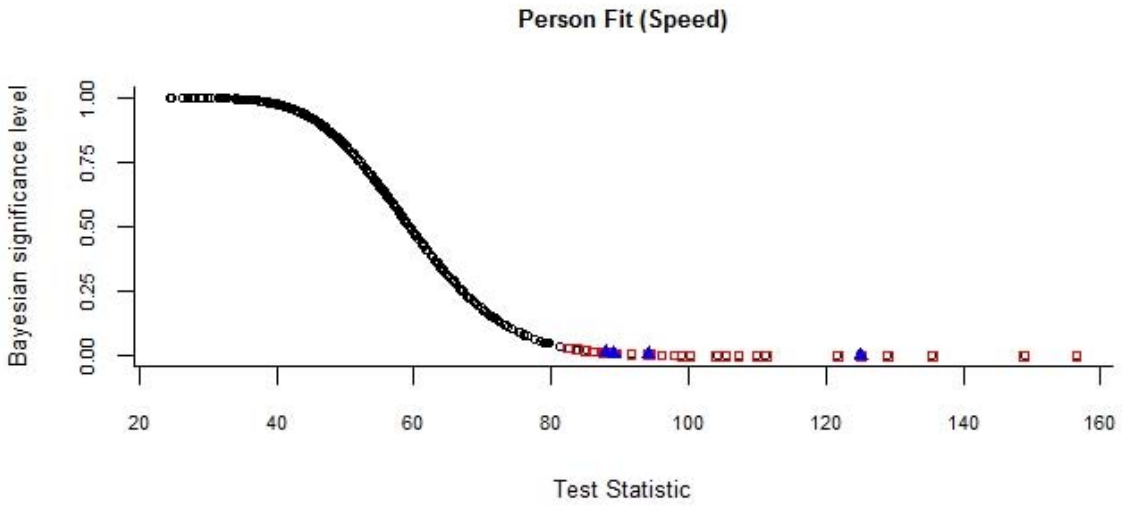
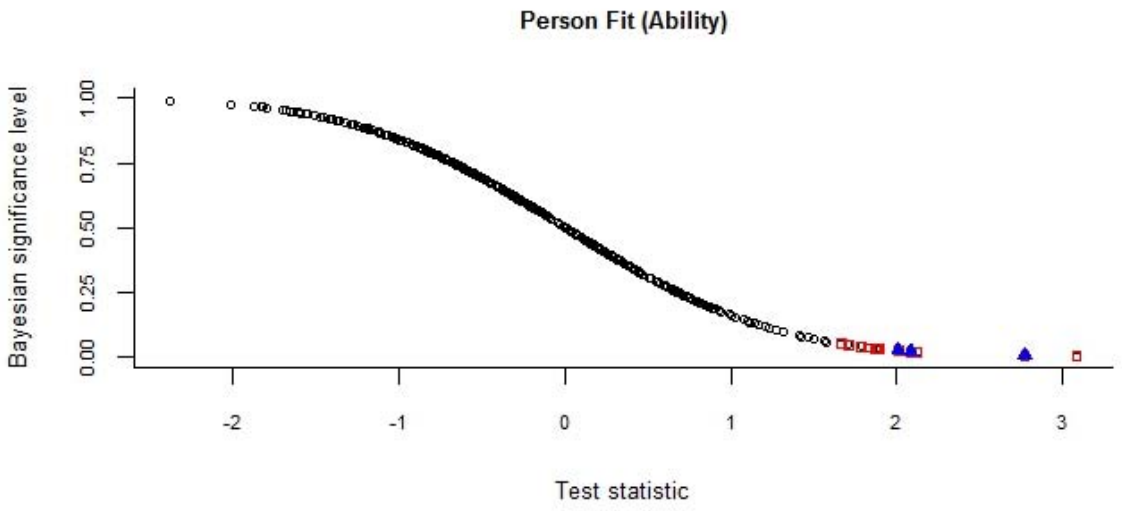


Figure 1. Three plots with the speed-accuracy tradeoff curves for the same test takers. Each test taker operates at a combination of speed and ability possible according to his personal tradeoff. The combination of choices leads to (a) positive, (b) zero and (c) negative correlation between speed and ability among the test takers. Reproduced with permission from W. J. van der Linden, (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46*, 247-272.


Distribution of Person Parameters
Distribution of Item Parameters
Item (labor intensity)
Person (speed)
Item (difficulty)
Person (ability)
RT
Response


Person Fit (Speed)
Bayesian significance level
0.00
0.25
0.50
0.75
1.00
20
40
60
80
100
120
140
160
Test Statistic


Person Fit (Ability)
Bayesian significance level
0.00
0.25
0.50
0.75
1.00
-2
-1
0
1
2
3
Test statistic