Psychological Review
2011, Vol. 118, No. 2, 339–356

0033-295X/11/$12.00 DOI: 10.1037/a0022749

Cognitive Psychology Meets Psychometric Theory: On the Relation Between Process Models for Decision Making and Latent Variable Models for Individual Differences

Han L. J. van der Maas and Dylan Molenaar
University of Amsterdam

Gunter Maris
Cito, Arnhem, The Netherlands, and University of Amsterdam

Rogier A. Kievit and Denny Borsboom
University of Amsterdam

This article analyzes latent variable models from a cognitive psychology perspective. We start by discussing work by Tuerlinckx and De Boeck (2005), who proved that a diffusion model for 2-choice response processes entails a 2-parameter logistic item response theory (IRT) model for individual differences in the response data. Following this line of reasoning, we discuss the appropriateness of IRT for measuring abilities and bipolar traits, such as pro versus contra attitudes. Surprisingly, if a diffusion model underlies the response processes, IRT models are appropriate for bipolar traits but not for ability tests. A reconsideration of the concept of ability that is appropriate for such situations leads to a new item response model for accuracy and speed based on the idea that ability has a natural zero point. The model implies fundamentally new ways to think about guessing, response speed, and person fit in IRT. We discuss the relation between this model and existing models as well as implications for psychology and psychometrics.

Keywords: ability, item response theory, diffusion model, response times, guessing

Item response theory (IRT) covers a family of measurement models for the analysis of test data, in which item responses or test scores are related to a latent variable. Specific models covered by the general IRT framework include models for dichotomous items and a continuous latent variable (Birnbaum, 1968; Lord, 1952; Mokken, 1971; Rasch, 1960), factor models for continuous items and a continuous latent variable (Jöreskog, 1971; Lawley & Maxwell, 1963; Mellenbergh, 1994), latent class models for dichotomous items and a categorical latent variable (Goodman, 1974; Lazarsfeld & Henry, 1968), and mixture models for continuous items and a categorical latent variable (Bartholomew, 1987; McLachlan & Peel, 2000).

Although these models are generally applicable, they have been especially successful in applications to psychological and educational testing, where primary attention has been devoted to the development of IRT models for dichotomous items and a continuous latent variable (e.g., see Fischer & Molenaar, 1995; Van der Linden & Hambleton, 1997). This class of models has proven to be extremely useful in test analysis, because it allows for model testing, equating, computer adaptive testing, and the investigation of differential item functioning or item bias.

In IRT models for dichotomous item responses, the probability of an item response is mathematically related to characteristics of the item and to characteristics of the respondent. For instance, in the two-parameter logistic (2PL) model (Birnbaum, 1968), the probability of a correct or affirmative response, P_+, depends on the difference between person ability (θ) and item difficulty (β), weighted by item discrimination (α) in the following way:

$$P_+ = \frac{e^{\alpha(\theta-\beta)}}{1 + e^{\alpha(\theta-\beta)}}. \quad (1)$$

Using marginal maximum likelihood, the item parameters can be estimated from a matrix of item responses, Y_{kj}, which consists of the responses of K persons to J items. The 2PL model is popular because it is more flexible than the one-parameter logistic (1PL) model (Rasch, 1960), in which all α_j are equal ($\alpha_j = \alpha$), but still gives a relatively parsimonious account of the association structure in the data. On the basis of these elementary IRT models, many more advanced IRT models have been proposed (Van der Linden & Hambleton, 1997).

IRT models, like the 1PL and 2PL models, are typically applied to item responses that result from human information processing. However, they bear no obvious connection to models that have been developed in cognitive psychology to represent the mechanisms that underlie such information processing. The Rasch model, for instance, is typically derived from statistical or measurement-theoretic assumptions (Fischer, 1995; Rasch, 1960; Roskam & Jansen, 1984). Such derivations are based on desirable properties (e.g., sufficiency of the total score for the latent variable, parameter separation, or additivity) rather than on a mathematical model of the psychological processes at play in responding to test items. IRT models are thus based on a set of assumptions concerning the relation between item responses and a set of person

This article was published Online First March 14, 2011.

Han L. J. van der Maas, Dylan Molenaar, Rogier A. Kievit, and Denny Borsboom, Department of Psychology, University of Amsterdam; Gunter Maris, Cito, Arnhem, The Netherlands, and Department of Psychology, University of Amsterdam.

Correspondence concerning this article should be addressed to Han L. J. van der Maas, UvA, Roetersstraat 15, 1018WB, Amsterdam, The Netherlands. E-mail: h.l.j.vandermaas@uva.nl

and item parameters, but these models do not address the question of how these item responses are generated. As Mislevy (2008) stated, neither "the genesis of performance nor the nature of the processes producing it are addressed in the metaphor or the attendant probability models of classical test theory or IRT" (p. 124).

It is important to emphasize that the lack of attendance to item response processes is not an inherent weakness in IRT. In many situations, item response processes are theoretically and practically intractable, and the fact that IRT models bypass assumptions concerning them may in such cases be seen as a strength rather than as a weakness. However, in investigations for which knowledge of the generating process is deemed relevant, the paucity of results that link such models to information-processing theories can become a problem. For instance, it has been argued that the primary locus for validity evidence lies in investigations that focus on the question of how a testing procedure works, that is, which processes transmit variation among individuals into variation in the item responses (Borsboom & Mellenbergh, 2007; Borsboom, Mellenbergh, & Van Heerden, 2004). It is evident that, in answering such questions, the presence of a theory that links IRT models to information-processing theories is essential. In general, such connections would seem to be invaluable for researchers who aim to augment IRT models with an explanatory component, that is, an account of how a latent variable can be conceptualized at the level of an individual person and how it may affect that person's item responses. In providing such an explanation, these accounts may also serve to bridge the gap between intraindividual process models and models for interindividual differences (Borsboom, Mellenbergh, & Van Heerden, 2003; Hamaker, Nesselroade, & Molenaar, 2007; Molenaar, 2004; van der Maas et al., 2006).

In the context of classic IRT models, which relate a continuous latent variable to a set of dichotomous item responses, Tuerlinckx and De Boeck (2005) carried out pioneering research in connecting process models to IRT. They showed that, if a diffusion process (Ratcliff, 1978) generates the item responses, then these item responses will conform to the structure of a 2PL model. In this article, we take their result as a starting point and augment it to accommodate different testing situations as they may arise in practice. Such extensions carry important consequences for the interpretation of existing models.

The structure of this article is as follows. We begin by briefly review the derivation presented by Tuerlinckx and De Boeck (2005). Then we derive three implications of their work that may be considered plausible for bipolar items (i.e., items with two *attractor* options) commonly used in attitude and personality tests but not for typical ability tests. To address this problem, we reconsider the ability concept and propose a new IRT model, the Q-diffusion model, which is applicable when respondents follow a diffusion process in deciding which answer is correct. In addition, we extend the model to accommodate for guessing and for multiple-choice items. We introduce techniques to fit the Q-diffusion model to data and illustrate these techniques with two empirical examples. Finally, we discuss the relation between the Q-diffusion model and other IRT models.

The Relation Between IRT Models and Diffusion Processes

Although the analysis of test data has mostly been the territory of psychometric models such as the 2PL, the development of formal models of human cognition has, in past decades, been mostly carried out by mathematical psychologists. The cross-fertilization of these fields has, to date, been rather meager, even though the fields evidently have much to offer to each other (Borsboom, 2006). The current work, however, is based on the conviction that (a) IRT models are ideally formal theories of item responses rather than just statistical modeling techniques and (b) such models should be based on the best formal models that mathematical psychology has to offer with respect to the cognitive processes that lie between item administration and item response.

Many item response processes used in cognitive and personality testing require the respondent to make a decision (e.g., to decide which response option is most likely to be correct, which description best fits the respondent, etc.). In the field of mathematical psychology, several models have been proposed for decision making (Busemeyer & Townsend, 1993). An important class of models applies the idea of sequential sampling of information (for a typology of these models, see Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006). In such models, noisy accumulation of information drives a decision process that stops when evidence for one of the response alternatives exceeds a threshold. The most influential model in this class of models is the diffusion model. The diffusion model is a continuous-time, continuous-state random-walk sequential sampling model (see Laming, 1968; Link, 1992; Ratcliff, 1978; Stone, 1960) that has been successfully applied to two-choice response time (RT) paradigms in studies of memory, perception, and language (see, e.g., Ratcliff, 1978, 2002; Ratcliff & Rouder, 1998; Ratcliff, Van Zandt, & McKoon, 1999).

One important reason for the popularity of the diffusion model is that it implements the sequential probability ratio test (Stone, 1960; Wald, 1947). This implies that the diffusion model optimizes expected accuracy given response time or, conversely, that it optimizes response time given the level of accuracy. Furthermore, as Bogacz et al. (2006) showed, more biologically realistic complex models with leakage, inhibition, and pooled inhibition can be reduced to the drift diffusion model and, in this way, can implement the sequential probability ratio test. Thus, the diffusion model is a widely applicable response model that may serve as a starting point for the analysis of test responses.

Figure 1 displays the basic ingredients of the diffusion model. When administered a two-choice response task, the respondent starts collecting evidence for the response options. This is formally modeled as a random walk that starts in the point z (sampled from a uniform distribution with range S_z) and stops when either of the boundaries at a or 0 is reached. Response X takes value 1 when accumulation of information terminates at bound a and takes value 0 when it terminates at the bound at 0. This termination determines the decision time (DT). Response time T is the sum of nondecision time (T_{er}), which may, for instance, cover perception of the stimulus and the time needed to execute a motor response as well as DT.

The information accumulation process that leads to DT depends on a drift rate parameter, which varies over trials with mean v and variance η. Drift rate is the mean amount of evidence accumulated over time and is thought to reflect the subject's ability for the task. Boundary separation, in contrast, is determined by the response caution of the subject, which may be influenced by instructions and rewards. If boundary separation is decreased, both DT and the probability of terminating at the correct boundary are reduced. In this way, the inverse relation between speed and accuracy (i.e., the *speed–accuracy trade-off*) is naturally accommodated in the model.

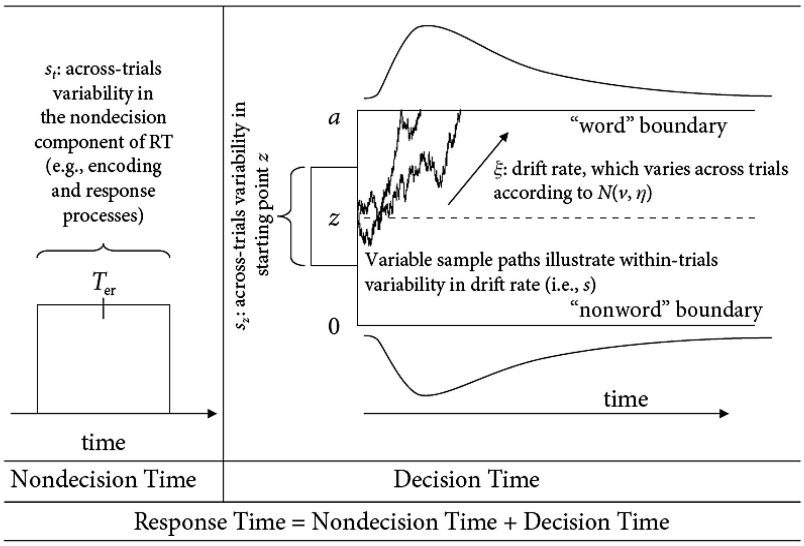

Figure 1. The drift diffusion model of two choice decisions. This example concerns the lexical decision about words and nonwords. The random walk, representing the noisy accumulation of evidence, starts at z and continues until the word or nonword boundary is hit at decision time DT. Response time RT is the sum of decision time and the time (T_{er}) required for other processes, such as the motor part of the response. Starting position z and drift rate x may vary over trials according to a uniform and a normal distribution, respectively.

The starting point z reflects the a priori bias of a participant for one or the other response. In response time modeling, this parameter is usually manipulated through payoff or proportion manipulations (Edwards, 1965). Such manipulations are generally not explicitly applied in psychological or educational testing. Variability in the accumulation of information depends on the parameter s, which is usually fixed at .1 or 1 to identify the model. Variability in the starting point and variability in drift rate over trials have been introduced in the model to account for errors faster or slower than correct responses. Further explanation of the model can be found in, for instance, Ratcliff and Rouder (1998). In this article, we use only the two most important parameters of the diffusion process: v, which denotes (mean) drift rate, and a, which denotes boundary separation. It is important to note that these two fundamental parameters, which feature in almost every sequential sampling model of choice, influence both accuracy and response time, as specified in the following equations.

The joint density of X (boundary chosen) and T (response time) has a rather complex mathematical form:

$$f_{X,T}(x,t) = \frac{\pi\sigma^2}{a^2}\exp\left(\frac{(ax-z)v}{\sigma^2} - \frac{v^2}{2\sigma^2}(t-T_{er})\right) \times \sum_{m=1}^{\infty} m \sin\left(\frac{\pi m(ax-2zx+z)}{a}\right) \times \exp\left(-\frac{1}{2}\frac{\pi^2\sigma^2m^2}{a^2}(t-T_{er})\right). \quad (2)$$

In contrast, the equations for the probability of a correct response and for the expectation of RT (i.e., of $DT + T_{er}$) are relatively easy. The probability of $X = 1$, which we denote by P_+ (i.e., terminating at the upper boundary) is as follows:

$$P_+ = P(X=1) = \frac{e^{-2zv}-1}{e^{-2av}-1} \quad (3)$$

(see Cox & Miller, 1970). In case of an unbiased decision process (i.e., $z = 1/2\ a$), this simplifies to a form very familiar to IRT modelers:

$$P_+ = \frac{e^{-av}-1}{e^{-2av}-1} = \frac{e^{av}}{1+e^{av}}. \quad (4)$$

The mean decision time can be expressed as[1]

$$E(DT) = \frac{a}{2v}\frac{1-e^{-av}}{1+e^{-av}}. \quad (5)$$

Tuerlinckx and De Boeck (2005) have connected Equation 4 to the 2PL (Equation 1). They have argued that v, the mean drift rate, can be decomposed into a person part (θ) and an item part (β), and have proposed a simple linear relation, such that $v = \theta - \beta$.[2] At first sight, this seems to be a reasonable step. However, as we show later, the resulting models, although plausible for attitude and personality tests, are highly implausible for ability tests.

Boundary separation, a, is equivalent to the discrimination parameter α in the 2PL model within the derivation of Tuerlinckx

[1] For unbiased ($z = 1/2a$) decision making, we can rewrite this to $E(DT) = (a/2v)\ (2P_+ - 1)$. The importance of the reduction in DT by the second term diminishes when P_+ takes extreme values close to zero or one. Thus, for high values of $|av|$, that is, P_+ close to zero or one, $E(DT)$ can be approximated by $|a/2v|$. The underestimation of $E(DT)$ in this approximation is not important in the qualitative analysis comparisons later. In fitting the Q-diffusion model to data, we do not apply this approximation.

[2] Throughout the article, diffusion model parameters are set in normal typeface, and IRT parameters are set in symbolic typeface.

and De Boeck (2005). This means that the discriminatory power of items depends on factors that determine boundary separation, such as instruction, rewards, and time limits, but also depends on person characteristics, such as response caution. Standard speed–accuracy trade-off studies have shown that boundary separation is influenced by time limits (Wickelgren, 1977). In practical test situations, boundary separation depends on how much time subjects get or take to answer items. This implies that increasing the time limit of items should increase the discriminatory power of items. This is an important consequence of the model that plays a key role in the present article.

Tuerlinckx and De Boeck (2005) described their work as fostering a new interpretation of item response models rather than as a derivation of item response models. Indeed, the most crucial step in their line of reasoning, which consists in equating the diffusion parameters with the 2PL parameters, is a primarily interpretative one. As we demonstrate later, alternative setups and interpretations are possible. On the other hand, in the work by Tuerlinckx and De Boeck, the 2PL is derived from assumptions about human information processing and two-choice decision tasks, which make the interpretative step quite reasonable. For instance, in view of this work, the choice for the logistic equation in the 2PL can be justified on considerations of substantive theory, and standard IRT parameters receive mechanistic interpretations in terms of one of the best information-processing models currently available. Thus, the work of Tuerlinckx and De Boeck lays out a general process model account of item response *processes* that is compatible with the standard IRT model for item response *data*. It is hard to overemphasize the psychometric importance of this result.

Implications of the Diffusion Interpretation of IRT Models for Ability Testing

If a diffusion model accurately describes the response processes that subjects follow when answering test items, this carries the implication that a standard 2PL IRT model should fit the data. However, additional implications follow from the model as well, because it yields predictions not only on the probability distribution of item responses but also on the distribution of response times. In this respect, the diffusion model makes strong predictions about the qualitative and quantitative properties of the response times (Ratcliff, 2006). Three qualitative implications are especially relevant in the current context.

A first implication that follows from the model, and one that will play an important role later in the article, is that reducing item administration time, in the limit, should yield $P_+ = 1/2$, irrespective of the value of θ. This is because, in the diffusion model, persons adjust boundary separation to handle very short time limits; as time limits become smaller, boundary separation approaches zero, and the probability of hitting either bound becomes 1/2. In IRT terms, item discrimination becomes zero, so that the item characteristic curve (ICC) becomes a flat line at $P_+ = 1/2$, equal for all levels of θ.

Clearly, this can happen only if items have two response options; otherwise, for M response options, the probability of any given response approaches $1/M$ as the time limit approaches zero. Hence, the diffusion interpretation of IRT models is not applicable to multiple-choice items with more than two response options. This is important because, even though IRT models are often applied to binary data, the dichotomies in the data result from scoring (incorrect–correct) rather than from the fact that the response process itself results from a two-choice situation. Thus, the derivation discussed earlier does not apply naturally to the ability tests for which IRT models are often used. We extend the model in this direction later in this article.

The second important implication of a diffusion interpretation of IRT models concerns the effect of changes in v (which equals $\theta - \beta$ in IRT terms). Under a diffusion model, response times are slowest when $v \approx 0$; hence, in the IRT context, response times should be slowest when $\theta = \beta$. Persons with $\theta \ll \beta$ are expected to be very fast; in fact, they should be as fast as persons with $\theta \gg \beta$. This is plausible for personality and attitude items but not for ability tests.

To see this, consider items such as *the death penalty is allowed* and *I stick to my decisions*, where *agree* and *disagree* responses are arbitrarily coded as $X = 1$ and $X = 0$. Subjects with extreme positions ($\theta \ll \beta$ or $\theta \gg \beta$) will probably answer confidently and quickly. Subjects with $\theta = \beta$ will be in doubt and are likely to respond more slowly (van der Maas, Kolstein, & van der Pligt, 2003). The latter effect is also known as the distance-difficulty hypothesis (Ferrando & Lorenzo-Seva, 2007).[3] For ability tests, this implication is extremely unlikely to hold: There is no reason to suppose that, for instance, individuals of very limited intelligence will be as fast in giving the incorrect response as highly intelligent individuals are in giving the correct response. In fact, in such cases one expects that individuals for whom the item is very hard will take longer than individuals for whom the item is easy.

A third implication concerns item discrimination, which in the diffusion context is determined by boundary separation. For positive v (i.e., whenever $\theta > \beta$), increases in boundary separation lead to increases in P_+. Because boundary separation is a function of the time limit imposed on the respondent, this means that if we allow able persons more time to think about their answer, the probability of a correct response will increase. However, the reverse is also true and considerably more surprising. For negative v (i.e., for $\theta < \beta$), allowing more time to think should *reduce* the probability of a correct response. This is illustrated in Figure 2.

This gives a special meaning to the point where $\theta = \beta$. Instead of just being the point where P_+ equals 1/2, it separates two qualitatively different regimes. Below this point, more time to solve the item will decrease P_+. Above this point, more time will increase P_+. For very long time limits, the item characteristic curve should approach that of a Guttman item (i.e., become a step function).

This may again be considered plausible in typical personality or attitude tests but not in ability tests. In personality and attitude testing, if θ is just below β and the subject has to respond quickly, the noise factor in the decision process will play a large role. However, the longer a subject thinks about his or her position, the more likely it becomes that he or she will select the answer option that best fits his or her latent state. For ability tests, however, the

[3] In fact, this phenomenon favors Tuerlinckx and De Boeck's (2005) model because their model explains this effect without any alterations to the model. The model of Ferrando and Lorenzo-Seva (2007) requires an adaptation of Thissen's (1983) model for response times to explain this effect.

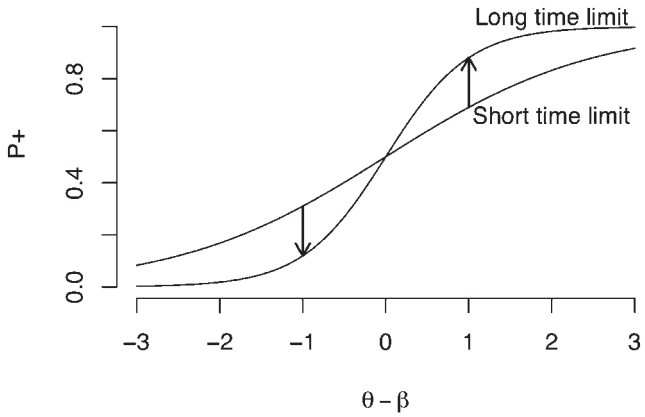

Figure 2. A longer time limit for an item will increase the probability correct (P_+) for persons with $\theta > \beta$ but will decrease this probability for persons with $\theta < \beta$.

lower end of the population cannot approach the limit of $P_+ = 0$ as a matter of principle: The worst they can do is guess, which means they approach the guessing probability of an item (0.5 for the binary case), not zero. In addition, one should expect that increasing the time limit will increase the probability of a correct response across the board.

Guessing is a widely recognized problem in IRT, where the standard solution to handle it is to introduce an extra parameter into the model. This results in the three-parameter logistic (3PL) model (Birnbaum, 1968), which is an extension of the 1PL or 2PL model: $P_+^{3PL} = c_j + (1 - c_j)\, P_+^{2PL}$ (see Equation 1). In this model, c is the lower asymptote of the item characteristic function, such that for any value of θ, $P_+ \geq c$. San Martin, del Pino, and De Boeck (2005), following Hutchinson (1991), discussed several interpretations based on a distinction in a p-process of searching for the correct answer and a g-process for guessing. One example of this line of reasoning is that a respondent first searches for the correct answer and only guesses when this search fails (assuming that the respondent recognizes this failure); however, an alternative order of processing is possible. It is not straightforward to derive this model from a diffusion model. We do not attempt such a derivation because we think a simpler and more fundamental solution is possible in terms of a single process model.

In conclusion, if a diffusion model governs the response processes in a testing situation, then the three qualitative implications just described make sense for two-choice attitude and personality tests but not for typical ability tests. We consider this to be surprising, because (a) the diffusion model would, at first glance, seem to give a reasonable approximation of the response processes in a typical ability testing situation, and (b) ability tests are the standard field of application for IRT models. What should we conclude from this?

One possible conclusion is that the 1PL and 2PL models can be applied only to personality and attitude tests and perhaps to some particular ability tests,[4] but not to the usual ability tests. This conclusion is correct but only when the diffusion model adequately describes the response process. It is debatable to what extent this is the case for typical ability tests (i.e., IQ tests). Thus, we leave open the possibility that the diffusion model does not describe the response processes in such cases but that whatever model does may imply the standard IRT model after all.

An alternative conclusion is that a diffusion model does, in fact, describe the item response processes in typical ability tests accurately. In that case, even though the standard 2PL model may be a valuable pragmatic or data-analytic tool, it cannot be considered to give an accurate theoretical account of the data structure. In this viewpoint, the suggested course of action is to investigate what kind of IRT model does follow from a diffusion account of the response process. The next paragraphs develop this line of reasoning and suggest a new IRT model that is consistent with a diffusion interpretation of the response process.

What Are Abilities?

If a sequential sampling model holds for a typical ability test item, the standard 2PL does not follow. First, the implication, that P_+ approaches 0.5 as the time limit approaches zero, cannot be correct whenever items have more than two response options; in that case, P_+ should approach the guessing probability, which is $1/M$ for equally attractive alternatives. Second, the model should be consistent with the fact that giving a subject more time will improve the probability of a correct response across all values of the latent variable. One could, of course, take these problems to be purely statistical in nature and craft solutions by considering models that evade them. However, we suspect that problems we encounter here indicate an underlying problem in the way that abilities are typically represented in psychometric theory. To see this, it is necessary to consider the deep structure of the ability concept in some detail.

In current psychometric theory, researchers commonly use the word *ability* to refer to the latent variable in a psychometric model like the 2PL. However, what such a latent variable represents is not an ability. A latent variable in a standard measurement model can represent only differences between levels of ability. That is, standard models of psychometric theory are restricted to the representation of individual differences (Borsboom et al., 2003), and the sentence "John has value θ_k on the ability measured by this test" derives its meaning exclusively from the relation between John and other test takers, real or imagined (Borsboom, Kievit, Cervone, & Hood, 2009). However, in our attempt to relate psychometric abilities to psychological processes, we are doing something entirely different from traditional approaches in psychometrics. For in the diffusion model, abilities are not merely instances of an individual differences variable. They are parameters at play in the actual process that a single individual follows when answering a test item. For this reason, we are forced to address a question that is virtually never raised in psychometric theory: What is ability at the level of an individual?

[4] Typical Piagetian conservation tests, in which children have to judge whether the amount of liquid remains the same when poured from a normal glass into a glass with a smaller diameter, could be consistent with these three implications, as (a) nonconservers typically score below chance level, because they believe the incorrect answer to be correct, (b) transitional children ($\theta = \beta$) are slower than nonconservers and conservers (van der Maas & Molenaar, 1992), and (c) the probability of a correct response is likely to decrease for nonconservers when they get more time to think over their answer.

To start with an uncontroversial and well-understood ability, consider the ability to walk. It refers to a capacity to do something, namely, to cover a certain distance by using a particular form of propulsion common to land animals. One immediate and striking feature of the ability in question is that it can be present or absent. That is, although some individuals can walk, others (e.g., babies and persons with disabilities but also fish, chairs, and elementary particles) cannot. In this sense, it is meaningful to say the ability to walk is essentially positive. We think that this characteristic is common to all abilities. In the terms of philosophers like Harré (Harré & Madden, 1975), at the individual level, to ascribe ability to an individual is to ascribe that individual a causal power. Although such powers may be present to a greater or lesser degree, they have a definite minimum, namely absence. Thus, in contrast to, say, your appraisal of a Mozart symphony or your liking of parties, abilities cannot be negative.[5]

A second observation about a simple ability, like the ability to walk, is that any task that can be said to measure this ability requires some of the ability. That is, a task that depends on the ability to walk (i.e., for which one requires the ability) depends on a positive amount of work to be done through the structures and processes that instantiate the ability. For instance, any task that depends on the ability to walk must require an individual to walk a positive distance; if the distance to be covered is zero, then the task simply cannot depend on the ability to walk (this is evident from the fact that all sorts of objects that do not have the ability in question, such as, say, the journal that you are currently holding in your hands, are able to meet the task just by staying where they are). This is also a key difference between ability testing on the one hand and personality and attitude testing on the other: To endorse an item in a personality questionnaire, or to comply with a statement in an attitude test, does not require any given level of a personality trait or attitude—one may, for instance, lie. That is, someone who walks a certain distance or solves an IQ item displays (some of) the ability in question, but someone who endorses a personality item does not. Thus, like abilities themselves, the difficulties of tasks that measure these abilities are essentially positive as well.

A third observation about simple abilities is that, if a task depends on that single ability and on no other ability, then that task can, in principle, be carried out by any individual who possesses the ability if only the individual is given sufficient time. Again, considering the ability to walk, we may identify tasks that depend only on this ability as tasks that require a person to walk a certain distance, without obstacles like rivers that require swimming or mountains that require climbing. If given enough time, any person who has the ability to walk may cover any distance unless, for some reason, that individual loses the ability along the way.

We propose that these characteristics may be considered axiomatic for simple abilities and the tasks that depend on them. It is evident, however, that this puts serious limits on what could count as a *process model* that describes abilities. In a diffusion model, for instance, it requires that drift rate is always positive and that the probability of hitting the boundary for a correct response should approach one if time limits are absent.

What happens if we assume these properties to hold and subsequently introduce individual differences into the model? Considering again the ability to walk, we may say that among individuals who have this ability, some are better at it than others. This means, in the diffusion model logic followed here, that these individuals (a) will have a higher probability of successfully completing a task (e.g., walking 100 meters) if there are time limits (which implies that some will not complete the task) and (b) will complete that task faster if there are no time limits (in this case, all individuals will complete the task). In the general situation, the time limits imposed will induce a speed–accuracy trade-off, meaning that under time limits, there will be both individual differences in the probability of completing a task and in the time in which they do so.

What kind of individual differences model follows from this line of reasoning? First, because abilities are essentially positive, it is sensible to represent them by positive numbers. Thus, the domain of θ is assumed to be R^+. Second, because all tasks require positive ability, the same holds for task difficulty, so that difficulty also has R^+ as its domain. Third, ability and difficulty should be combined in such a way that resulting drift rate is always positive. In the next section, we propose a new way to combine ability and difficulty to establish this. Fourth, if the time limit approaches infinity, the probability of a correct response should approach unity for all levels of ability. Fifth, as time pressure increases, the probability of a correct response should approach $1/M$ for M equally attractive answer options, with $P_+ = 0.5$ for the two-choice case. This implies a solution for the problem of guessing in the 1PL and 2PL model for abilities. Given that drift rate is always nonnegative, below chance scoring cannot occur. We now derive a model that has all of these properties.

The Positive Ability Model

Here, we take characteristics of ability outlined earlier as axiomatic and propose a class of models that respects these axioms. Naturally, there exists a wide class of models for which this is the case; we propose to let these models fall under the general *positive ability model*. We focus on specific subtypes of these models that may be useful in scientific research because they have clear relations with existing process and individual differences models. As before, we take the diffusion model as our starting point.

First, however, an observation about the parameters of the diffusion model is in order. Tuerlinckx and De Boeck (2005) used $\text{logit}(P_+) = av$ to derive the 2PL model; they equated the diffusion parameter a with the IRT parameter α and equated the diffusion parameter v with the linear combination of IRT parameters $\theta - \beta$. From the latter identification, it is clear that the diffusion model has no clear separation between item and person parameters, whereas in IRT this distinction is essential (van der Linden, 2009). A similar problem occurs with the a parameter, for which it is unclear whether it should be considered a person, item, or item by person parameter (in IRT, the corresponding parameter α is an item by person interaction parameter). To derive an IRT model for ability from the diffusion hypothesis, however, we need to be able to separate the person and item contributions to performance as in IRT.

[5] A standard reply of psychometricians to the requirement of nonnegative ability is that we can apply Rasch's transformation $\theta^* = e^\theta$ to scale the latent trait values to positive values. However, this does not fully solve the problem, as it still allows some θ to be less than some β, implying that subjects score below chance.

To solve this problem, we decompose drift rate and boundary separation into a person and an item part. For drift rate, Tuerlinckx and De Boeck (2005) proposed to distinguish between ability and difficulty, which we here denote by v^p and v^i, respectively; both are required to be positive, in keeping with our reconceptualization of the ability concept. For boundary separation, we propose a similar decomposition of the a parameter into a person and item part (response caution and time pressure), denoted by a^p and a^i, respectively. Response caution is then taken to be a person characteristic. Individual differences in response caution may, for instance, relate to personality traits. Time pressure depends on the setup of the test and test instructions. In a computerized test with fixed equal time limits per item, it can be equal for all items. In a test with a time limit for the whole test, time pressure may increase during the test. These and other scenarios are discussed later.

Next, we need functions $v = f(v^p,v^i)$ and $a = g(a^p,a^i)$ to combine the person and item parts into the ordinary diffusion parameters. Several constraints on these functions can be formulated. First, v and a must be positive: Boundary separation must be positive by definition, and drift rate must be positive because of the requirement that ability is positive. Note that for this reason, the difference function proposed by Tuerlinckx & De Boeck, 2005, does not work here. Second, the function f should be monotonically increasing in v^p and monotonically decreasing in v^i. Third, if v^p approaches infinity or v^i approaches zero, f should approach infinity, such that P_+ goes to 1. Fourth, if v^p approaches zero or v^i goes to infinity, f should approach zero, such that P_+ approaches 1/2.

A wide class of functions have these properties and, thus, instantiate submodels of the positive ability model. One example of a function that respects the constraints just described and that we use in the following is the quotient function, $v = v^p/v^i$.[6] We propose to apply the quotient function for both v and a, such that $v = v^p/v^i$ and $a = a^p/a^i$. In this case, v and a are always positive when v^p, v^i, a^p, and a^i are positive. Also, drift rate increases with increasing ability and decreases with increasing difficulty. For boundary separation, the quotient function works well too. Given these definitions, we arrive at the *sequential sampling based item response model for positive ability*:[7]

$$P_+ = \frac{e^{\frac{a_k^p v_k^p}{a_j^i v_j^i}}}{1 + e^{\frac{a_k^p v_k^p}{a_j^i v_j^i}}}. \tag{6}$$

This model adheres to the properties attributed to simple abilities in the previous paragraph: both ability and difficulty are positive and drift rate, $v = v^p/v^i$, is always positive, so that as the time limit becomes larger, the probability of a correct response increases for all levels of ability; the limiting case is the situation without a time limit, in which the probability of a correct response equals one. In Equation 6, participants cannot systematically score below chance level. This model has the proper ingredients to serve as a model for ability that is congruent with the hypothesis that a diffusion process generates the item responses.

The model proposed in Equation 6 is naturally applicable to two-choice tests; however, as we argued earlier, it is important to accommodate tests with multiple response options. For this purpose, we propose to apply an extension of the 2PL that covers multiple-choice options. We derive this extension from Bock's (1972) nominal response model (an IRT model for multiple-choice items with unordered categories). For M alternatives, the probability of response m in this model is as follows:

$$P_m = \frac{e^{\beta_m^* + \alpha_m^* \theta}}{\sum_{k=1}^{M} e^{\beta_k^* + \alpha_k^* \theta}}. \tag{7}$$

The parameters β_m^* (intercepts) and α_m^* (slopes) are item parameters that determine the attractiveness, $e^{b_m^* + a_m^* \theta}$, of each alternative m as a function of θ. Under the assumptions that incorrect alternatives are all equally attractive and that $m = M$ is the correct answer (i.e., by setting $\alpha_1^* \ldots \alpha_{M-1}^*$ and $\beta_1^* \ldots \beta_{M-1}^*$ to zero), it follows that

$$P_m = \frac{e^{\beta_m^* + \alpha_m^* \theta}}{(M-1)e^{0+0\theta} + e^{\beta_m^* + \alpha_m^* \theta}} = \frac{e^{\beta_m^* + a_m^* \theta}}{(M-1) + e^{\beta_m^* + \alpha_m^* \theta}}$$

$$= \frac{\frac{e^{\beta_m^* + \alpha_m^* \theta}}{e^{\ln(M-1)}}}{\frac{e^{\ln(M-1)}}{e^{\ln(M-1)}} + \frac{e^{\beta_m^* + \alpha_m^* \theta}}{e^{\ln(M-1)}}} = \frac{e^{\beta_m^* + \alpha_m^* \theta - \ln(M-1)}}{1 + e^{\beta_m^* + \alpha_m^* \theta - \ln(M-1)}}. \tag{8}$$

By setting $\alpha^* = \alpha$ and $\beta^* = -\beta\alpha$, this can be rewritten as a modified 2PL model:

$$P_+ = \frac{e^{\alpha(\theta - \beta) - \ln(M-1)}}{1 + e^{\alpha(\theta - \beta) - \ln(M-1)}} \tag{9}$$

Alternatively, applied to the positive ability model it rewrites to the following:

$$P_+ = \frac{e^{\frac{a_k^p v_k^p}{a_j^i v_j^i}}}{M - 1 + e^{\frac{a_k^p v_k^p}{a_j^i v_j^i}}} = \frac{e^{\frac{a_k^p v_k^p}{a_j^i v_j^i} - \ln(M_j - 1)}}{1 + e^{\frac{a_k^p v_k^p}{a_j^i v_j^i} - \ln(M_j - 1)}}. \tag{10}$$

Here, when ability approaches zero, the probability of a correct response does not approach 0.5, as in the original derivation of Tuerlinckx and De Boeck (2005), but approaches $1/M$, which is precisely what is desired.[8]

The model in Equation 10 still has the structure of a 2PL model. Given $\theta_k = a_k^p v_k^p; \alpha_j = 1/a_j^i v_j^i; \beta_j = \ln(M_j - 1) a_j^i v_j^i$, Equation 10 leads to Equation 1, the standard 2PL model, with the constraints that both θ and α are nonnegative. For the linear setup of

[6] A metaphorical way to think about this relation is in terms of the famous Newtonian relation speed (drift rate) = power (ability)/force (difficulty). It is interesting that Rasch (1960) proposed that the speed parameter be decomposed in his model for reading speed in reading ability/difficulty (see van der Linden, 2009, for a recent discussion).

[7] Here, superscripts p and i are used to indicate the person and item part of drift and boundary separation. Subscripts k and j indicate subject and item number.

[8] Tuerlinckx and De Boeck's (2005) derivation of the 2PL model was based on the relation $\text{logit}(P_+) = a\,v = \alpha(\theta - \beta)$. For multiple choice, we now get $a\,v = \alpha(\theta - \beta) - \ln(M - 1)$. Assuming that an increase in number of alternatives in first instance reduces the drift rate and not boundary

the 2PL, $\text{logit}(P_+) = \alpha_j^*\theta_k + \beta_j^*$ (as used in the nominal response model), β_j^* further reduces to $-\ln(M_j - 1)$ and $\alpha_j^* = \alpha_j = 1/a_j^i v_j^i$.

We propose that Model Equation 10 may be taken as a general framework to model item responses for simple abilities, under the assumption that a diffusion process gives rise to the item responses. To name this model, we link it to an earlier model developed by Ramsay (1989), which was proposed on a different basis and which he called the *quotient model* (QM):

$$P_+ = \frac{e^{\theta_k/\beta_j}}{K + e^{\theta_k/\beta_j}},\ \theta_k \geq 0,\ \beta_j > 0. \tag{11}$$

When $\theta = a^p v^p$, $\beta = a^i v^i$, and $K = M - 1$, Ramsay's QM and the model in Equation 10 are equivalent. Therefore, the model we proposed in Equation 10 can be considered a specification of Ramsay's QM on a diffusion model basis. For this reason, we propose to call the model the *Q-diffusion model*.[9]

Properties of the Q-Diffusion Model

An attractive property of the Q-diffusion model is that it explicates the meaning of the usual IRT parameters θ, α, and β. The role of the speed–accuracy trade-off in testing is clarified by the definition of ability as the product of boundary separation and drift rate. The success rate in solving items depends not only on the information powers of the subject but also on his or her response caution. We can use response times to separate these two factors. Improvements in test scores that result from increases in response caution should increase response times, whereas improvements that result from increases in drift rate should lead to lower response times. In addition, the discrimination parameter α obtains a radically new interpretation as the easiness parameter, which is the product of the time pressure and drift rate of the item. The role of β (or actually β^*) is limited to being an intercept guessing parameter for subjects with zero ability.

The predicted item response probabilities in standard IRT models are invariant under linear transformations of the θ parameter (i.e., the scale of θ has an arbitrary zero). As a result, a value of zero could mean very different things depending on the values for other subjects and the item parameters. In the Q-diffusion model, however, $\theta = 0$ always means the *absence of ability*, which implies performance at chance level. This gives the Q-diffusion model certain ratio properties: For instance, if $M = 2$ a doubling of θ gives a doubling of the logit score. A similar line of reasoning applies to the α parameter. For example, for the ability of addition, we could think of 1 + 1 as an item for which the parameter v^i is nearly zero, which implies that it has very high α. For extremely difficult items, on the other hand, α approaches zero.

The ICCs of the Q-diffusion model are simple. Because all a and v are positive, all θ and α are nonnegative. This implies that all ICCs reside in the first quadrant and all ICCs are monotonically nondecreasing. The intercept is simply $1/M$. Slopes are determined by $\alpha_j = 1/a_j^i v_j^i$. Figure 3 gives examples for $M = 2$, $M = 4$, and $M = 500$ (as an approximation to open questions).[10]

The ICCs in Figure 3 demonstrate the restrictiveness of the Q-diffusion model. Both the 2PL and the 3PL are much more flexible. The additional restrictions of the Q-diffusion model have the following sources. First, in contrast to 2PL and 3PL models, the Q-diffusion model excludes the possibility that subjects score below chance level. Second, if $M_j = M$, then the model implies that ICCs do not cross. As in Rasch's (1960) model, this restriction has both statistical and interpretive advantages, although it makes the model less flexible as a data-fitting tool; needless to say, however, the present article does not work within the tradition that views data-fitting flexibility as the prime virtue of a measurement model. Third, all ICCs start increasing at $\theta = 0$ (see Figure 3, where M is 2 and 4). In the 1PL, 2PL, and 3PL, the ICCs of difficult items first remain at zero (or chance) level and begin to increase only when θ reaches the β of the item.

Another way to demonstrate the ICCs of the Q-diffusion model is to compare them with the ICCs of the so-called *Rasch model with guessing*, which is a 3PL model with equal discrimination and fixed (at $1/M$) guessing parameters for the items. In Figure 4, we display both ICCs on the exp and log scale to ease comparison.

A note on the limitations of the Q-diffusion model is in order at this point. The structure imposed on the ICCs clearly restricts the applicability of the Q-diffusion model in certain practical applications where IRT models are routinely used today. For instance, one may solve some items of a test measuring one's proficiency in physics, but one will certainly fail some more difficult items, even when one is allowed to ponder them for the rest of one's life; such a situation violates the Q-diffusion model. This may mean that a diffusion process does not accurately describe the ability in ques-

separation, we can express the effective drift rate for multiple choice as $v^* = v - \ln(M - 1)/a$. An alternative way to incorporate multiple choice in the diffusion model is to change starting point Z to a/M. In both cases the probability of a correct guess (when $v = 0$) is, as desired, $1/M$. However, in the latter case the mean response time of an incorrect answer is much lower than the mean response time of correct answers. This is not desirable and is one of the main reasons why we prefer to correct for multiple choice by adapting the drift rate. Note that the upper boundary in this multiple-choice diffusion model represents the correct response, whereas the lower boundary corresponds to all incorrect options together. Assigning one boundary to all incorrect response options is not completely satisfactory. Ultimately, it would be preferable to derive a Q-diffusion type of model from a multiple-choice stochastic sequential sampling model for decision making.

[9] As Ramsay (1989) noted, the quotient model could also be called an exponential difference equation, when logarithmic transformation to ability and difficulty are applied. Earlier, Cressie and Holland (1983) used this model in a slightly different form:

$$P_+ = \frac{ce^{\theta_k^* - \beta_j^*}}{(1 - c) + ce^{\theta_k^* - \beta_j^*}},$$

which, given $c = 1/M$ and logarithmic transformations of θ* and β*, is again equivalent to Ramsay's QM and thus to Equation 10.

[10] This approximation can be defended because the set of possible answers to open questions is often restricted. A chess item asking for the best move in a chess position, for instance, is an open question with a limited set of legal moves (often < 100) as possible answers. However, it is well known (e.g., de Groot, 1978) that chess players, in contrast to chess computers programs, consider only a very limited set of possible moves. The initial selection of candidate moves is an important and still not well-understood part of the solution process, which is not covered in the diffusion part of the decision process.

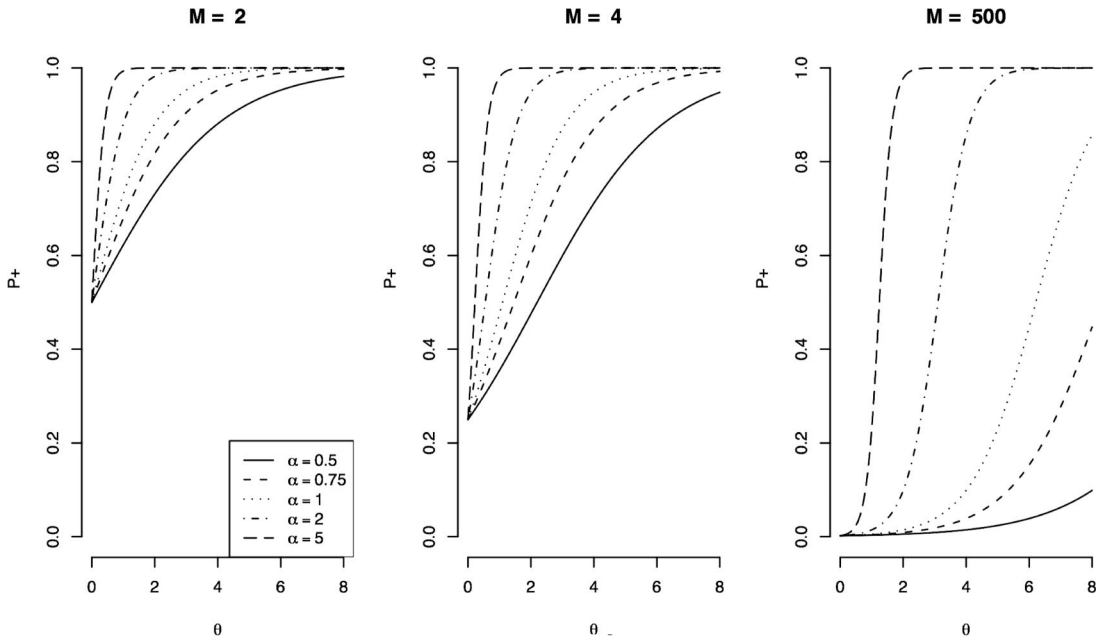

Figure 3. Item characteristic curve (ICC) of a restricted Q-diffusion model for ability testing. All θ ($v^P a^P$) and α ($1/v^i a^i$) are, by definition, positive. Intercepts are equal to $1/M$, that is, subjects with ability zero guess. Note that for high M the typical logistic form of the ICC is recovered. High M is used to model open questions.

tion, that the test does not measure a single ability, or both. In many practical testing situations, we do suspect that scores depend on a host of related and hierarchically nested abilities (van der Maas et al., 2006). Thus, in this case the Q-diffusion model may not describe the data well because the conceptualization of the test items as measuring a single ability is incorrect. Even though the item scores may approach unidimensionality in a statistical fashion, from a substantive point of view, they depend on discretely

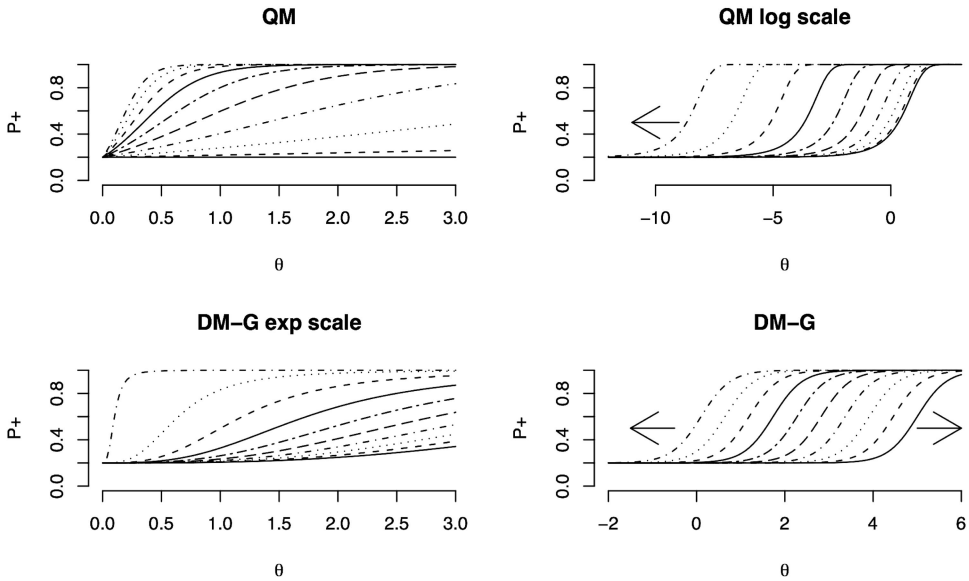

Figure 4. The quotient model (QM) compared with the Rasch (difference) model with guessing (DM-G). To ease comparison, the QM is also displayed on a log scale, and the DM-G is displayed on an exponential scale. A subtle difference between the QM (log scale) and DM-G is the asymmetry in the QM-log item characteristic curves (ICCs). They leave the lower asymptote more slowly than they approach the upper asymptote. A larger difference between the models is indicated by the arrows. DM-G ICCs can be added left and right to the current set of items. In the QM-log the item on the right is the most difficult item possible.

separable abilities. Thus, such a test may be treated with, for instance, a conjunctive multidimensional or multicomponent Q-diffusion model (de la Torre, 2009; Embretson, 1984; Hoskens & De Boeck, 2001; Maris, 1995). The construction of such a model requires further theoretical work.

Finally, as a typical advantage of the Q-diffusion model we mention the following case. Suppose we constructed a computer item bank for items with a time limit of 30 s per item. For some reason, we decide to apply this item bank in a test with a different time limit, say 1 min per item. Rescaling of the item parameters is hard to achieve within the standard interpretation of IRT parameters. We generally expect items to be more easy, but the change in difficulty probably also depends on difficulty itself. According to the diffusion interpretation of the 2PL, the answer is simple and surprising. The βs do not change at all. Only the αs change, because they increase when the time limit is relaxed.

Fitting the Q-Diffusion Model

In this paragraph, we develop statistical methodology to fit the Q-diffusion model and present two examples. First, note that the full Q-diffusion model cannot be fitted with accuracy data only, because response times are required to distinguish between the *a* and *v* part of persons and items. If response times are not available, however, we may still fit a partial Q-diffusion model to the accuracy data only (see also Ramsay, 1989).

If response times are available, different techniques developed to fit the diffusion model to data (Vandekerckhove & Tuerlinckx, 2007; Voss & Voss, 2007; Wagenmakers, van der Maas, & Grasman, 2007) are available but are not always appropriate for psychometric applications. One reason for this is that, in typical experimental psychology applications where standard techniques are used, subjects are treated as equal ($a_k^p = a^p$, $v_k^p = v^p$), and items are divided into a small number of types, say *A* and *B* ($v_j^i = v_A^i$ or $v_j^i = v_B^i$). Moreover, because items are submitted in mixed blocks, boundary separation is equal over item types ($a_j^i = a^i$). Because *M* is also known, only two parameters have to be estimated, often based on many observations per subject per item (type).

In psychometrics, however, subjects and items differ from each other, so that other constraints are required. If time pressure is equal for all items, then $a_j^i = a^i$. The α estimate then reflects differences in easiness (i.e., 1/difficulty). For many tests, such as exams, however, time pressure increases during the test, which, if not corrected for, will bias the estimates of v_j^i. We developed a Bayesian fit technique to meet the special requirements of psychometric data.

Example 1: Mental Rotation

We collected data from 121 subjects in the context of a mental rotation task (Kievit, 2010). Subjects had to identify a stimulus as either identical to or different from a rotated presentation of the same stimulus or of a different stimulus. In such a mental rotation task, item difficulty varies with rotation angle. For more details, see Borst, Kievit, Thompson, and Kosslyn (2011), who used the same experimental setup.

Responses were dichotomous (correct vs. incorrect), and response times were recorded. We randomly selected 10 out of 280 items of three different rotation angles (50^o, 100^o, 150^o) for the model-fitting analysis. We discuss the analysis at two levels that may arise in practice. First, we analyze the accuracy data only. This allows for comparison against standard IRT models, which do not have implications for the response times; in addition, these analysis can be executed with standard software. Second, we analyze the full data set, including response times. This analysis presents more difficult problems; we used a Bayesian analysis to tackle these problems. Both examples and scripts are available on the website of the first author.

Estimation without response times. Without response times we cannot distinguish between the *a* and *v* parameters of persons and items, but analyzing only the accuracy does allow for a comparison with standard IRT models. Therefore, we choose the linear setup of the 2PL formulation of Equation 10, in which $\theta_k = a_k^p v_k^p$, $\alpha_j = 1/a_j^i v_j^i$, and $\beta_j = \ln(M_j - 1)$. The resulting model may be fitted with the nonlinear mixed effect setup for IRT models (De Boeck & Wilson, 2004) through the SAS procedure NLMIXED (SAS Institute, 2000), where we attain positive values of ability parameters by exponentiation. The NLMIXED procedure provides maximum likelihood estimates with standard errors. M_j can be fixed or estimated, and information criteria, such as the Akaike information criterion and the Bayesian information criterion, can be used to compare models.

Results are represented in Table 1 and indicate that the Q-diffusion model outperforms the 1PL, the 2PL, the 3PL, and the 1PL with guessing. Both the Akaike information criterion and the Bayesian information criterion favor the Q-diffusion model.

Estimation with response times. Fitting the full Q-diffusion model requires the evaluation of rather complex mathematical functions, some involving triple integrals. Hence, we take a Bayesian approach to model fitting, which renders these functions tractable. We note that it should be possible, in principle, to implement the Q-diffusion model in a frequentist framework as well; because we use uninformative priors for the parameters in the Q-diffusion model, results are generally the same.

Table 1
Fit Statistics for the Q-Diffusion Item Response Model and Several Standard Item Response Models

Model	−2LL	AIC	BIC
	Mental rotation example		
Q-diffusion	832.2	*852.2*	*880.2*
1PL	835.1	857.1	887.9
1PL guessing	846.5	866.5	894.5
2PL	830.5	870.5	926.4
3PL full	819.2	859.2	915.1
	Chess ability example		
Q-diffusion	4,263.0	4,341.0	4,479.2
1PL	4,309.3	4,351.3	4,425.7
1PL guessing	4,214.8	4,294.8	4,436.7
2PL	4,178.4	*4,258.4*	*4,400.3*
3PL full	4,164.8	4,282.8	4,491.9

Note. Models were fitted with the accuracy data of mental rotation (Example 1) and chess ability (Example 2). The fit statistics of the best fitting models are set in italics. LL = log likelihood; AIC = Akaike information criterion; BIC = Bayesian information criterion; 1PL model = one-parameter logistic model; 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model.

The model setup is as follows. First, we assume that the binary ($M = 2$) responses are Bernoulli distributed according to Equation 6. We approximate the response time distribution with a lognormal distribution, that is, $\log(RT_{kj}) \sim \text{normal}(\mu_{kj}, \sigma^2_{kj})$. Parameters μ and σ can be obtained if the values of mean and variance are known:

$$u_{kj} = \log[E(RT_{kj})] - \frac{1}{2}\left[1 + \frac{\text{var}(RT_{kj})}{E(RT_{kj})^2}\right]$$

$$\sigma^2_{kj} = \log\left[1 + \frac{\text{var}(RT_{kj})}{E(RT_{kj})^2}\right]. \quad (12)$$

The mean and variance are derived in Wagenmakers, Grasman, and Molenaar (2005) and are functions of the Q-diffusion model parameters:

$$E(RT_{kj}) = \frac{a^p_k v^i_j}{2a^i_j v^p_k}\frac{1 - e^{h_{kj}}}{1 + e^{h_{kj}}} + Ter_k$$

$$\text{var}(RT_{kj}) = \frac{a^p_k}{2a^i_j}\left(\frac{v^i_j}{v^p_k}\right)^3\left[\frac{2h_{kj}e^{h_{kj}} - e^{2h_{kj}} + 1}{(e^{h_{kj}} + 1)^2}\right]$$

$$h_{kj} = -\frac{v^p_k a^p_k}{v^i_j a^i_j}. \quad (13)$$

In a simulation study, we established good recovery of the parameters of the full model including T^p_{er} (i.e., nondecision time varying by person).

As uninformative item priors for a^i and v^i, we use uniform distributions from .01 to .50 and from .01 to 100, respectively. As person priors for a^p, v^p, and T^p_{er}, we choose lognormal distributions with μ and σ of 0 and 1, respectively. Note that compared with the other priors, the prior of a^i appears to be relatively narrow. However, from experiences with fitting the Q-diffusion model, we encountered that a^i was always roughly in the range of .25 to .35. We drew 10,000 samples from the posterior distribution and discarded the first 5,000 as burn-in. Judged from the plots of the Markov chain Monte Carlo output, all chains converged.

The fit of the Q-diffusion model was evaluated by means of the posterior predictive distribution (Gelman, Carlin, Stern, & Rubin, 2004), which is the distribution that is predicted for the observed data given the estimated model. If the model is the true model, the posterior predictive distribution and the observed distribution are asymptotically equivalent. Figure 5 demonstrates a high degree of equivalence for the mental rotation data. Taken together, these results indicate that the Q-diffusion model is feasible and support the hypothesis that mental rotation is a simple ability.

Example 2: Chess

The mental rotation example is a theoretically interesting one but does not provide us with a strong external criterion that would allow us to examine the predictive properties of Q-diffusion model parameters. An example data set that does allow for such comparison is composed of a subset of data derived from a chess ability test (i.e., the Amsterdam Chess Test; van der Maas & Wagenmakers, 2005). In this test, chess players solve chess puzzles that require the respondent to, for instance, select the best move given a certain configuration of pieces on the board. The advantage of the chess data set is that it contains a very strong criterion measure in the form of Elo ratings (Elo, 1978). The Elo rating is based on the number of wins and losses of a given chess player and is updated on the basis of the outcomes of officially played games. Elo ratings are very good predictors of game results. Having a strong external criterion measure allows us to evaluate the fit of the Q-diffusion model in a direct way.

Some aspects of the data set are challenging. First, the item format of these chess puzzles was open ended. Because the number of legitimate sensible moves in chess is limited, we can interpret the item format as a multiple-choice format with an unknown number of options. For this reason, we apply Equation 7 to the binary scored responses. The correction for guessing used in that equation, however, cannot easily be applied to response times (see Equation 5). We followed the line of reasoning given in footnote 8, assuming that an increase in alternatives primarily increases v^i, the item drift rate or difficulty of the item. The corrected item drift rate v^{i*} is then a function of M^i and all other person and item parameters, where v^i represents the item drift rate for two alternatives:

$$a^{p^*}_k = a^p_k;\ v^{p^*}_k = v^p_k;\ a^{i^*}_j = a^i_j$$

$$v^{i^*}_j = \frac{a^p_k v^p_k v^i_j}{a^p_k v^p_k - \ln(M^i - 1)a^i_j v^i_j}. \quad (14)$$

These transformed parameters are then used to compute $E(RT)$ and $var(RT)$ according to Equation 12. We were able to implement this setup in Winbugs and applied it to this example[11], estimating M as a separate parameter for each item.

An additional problem is that chess-playing ability in general consists of many different abilities, as well as specific bits of knowledge. For certain types of items, such as endgame puzzles, one needs to know specific solution rules. If not, one will certainly fail such items, even when one is allowed lots of time. For these items, the Q-diffusion model cannot be correct. However, some subtests of the chess test consist of so-called tactical items in which knowledge is relatively unimportant. We limit our analyses to 20 chess items (the 20 tactical items of the Choose-a-Move Task, Test A, from the Amsterdam Chess Test; van der Maas & Wagenmakers, 2005) and analyze the data with the full Q-diffusion model according to the Bayesian setup described earlier.

In the model-fitting procedure, we used a uniform distribution ranging from 1.01 to 500 for M. Other priors were equal to those applied in the mental rotation example. We drew 10,000 samples from the posterior distribution and discarded the first 5,000 as burn-in. Judging from the plots of the Markov chain Monte Carlo output, all chains converged. For all items, the posterior predictive distributions and the observed distribution are similar.

Table 2 displays the relations of Elo, tournament ratings, and age with sum scores, mean response times, person drift rates, person response cautions, and nondecision time. Using drift rate

[11] When $M > 2$, equations get numerically too complex for the standard WinBUGS program (i.e., sampling proceeds extremely slow or is not possible at all). Therefore, when $M > 2$, we implemented the Q-diffusion model in the WinBUGS Development Interface (WBDev; Lunn, 2003), which is a freely available WinBUGS add-on program.

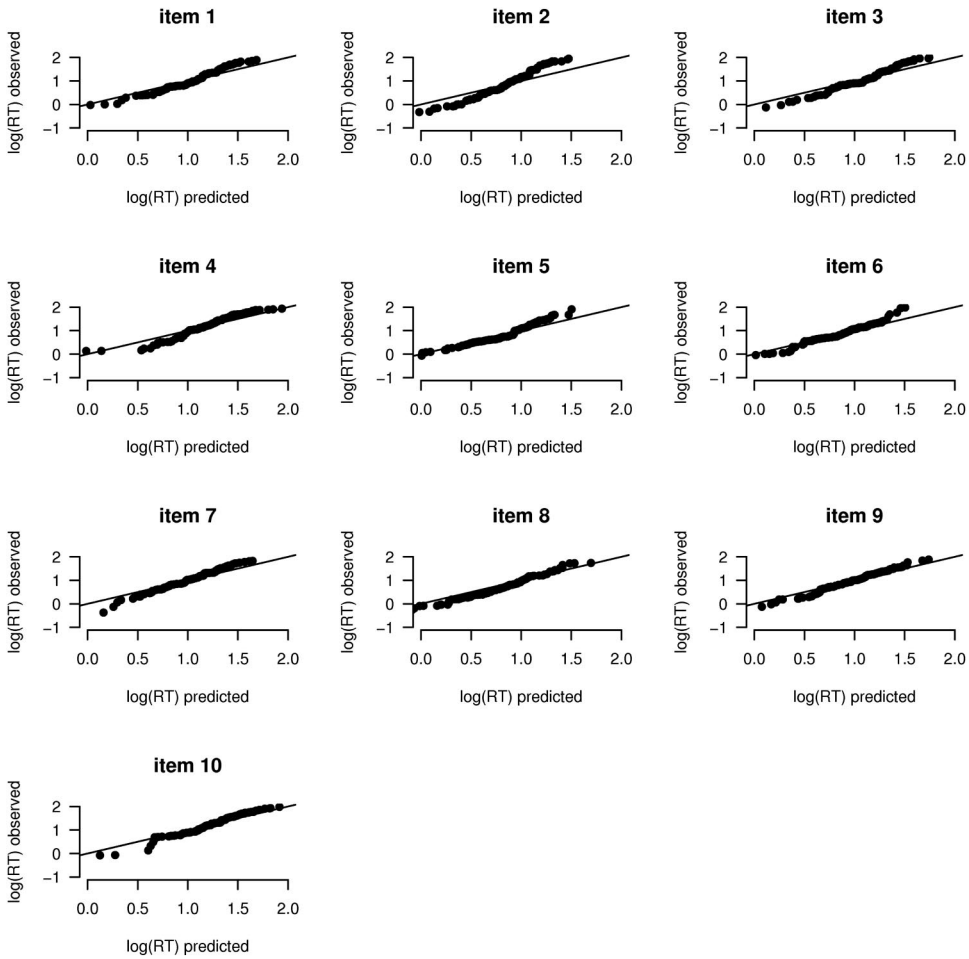

Figure 5. Quantile–quantile plot comparing the predicted quantiles of the log(*RT*) distribution according to the Q-diffusion model (*x*-axis) to the observed quantiles of the log(*RT*) distribution (*y*-axis).

instead of test score improves validity slightly. Hence, the use of response times appears to be beneficial. However, it is prudent to note that, with the accuracy data only, the 2PL model provided a better fit to the data than the Q-diffusion model. Hence, the evidence that we are dealing with a simple ability here is less convincing than in the mental rotation example. Note that age correlates most highly with T_{er}, a result also found in other applications of diffusion modeling (e.g., Ratcliff, Tapar, & McKoon, 2001).

Relation to Other IRT Models

Historically, there have been many attempts to incorporate response times into IRT models. This is an ongoing research topic. In addition, there are several relatively novel IRT models that handle response time in a new way (van der Linden, 2009). Also, different models have been suggested for guessing. In this section, we compare the Q-diffusion model with such models.

Table 2
Correlations of the Standard Test Statistics, Person Estimates According to the 1PL and 2PL Models, and the Q-Diffusion Parameters Person Drift Rate (v), Response Caution (a), and Nondecision Time (T_{er}), With the Elo Ratings and Ages of Chess Players

Person	Test score	Response time	1PL θ	2PL θ	v^p	a^p	T_{er}
Elo rating	0.68	−0.44	0.67	0.69	0.72	−0.38	−0.17
Age	−0.35	0.54	−0.35	−0.33	−0.34	0.24	0.60

Note. Data are from van der Maas and Wagenmakers's (2005) chess study. 1PL model = one-parameter logistic model; 2PL model = two-parameter logistic model; θ = person ability.

The D-Diffusion IRM

We first return to the interpretation of the 2PL in terms of the diffusion parameters by Tuerlinckx and De Boeck (2005). We have argued that this interpretation makes sense for attitude and personality tests, because attitudes and personality traits often suggest a bipolar structure; examples are pro versus contra attitudes and introverted versus extraverted personalities. In such cases, item and person drift rates can be negative; therefore, the difference function $v = v^p - v^i$ seems reasonable. Because a has to be positive, the ratio function seems appropriate for $a = g(a^p, a^i)$. Hence, for bipolar traits we propose the following:

$$P_+ = \frac{e^{\frac{a_k^p}{a_j^i}(v_k^p - v_j^i) - \ln(M_j - 1)}}{1 + e^{\frac{a_k^p}{a_j^i}(v_k^p - v_j^i) - \ln(M_j - 1)}}. \quad (15)$$

This is a minor extension of Tuerlinckx and De Boeck's model, but it clarifies the role of the person (response caution) and item (time pressure) part of the discrimination parameter. We call this model the *D-diffusion item response model*, where D denotes difference.

An interesting consequence of this separation of the person and item role in boundary separation for both the Q- and D-diffusion item response models, is that it allows one to model person fit. Strandmark and Linn (1987) presented a 2PL model for person fit (see also Ferrando, 2007; Reise, 2000). Of interest, their 2PL model is similar to the D-diffusion item response model. In their model, the discrimination parameter of the 2PL is equal to the product of a person and an item discrimination parameter. However, both the interpretation and the estimation of Strandmark and Linn's model are problematic. We think that the concepts of response caution and time pressure may help to interpret the person and item discrimination parameter in this model. Such an interpretation directly suggests other means of testing the interpretation, such as experimental manipulations of response caution or time pressure and/or using response time data.

When we define person fit in terms of response caution, we can also see that person fit plays a different role in attitude testing than in ability testing. In the D-diffusion model, a lower level of response caution may increase P(agree) for $\theta < \beta$. In the Q-diffusion model, however, a lower level of response caution is always counterproductive, because it decreases P_+ for all (positive) values of θ.

IRT Models for Response Times

A number of IRT-type models for the joint analysis of accuracy and response time were reviewed by van der Linden (2009). One of the key models in this tradition is Rasch's (1960) model for misreadings and reading speed. In the model part for misreadings, the expected number of misreadings depends on a probability, $\theta_{jk} = \delta_j \xi_k$, that is, the ratio of text difficulty and person reading ability. In the model part for RT, the speed parameter (the expected number of words read per time unit) equals $\lambda_{jk} = \delta_j \xi_k$. This suggests that the same parameters explain accuracy and response times. However, Rasch did not necessarily believe this to be the case and proposed that this be decided empirically. This is exactly the line followed by van der Linden and his collaborators; van der Linden (2007) proposed a hierarchical approach, in which RT and responses are modeled separately by item and person parameters that can be related in different ways at a second level of modeling, depending on the data. Because his model is currently the most promising approach within IRT, we compare our model to the hierarchical approach.[12]

A key idea in van der Linden's (2007) and most other IRT response time models is that, in addition to the usual ability parameter, a new latent construct is required, usually in the form of a speed parameter. He distinguished between item time intensity and person speed parameters on the one hand and item difficulty and person ability parameters on the other. Item time intensity and person speed are used to model response times, whereas item difficulty and person ability figure in the accuracy model. Item difficulty and time intensity are latent parameters that derive their meaning entirely from the fact that they represent the effects of the items on the probability of a correct response and the time spent on items, respectively. Because these are different quantities, the two types of effects are different, although they may correlate across items (van der Linden, 2009). Hence, the item and person parameters are combined only at the second order level of van der Linden's hierarchical model.

An important underlying idea in van der Linden's (2009) model is the fundamental equation of response time modeling. According to van der Linden, the person speed parameter equals the amount of labor required to solve the item divided by response time. Hence, response time is the ratio of amount of labor and speed. A logarithmic transformation then gives $E[\ln(RT_{jk})] = \xi_j - \tau_k$. This equation is the basis for the response time part of van der Linden's model.

How does this model relate to the Q-diffusion model? First of all, we note that parameters introduced in van der Linden's (2007) model are typical latent variables affecting either accuracy or response time. In the Q-diffusion model, we have no such variables. Drift rate and boundary separation fundamentally differ from the speed and ability parameters of the model types discussed earlier. This is because, in the Q-diffusion model, these are *process parameters*, which are of equal importance to accuracy and to response time. Because they are not defined by their effects on the probability of a correct response and time spent on items, as the standard IRT parameters are, there is no objection to using them within the same level of modeling.

Of course, an important advantage of the Q-diffusion model is that its extension to response time modeling need not be crafted, because it is the very basis of the model. Given appropriate data,

[12] van der Linden (2009) criticized models that integrate response time and accuracy at one level. An example is Roskam's (1997) model in which the log of response time is added to the $\theta - \beta$ part of the 1PL model, which increases P_+ when more time is spent on the item. In other models this correction is replaced by a speed parameter (Verhelst, Verstralen, & Jansen, 1997) or modified by person and item parameters (Wang & Hanson, 2005). The general form of these types of models is $\text{logit}(P_+) = \theta_k - \tau_k - \beta_j$. An example of a model that integrates accuracy parameters in a model for response time is Thissen's (1983) model: $\ln(RT_{jk}) = \mu + \tau_k + \xi_j - \rho(\alpha_j \theta_k - \beta_j) + \varepsilon_{jk}$, where μ is a general intercept, τ_k and ξ_j are slowness parameters of person i and item j, respectively, ρ determines the influence of the usual 2PL response parameter structure and, ε is a normally distributed error term.

we could fit the joint distribution of accuracy and response time as specified in Equation 2. We can also use the equation for the mean response time (Equation 5). It is informative to relate the mean response time prediction of the diffusion model to van der Linden's (2009) fundamental equation of response time modeling. As explained in footnote 1, for reasonably high values of av, expected response time equals $a/2v$. Hence,

$$E(DT) = \frac{\xi_j^*}{\tau_k^*} \approx \frac{a}{2v} = \frac{1}{2}\frac{a_k^p/a_j^i}{v_k^p/v_j^i} = \frac{1}{2}\frac{v_j^i/a_j^i}{v_k^p/a_k^p}$$

$$E(\ln(DT)) = \xi_j - \tau_k \approx -\ln(2) + \ln\frac{v_j^i}{a_j^i} - \ln\frac{v_k^p}{a_k^p}$$

$$\xi_j \approx \ln\frac{v_j^i}{a_j^i};\ \tau_k \approx \ln\frac{v_k^p}{a_k^p}. \qquad (16)$$

In words, the time intensity and speed parameters in van der Linden's model relate linearly to the logarithm of the ratio of drift rates and item boundary separations for items and persons, respectively. Because van der Linden applied standard 2PL or 3PL to the response data, the translation of his difficulty and ability parameters is already specified in Equation 16, as the products of item and person drift rates and boundary separations, respectively (see van der Linden, 2009).

These relations constrain the model at the second level of van der Linden's (2007) hierarchical model, which in turn allows for a test of the Q-diffusion model within the modeling approach of van der Linden (Fox, Klein Entink, & van der Linden, 2007).

For instance, if speed and ability parameters at the second level in van der Linden's model are positively correlated, then individual differences are primarily due to differences in drift rate (e.g., see van der Linden, 2007). If these parameters correlate negatively, the individual differences in drift rate are probably similar across subjects, and differences are mainly due to differences in response caution (see Example 2 of Klein Entink, Fox, & van der Linden, 2009).

There is one important problem for the Q-diffusion model that emerges from this comparison. Dimensional analysis, as applied to Equation 3, leads to sound results. Boundary separation is measured in units of information, and drift rate is a speed measure (units per seconds), so that their ratio RT (DT) is measured in seconds. However, because we modeled v and a as ratios of the person and item parameters, they become dimensionless quantities. As a consequence, RT becomes dimensionless too.

To solve this problem, we have to reconsider our choices of $v = f(v^p, v^i)$ and $a = g(a^p, a^i)$. For v, we suggested a solution in footnote 6. If v^p is the information-processing power of the person, and v^i is the force required to solve the item, their ratio v is a speed measure. For g, as a provisional solution, we suggest that a^p be defined in units of information, and we suggest that time pressure, a^i, be viewed as a dimensionless quantity modifying person response caution. Clearly, more work is required here, involving precise analysis of how people set their response boundaries and how they are influenced by task factors, such as instruction, rewards, and time limits. This is currently an active area of research in experimental psychology (e.g., Simen et al., 2009; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

IRT Models for Guessing

The Q-diffusion model handles guessing in a principled way: The guessing probability for equally attractive response options equals 1/M for zero ability. We note that it is quite remarkable that guessing can, in fact, be handled by restricting the 2PL model (to positive θ) instead of extending the model with a guessing parameter, as is common in the IRT tradition. How, precisely, does the Q-diffusion model relate to other models for guessing?

Currently, the 3PL is by far the most popular IRT model to accommodate guessing. Yet, this model suffers from at least two problems. First, the estimates of 3PL parameters are unstable, especially for small samples (see San Martin et al., 2006, for an overview). If the guessing parameters are constrained to be equal to each other, or to $1/M$, this problem is less severe. If, additionally, all discrimination parameters are set to unity, we obtain the difference model with guessing (DM-G), or the Rasch model with guessing discussed earlier. However, this model seems to have worse fit in the comparison of models by Ramsay (1989).

The second problem of the 3PL, which also plagues the DM-G, is its interpretation. As noted earlier, different interpretations based on a distinction in a p-process of searching for the correct answer, and a g-process for guessing are possible (Hutchinson, 1991). Subjects may search for the correct answer, and guess when this search fails (assuming that they recognize their failure); however, other serial or parallel setups are possible. According to San Martin et al. (2006), the success of guessing also depends on ability. They therefore introduce ability-based guessing within the DM-G model, by making the success of the g-process dependent on ability, weighted by a discrimination parameter α:

$$P_+ = \frac{e^{(\theta_k-\beta_j)}}{1+e^{(\theta_k-\beta_j)}} + \left(1 - \frac{e^{(\theta_k-\beta_j)}}{1+e^{(\theta_k-\beta_j)}}\right)\frac{e^{\alpha_j\theta_k+\gamma_j}}{1+e^{\alpha_j\theta_k+\gamma_j}}. \qquad (17)$$

For $\alpha = 0$, this model reduces to the DM-G with guessing parameter equal to the expit of γ. In an empirical example, they show that this data-mining model with ability-based guessing (DM-AG) better explains the data than the DM or the DM-G. However, a disadvantage of the DM-AG (or 1PL-AG, as San Martin et al., 2005, called it) is that the p-process and g-process are not well separated anymore. If ability plays a too large role in guessing, the success of the g-process could be equal to or even greater than the success of the p-process. Also, for very low ability, the model predicts below chance responding (Property 3; see San Martin et al., 2005).

As Ramsay (1989) remarked, it would be more elegant to have a model based on a single process model. Especially when the p- and g-process get mixed up, a single process model might be preferable. We think that Ramsay's QM and our Q-diffusion model are more attractive for this reason. In the QM, guessing depends on ability, and there is a natural transition form accurate responding to guessing: The lower the level of ability, the lower the probability of a correct response; this probability has its lower asymptote at $1/M$ for an ability level of zero (which represents no ability). In the QM, there is just one probabilistic process: ability-based guessing. Pure guessing, educated guessing ($P_+ > 1/M$), and correct responding are explained by the same underlying mechanism.

Of course, sometimes a two-process description of item responding might be more accurate. Cao and Stokes (2008) and

Bechger, Maris, and Verstralen (2005) discussed several guessing scenarios and the associated IRT models. It would be very interesting to attempt to derive these models from sequential sampling models of decision making that include a guessing process (see, e.g., Ratcliff, 2006). The DM-AG may serve as a basic model here.

Another model for guessing was introduced by Hessen (2004, 2005). Hessen investigated a subclass of the four parameter logistic item response model, for which both specific objectivity and sufficiency of the total score for the latent trait hold. One special case that Hessen (2004) proposed is as follows:

$$P_{+} = \frac{\lambda/e^{\beta_j} + e^{\theta_k - \beta_j}}{1 + e^{\theta_k - \beta_j}}. \quad (18)$$

This IRT model has upper asymptotes at 1 but lower asymptotes at λ/e^{β_j}. Hence, the easier the item is, the higher the guessing asymptote becomes. This, in a sense, mirrors the DM-AG in that the success of guessing depends on the difficulty of the item and not on the ability of the subject. We therefore call this model the DM-DG, for difficulty-based guessing.

The prime attractiveness of Hessen's (2004) model lies in the statistical properties of specific objectivity and sufficiency. These advantages also apply to the DM-G and, of course, the DM but do not apply to the DM-AG or the QM. Ramsay (1989) discussed a way in which the QM permits specific objective comparisons, but clearly sufficiency of the total score for ability is missing. This is a disadvantage, but we agree with Ramsay's remark that we should not overemphasize statistical convenience.

Discussion

In introductions to IRT, the preference for the logistic equation is typically explained in terms of statistical or measurement-theoretical convenience. However, from a substantive point of view, the lack of a psychological justification for this key property of the measurement model compromises test validity. The reason is that validity requires a causal mechanism linking the trait or ability with the item responses (Borsboom & Mellenbergh, 2007; Borsboom et al., 2004). In the absence of such a mechanism, the relation between the targeted attribute and the item responses is essentially a black box, and the psychological appropriateness of the function that describes this relation becomes an article of faith. Because the processes that lie between item administration and item response are psychological in nature, the only way to remedy this situation is to construct psychological theories of item response processes and to link these to models for individual differences. In the positive ability model, we make this important step. Apart from the fact that the ensuing investigation, in our view, has produced many unexpected implications and surprising results, the most important aspect of this article may be that it gives proof of concept: It is possible to systematically connect latent variables to item responses through process models to get a *substantive* handle on the measurement problem in psychology.

Clearly, however, we have only begun to investigate the common ground covered by IRT and cognitive process models. We consider the further exploration of this territory to be of significant importance for both IRT modeling and formal models of cognitive psychology. Some possibilities for advances along these lines are the following.

First, in the current article, we proposed a fundamental difference between ability tests on the one hand and personality and attitude tests on the other, by noting that a diffusion process renders a traditional IRT model unlikely for abilities but plausible for personality and attitudes. This is surprising because IRT models have been traditionally proposed for, and applied in the context of, ability testing: Although applications to personality and attitudes have become more frequent in the past decades, these are clearly spin-offs of the ability testing approach. It turns out, however, that the situation might as well have been reversed: From a process-modeling point of view, standard IRT models are plausible for attitudes and personality but not necessarily for ability tests. This, of course, invites further investigations into the substantive nature of abilities versus personality traits and attitudes and into the methodological and psychometric treatment of test scores that is consistent with that nature.

The item response model that we argue is required for abilities radically differs from standard item response models in a number of respects. In particular, the postulate that ability is essentially positive has far-reaching implications. It leads to scales with natural zero points, inviting further analysis concerning the measurement properties of such scales. It also leads to an item response model that incorporates guessing as part of the decision process. In contrast to other item response models of guessing, the positive ability model can accommodate for guessing by restricting instead of extending the standard 2PL model. This is a remarkable result. Finally, the modeling framework leads to novel interpretations of standard item response parameters. For instance, it is surprising that the positive ability model has no standard β parameter. Instead, the difficulty of the item is incorporated in the discrimination parameter. At the same time, we have shown that standard ability and discrimination parameters, as well as van der Linden's (2007) time intensity and speed parameters, can be translated to the basic diffusion parameters of drift rate and boundary separation. The approach of van der Linden stands out as the best current psychometric model for accuracy and response times; therefore, the fact that we were able to find such strong relations between his approach and ours is promising.

Considering abilities and their measurement, we think that the positive ability model strongly suggests that the appearance of unidimensionality for broad sets of items, as is observed, for instance, in intelligence testing or educational measurement, is just that: appearance. Arithmetic items for addition and multiplication simply cannot be unidimensional, because they depend on discretely separable abilities, each of which is plausibly governed by a separate positive ability model. The appearance of unidimensionality probably results from the fact that these abilities are strongly intertwined and arranged in a hierarchical fashion; this, in turn, results in data that appear to be unidimensional because of the implied strong positive association between item responses. However, this should not be mistaken for evidence that a single ability is in play. It merely means that individual differences in performance can be reasonably described by a scalar variable. Statistical unidimensionality, thus, does not imply that a single ability is measured. It is important to stress this fact, because in both the psychometric and substantive literatures, the concepts of a psychometric latent variable and a substantive ability have become conflated, as have the activities of fitting a unidimensional model and measuring a single ability; these concepts and activities do not

coincide and should be separated clearly (van der Maas et al., 2006). The diffusion account may be useful in disentangling different abilities, because it extends the standard IRT paradigm with predictions on the behavior of response time data. Building up an account of the relations between distinct abilities in typical tests should be a main point on the psychometric and psychological research agendas for the next decades.

Naturally, the presented modeling approach hinges on the appropriateness of the chosen process model. Thus, one possible critique of our model could be based on the fact that the diffusion model is normally applied to simple fast decisions, as in perceptual tasks or lexical decision research, rather than to the type of decisions found in tests used in differential psychology. For example, in many ability tests that are analyzed with IRT, the decision process may consist of longer, perhaps sequential, stages that could probably not be reduced to one simple random walk of information accumulation. In this case, a simple random walk is at best a rough approximation of the underlying decision process. On the other hand, given reasonable assumptions, more complex decision models reduce to the diffusion model (Bogacz et al., 2006). Also, some slow responses to certain knowledge-based questions (What is the biggest country of Europe?) may be well described by a simple random walk process. For other decisions, for instance, those involving a series of computations, a simple random walk may be too simplistic but may still serve as a reasonable first approximation. In addition, it has been shown that optimal decision making, even in more complex decision models, may be best described by the diffusion model (Bogacz et al, 2006; Ratcliff et al., 1999).

Another strong assumption in our approach concerns the fact that ability is essentially positive. First, it could be argued that some abilities do, in fact, have a bipolar structure, which admits for both positive and negative values in the model. A classic example is the Piagetian ability to conserve quantitative properties as number mass and volume, in spite of changes in form. Children who do not understand conservation systematically score below chance level on multiple-choice items of a conservation test. In such a case, all qualitative implications of Tuerlinckx and De Boeck's (2005) model make perfect sense, as explained in footnote 4, and we would recommend use of the D-diffusion item response model in this case (Equation 15). However, we do not believe this to be a counterexample to our thesis of ability being essentially positive, because children's responses to conservation items are determined by two mutual exclusive strategies or abilities causing sudden transitions in developmental trajectories (van der Maas & Molenaar, 1992), and each of these should be constructed as being an essentially positive ability.

Second, the Q-diffusion model requires tests that measure a single ability. Because some psychological and educational tests clearly measure a host of related abilities, the applicability of the Q-diffusion model to such cases may be limited. Perhaps a conjunctive multidimensional or multicomponent Q-diffusion model could be developed for such situations. On the one hand, this represents an opportunity for further research. On the other hand, one could also argue that we should reconsider the use of tests that depend on multiple related abilities. From a measurement point of view, single ability tests should be preferred. The integration of simple abilities into higher order abilities, perhaps culminating in what seems to be a single overarching dimension, should ideally be explicitly modeled; this would arguably be a more transparent approach than the current practice, in which multiple related abilities are implicitly grouped together in a single dimension. The disadvantage of the latter procedure is that the emergent single dimension no longer represents a theoretically transparent psychological concept.

Cognitive diagnosis models (de la Torre, 2009; Embretson, 1984; Hoskens & De Boeck, 2001; Maris, 1995) share the basic idea that psychometric models should ideally be formalized substantive theories of item responses. In addition, in cognitive diagnostic models, the item response probabilities are modeled as a function of a constellation of skills that have the character of essentially positive abilities. For instance, these models may analyze the response to an item like $(2 + 2)/6 \times 3 = ?$ by decomposing it into skills for addition, multiplication, and fractions. The current model, however, provides a process level account of the simple abilities themselves; that is, it applies to simple abilities in isolation. It should be possible to integrate these two approaches through the construction of hierarchical models involving the interplay between distinct simple abilities, and this represents an interesting avenue for further research.

Third, a rather radical consequence of the definition of ability in the Q-diffusion model is that any able person will eventually solve all items of test, even if ability is very small and the item extremely difficult. This consequence is theoretically valuable because it represents an important testable prediction that flows naturally from the model. It is also useful to characterize the nature of simple abilities. However, in psychometric practice, it may not be appropriate in certain situations. In these cases, extensions of the Ratcliff diffusion model (e.g., introducing variance in drift rates) can be used to eliminate this property of the model so that it allows for response errors even when boundary separation (available time to respond) approaches infinity.

A final limitation of the Q-diffusion model concerns the solution we propose to deal with multiple-choice items. We have derived the correction for multiple-choice from Bock's (1972) nominal response model, which leads to a simple extension of the Q-diffusion model that can handle multiple-choice items. However, the resulting transformation is not so easily applicable to the formula for response times. As a consequence, the Bayesian fit procedure for $M > 2$ is complicated and requires additional assumptions. Thus, it would be preferable to derive a Q-diffusion type model from a multiple-choice stochastic sequential sampling model for decision making. In view of the recent interest in multiple-choice decision models in mathematical psychology (e.g., McMillen & Holmes, 2006), we are hopeful that such an approach is within reach.

Clearly, the connection between process models for decision making and IRT models for individual differences is an extremely fruitful one. It allows researchers in individual differences research to craft process models for their item responses as well as response times and to develop new research strategies and hypotheses that may function to elucidate how tests work. The proposed systematic connection between psychological processes and psychometric latent variables may allow researchers to address their validity problems by uncovering how their tests work, that is, by explicitly modeling the processes that lie between item administration and item response (Borsboom et al., 2004). For this reason, the present investigations may do much more than merely extend the family of

IRT models with some new members; they may serve to finally get a grip on the validity issues that have plagued psychological testing for the past century.

References

Bartholomew, D. J. (1987). *Latent variable models and factor analysis.* London, England: Griffin.

Bechger, T. M., Maris, G., & Verstralen, H. H. M. (2005). The Nedelsky model for multiple choice items. In A. van der Ark, M. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 187–206). Mahwah, NJ: Erlbaum.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores. Reading* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51. doi:10.1007/BF02291411

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. C. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113,* 700–765. doi:10.1037/0033-295X.113.4.700

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440.

Borsboom, D., Kievit, R., Cervone, D. P., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudary (Eds.), *Developmental process methodology in the social and developmental sciences* (pp. 67–98). New York, NY: Springer.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity and cognitive assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85–116). Cambridge, England: Cambridge University Press. doi:10.1017/CBO9780511611186.004

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110,* 203–219. doi:10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071. doi:10.1037/0033-295X.111.4.1061

Borst, G., Kievit, R. A., Thompson, W. L., & Kosslyn, S. M. (2011). Mental rotation is not easily cognitively penetrable. *Journal of Cognitive Psychology, 23,* 60–75.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459. doi:10.1037/0033-295X.100.3.432

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, 73,* 209–230. doi:10.1007/s11336-007-9045-9

Cox, D. R., & Miller, H. D. (1970). *The theory of stochastic processes.* London, England: Chapman & Hall/CRC Press.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48,* 129–141. doi:10.1007/BF02314681

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

de Groot, A. D. (1978). *Thought and choice in chess.* The Hague, The Netherlands: Mouton. (Original work published 1946)

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement, 33,* 163–183. doi:10.1177/0146621608320523

Edwards, W. (1965). Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *Journal of Mathematical Psychology, 2,* 312–329. doi:10.1016/0022-2496(65)90007-6

Elo, A. (1978). *The rating of chessplayers, past and present.* New York, NY: Arco.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186. doi:10.1007/BF02294171

Ferrando, P. J. (2007). A Pearson-Type-VII item response model for assessing person fluctuation. *Psychometrika, 72,* 25–41. doi:10.1007/s11336-004-1170-0

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item-response model incorporating response time data in binary personality items. *Applied Psychological Measurement, 31,* 525–543. doi:10.1177/0146621606295197

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 157–180). New York, NY: Springer.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York, NY: Springer-Verlag.

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., & Rubin, D. B. (2004). *Bayesian data analysis.* London, England: Chapman & Hall.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215–231. doi:10.1093/biomet/61.2.215

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated state–space model. *Journal of Research in Personality, 41,* 295–315. doi:10.1016/j.jrp.2006.04.003

Harré, R., & Madden, E. H. (1975). *Causal powers.* Oxford, England: Blackwell.

Hessen, D. J. (2004). A new class of parametric IRT models for dichotomous item scores. *Journal of Applied Measurement, 5,* 385–397.

Hessen, D. J. (2005). Constant latent odds-ratios models and the Mantel-Haenszel null hypothesis. *Psychometrika, 70,* 497–516. doi:10.1007/s11336-002-1040-6

Hoskens, M., & De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25,* 19–37. doi:10.1177/01466216010251002

Hutchinson, T. P. (1991). *Ability, partial information and guessing: Statistical modelling applied to multiple-choice tests.* Rundle Mall, South Australia: Rumsby Scientific.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133. doi:10.1007/BF02291393

Kievit, R. A. (2010). *Representational inertia: The influence of associative knowledge on 3D mental transformations.* Unpublished manuscript, Department of Psychology, University of Amsterdam, The Netherlands.

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. doi:10.1007/s11336-008-9075-y

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* London, England: Academic Press.

Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method.* London, England: Butterworth.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston, MA: Houghton Mifflin.

Link, S. W. (1992). *The wave theory of difference and similarity.* Hillsdale, NJ: Erlbaum.

Lord, F. M. (1952). *A theory of test scores.* New York, NY: Psychometric Society.

Lunn, D. J. (2003). WinBUGS development interface (WBDev). *ISBA Bulletin, 10,* 10–11.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547. doi:10.1007/BF02294327

McLachlan, G., & Peel, D. (2000). *Finite mixture models.* New York, NY: Wiley.

McMillen, T., & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology, 50,* 30–57. doi:10.1016/j.jmp.2005.10.003

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307. doi:10.1037/0033-2909.115.2.300

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives, 6*(1–2), 124.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research.* The Hague, The Netherlands: Mouton.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2,* 201–218.

Ramsay, J. O. (1989). A comparison of three simple test theory models. *Psychometrika, 54,* 487–499. doi:10.1007/BF02294631

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85,* 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278–291.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology, 53,* 195–237. doi:10.1016/j.cogpsych.2005.10.002

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. doi:10.1111/1467-9280.00067

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging, 16,* 323–341. doi:10.1037/0882-7974.16.2.323

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Comparing connectionists and diffusion models of reaction time. *Psychological Review, 106,* 261–300. doi:10.1037/0033-295X.106.2.261

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543–568. doi:10.1207/S15327906MBR3504_06

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, NY: Springer.

Roskam, E. E., & Jansen, P. G. W. (1984). A new derivation of the Rasch model. In E. DeGreef & J. van Buggenhaut (Eds.), *Trends in mathematical psychology* (pp. 293–307). Amsterdam, The Netherlands: Elsevier. doi:10.1016/S0166-4115(08)62094-4

San Martin, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability based guessing. *Applied Psychological Measurement, 30,* 183–203.

SAS Institute. (2000). *SAS/STAT user's guide* (Version 8). Cary, NC: SAS Institute.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward-rate optimization in two-alternative decision making: Empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 1865–1897. doi:10.1037/a0016926

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25,* 251–260. doi:10.1007/BF02289729

Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement, 11,* 355–370. doi:10.1177/014662168701100402

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, NY: Academic Press.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70,* 629–650. doi:10.1007/s11336-000-0810-3

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14,* 1011–1026.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308. doi:10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement, 46,* 247–272. doi:10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York, NY: Springer.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113,* 842–861. doi:10.1037/0033-295X.113.4.842

van der Maas, H. L. J., Kolstein, R., & van der Pligt, J. (2003). Sudden jumps in attitudes. *Sociological Methods & Research, 32,* 125–152. doi:10.1177/0049124103253773

van der Maas, H. L. J., & Molenaar, P. C. M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99,* 395–417. doi:10.1037/0033-295X.99.3.395

van der Maas, H. L. J., & Wagenmakers, E. J. (2005). The Amsterdam Chess Test: A psychometric analysis of chess expertise. *American Journal of Psychology, 118,* 29–60.

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

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767–775.

Wagenmakers, E. J., Grasman, R., & Molenaar, P. C. M. (2005). On the relation between the mean and the variance of a diffusion model response time distribution. *Journal of Mathematical Psychology, 49,* 195–204. doi:10.1016/j.jmp.2005.02.003

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58,* 140–159. doi:10.1016/j.jml.2007.04.006

Wagenmakers, E. J., van der Maas, H. L. J., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14,* 3–22.

Wald, A. (1947). *Sequential analysis.* New York, NY: Wiley.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29,* 323–339. doi:10.1177/0146621605275984

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85. doi:10.1016/0001-6918(77)90012-9

Received January 21, 2010
Revision received December 16, 2010
Accepted December 17, 2010 ■