Abstract

This manuscript 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 two-choice response processes entails a two-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/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 item response theory. 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, decision making, 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 are, for instance, the models of Rasch (1960), Birnbaum (1968), Lord (1952), and Mokken (1971) for dichotomous items and a continuous latent variable; factor models for continuous items and a continuous latent variable (Lawley & Maxwell, 1963; Jöreskog, 1971; 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 (McLachlan & Peel, 2000; Bartholomew, 1987).

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 (\( \theta \)) and item difficulty (\( \beta \)), weighted by item discrimination (\( \alpha \)) in the following way:

\[
P_+ = \frac{e^{\alpha(\theta - \beta)}}{1 + e^{\alpha(\theta - \beta)}}
\]

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 \( \alpha_j \) 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 model, 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- and item parameters, but do not speak on the question how these item responses are generated; as Mislevy, (2008) states, ‘[n]either 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.’

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 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 how a testing procedure works, i.e., which processes transmit variation among individuals into variation in the item responses (Borsboom, Mellenbergh, & Van Heerden, 2004; Borsboom & Mellenbergh, 2007). 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, i.e., 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 intra-individual process models and models for inter-individual differences (Molenaar, 2004; Hamaker, Nesselroade & Molenaar, 2007; Borsboom, Mellenbergh, & Van Heerden, 2003; 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 what follows, we take their result as a starting point, and augment it to accommodate different testing situations as they may arise in practice. As will become apparent, such extensions carry important consequences for the interpretation of existing models.

The structure of this paper is as follows. First, we shortly 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 Item Response Theory models and diffusion processes**

While 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 the 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 and 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, the SPRT (Stone, 1960; Wald, 1947). This implies that the diffusion model optimizes expected accuracy given RT, or conversely, optimizes RT given the level of accuracy. Furthermore, as Bogacz et al. (2006) show, more biologically realistic complex models with leakage, inhibition and pooled inhibition can be reduced to the drift diffusion model, and in this way implement the SPRT. 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: 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 \( (T_w) \) 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.
Figure 1 displays the basic ingredients of the diffusion model. Upon being 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 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 \( (T_{ev}) \), which may for instance cover perception of the stimulus and the time needed to execute a motor response, and decision time \( (DT) \).

The information accumulation process that leads to \( DT \) depends on a drift rate parameter, which varies over trials with mean \( v \) and variance \( \eta \). 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 reduce. In this way, the inverse relation between speed and accuracy, i.e., the speed accuracy trade-off, is naturally accommodated in the model.

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 via payoff or proportion manipulations (Edwards, 1965). Such manipulations are generally not 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 paper we only use 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:

\[
\begin{align*}
\pi(x, t) &= \frac{\pi \sigma^2}{a^2} \exp\left(\frac{ax - z}{\sigma^2}\right) - \frac{v^2}{2\sigma^2}(t - T_{er}) \\
\times \sum_{a=1}^{\infty} m \sin\left(\frac{\pi m (ax - 2z + z)}{a}\right) \exp\left(\frac{-1}{2} \frac{\pi^2 a^2 m^2}{2} (t - T_{er})\right)
\end{align*}
\]

(2)

In contrast, the equations for the probability of a correct response and for the expectation of \( RT \) (i.e., of \( DT + T_{ev} \)) are relatively easy. The probability of \( X=1 \), which we denote by \( P_+ \), (i.e., terminating at the upper boundary) is (Cox & Miller, 1970):

\[
P_+ = P(X = 1) = \frac{e^{-2av}}{e^{-2av} - 1} \tag{3}
\]

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

\[
P_+ = \frac{e^{-av}}{e^{-2av} - 1} = \frac{e^{av}}{1 + e^{av}} \tag{4}
\]

The mean \( DT \) can be expressed\(^1\) as

\[E(DT) = \frac{a}{2v}\] 

(rewritten from Cox & Miller, 1970) where \( P_+ \) is the probability of a correct response, and \( E(DT) \) is the expected decision time. The underestimation of \( E(DT) \) in this approximation is not

\(^1\) For unbiased \( (z = 1/2 a) \) decision making we can rewrite this to \( E(DT) = (2 / 2v) (2 P_+ - 1) \). 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 \( |av|/2v \).
E(DT) = \frac{a \left(1 - e^{-av}\right)}{2v \left(1 + e^{-av}\right)} \tag{5}

Tuerlinckx and De Boeck (2005) have connected Equation (4) to the 2PL (1). They argue that \(v\), the mean drift rate, can be decomposed into a person part (\(\theta\)) and an item part (\(\beta\)), and propose a simple linear relation such that \(v = \theta - \beta\). At first sight, this seems to be a reasonable step. However, as will be shown below, 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 \(\alpha\) in the 2PL model within the derivation of Tuerlinckx 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 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 will depend 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 will play a key role in the present article.

Tuerlinckx and De Boeck (2005) describe their work as fostering a new interpretation of item response models, rather than as a derivation. 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 will demonstrate later, alternative setups and interpretations are possible. On the other hand, in the work by Tuerlinckx and De Boeck (2005), 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 (2005) 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 not only yields predictions 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 RT’s (Ratcliff, 2006). Three qualitative implications are especially relevant in the current context.

A first implication that follows from the model, and that will play an important role below, is that reducing item administration time, in the limit, should yield \(P_r = 1/2\), irrespective of the value of \(\theta\). 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 ICC becomes a flat line at $P_\tau=1/2$, equal for all levels of $\theta$.

Clearly, this can only happen 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 above does not apply naturally to the ability tests for which IRT models are often used. We will extend the model in this direction later in this paper.

The second important implication of a diffusion interpretation of IRT models concerns the effect of changes in $\nu$ (which equals $\theta-\beta$ in IRT terms). Under a diffusion model, RT's are slowest when $\nu\approx0$, and hence in the IRT context should be slowest when $\theta$ equals $\beta$. Persons with $\theta<<\beta$ are expected to be very fast; in fact, they should be as fast as persons with $\theta>>\beta$. 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' are arbitrarily coded as $X=1$ and $X=0$. Subjects with extreme positions ($\theta<<\beta$ or $\theta>>\beta$) will probably answer confidently and quickly. Subjects with $\theta=\beta$ will be in doubt and are likely to respond slower (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 $\nu$, that is, whenever $\theta>\beta$, increases in boundary separation lead to increases in $P_\tau$. 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 $\nu$, 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 $P_\tau$ equals 1/2, it separates two qualitatively different regimes. Below this point, more time to solve the item will decrease $P_\tau$. Above this point, it will increase $P_\tau$. For very long time limits, the item characteristic curve should approach that of a Guttman item, i.e., become a step function.

---

3 In fact, this phenomenon favors Tuerlinckx and De Boeck's model since their model explains this effect without any alterations of the model. The model of Ferrando and Lorenzo-Seva requires an adaptation of Thissen's (1983) model for RT's to explain this effect.
This may again be considered plausible in typical personality or attitude tests, but not in ability tests. In personality and attitude testing, if $\theta$ is just below $\beta$, and the subject has to respond quickly, the noise factor in the decision process will play a large role. But 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 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 a extra parameter into the model. This results in the 3PL model (Birnbaum, 1968), which is an extension of the 1 or 2PL: $P_{3pl}=c_j+(1-c_j)P_{2pl}$, (see Equation 1). In this model $c$ is the lower asymptote of the item characteristic function, such that for any value of $\theta$, $P_3 \geq c$. San Martin, del Pino and De Boeck (2005), following Hutchinson (1991), discuss 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 recognizes this failure); however, an alternative order of processing is possible. It is not straightforward to derive this model from a diffusion model. We will 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 above make sense for two-choice attitude and personality tests, but not for typical ability tests. We consider this to be surprising, because 1) the diffusion model would, at first glance, seem to give a reasonable approximation of the response processes in a typical ability testing situation, and 2) 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 model can only be applied 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

\(^4\) 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 1) nonconservers typically score below chance level, because they believe the incorrect answer to be correct, 2) transitional children ($\theta=\beta$) are slower than nonconservers and conservers (van der Maas & Molenaar, 1992), and 3) the probability of a correct response is likely to decrease for nonconservers when they get more time to think over their answer.
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 only represent differences between levels of ability. That is, standard models of psychometric theory are restricted to the representation of individual differences (Borsboom, Mellenbergh, & Van Heerden, 2003), and the sentence 'John has value \( \theta_X \) 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?

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, while 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 an ability to an individual is to ascribe that individual a causal power. Although such powers may be present in 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 the above characteristics may be considered axiomatic for simple abilities and the tasks that depend upon 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' than others. This means, in the diffusion model logic followed here, that these individuals will a) 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

---

5 A standard reply of psychometricians to the requirement of nonnegative ability is that we can apply Rasch’s transformation $\theta^* = e^\theta$ in order to scale the latent trait values to positive values. However, this does not fully solve the problem, as it still allows some $\theta$ to be less than some $\beta$, implying subject scoring below chance.
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 $\theta$ will be assumed to be $\mathbb{R}^+$. Second, because all tasks require positive ability, the same holds for task 'difficulty', so that difficulty also has $\mathbb{R}^+$ 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 will 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 $P_\alpha = 0.5$ 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 will now derive a model that has all of these properties.

**The positive ability model**

In this paragraph, we take characteristics of ability outlined above 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**. Here, 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 $\text{logit}(P_\alpha) = av$ to derive the 2PL model; they equated the diffusion parameter $a$ with the IRT parameter $\alpha$, and 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 item response theory 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 x person parameter (in IRT, the corresponding parameter $\alpha$ is an item x person interaction parameter). In order 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.

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) propose 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 will be 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 because of the requirement that ability is positive (note that for this reason, the difference function proposed by Tuerlinckx and 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.

There exists a wide class of functions that have these properties, and thus instantiate submodels of the positive ability model. One example of a function that respects the above constraints, and that we will use in the following, is the quotient function, $v = v^p / v^i$. 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$^7$:

$$P_* = \frac{\frac{a^p v^p}{e^{a^p v^p}}}{1 + e^{a^i v^i}}$$

$\text{(6)}$

$^6$ A metaphorical way to think about this relation is in terms of the famous Newtonian relation speed (drift rate) = power (ability)/force (difficulty). Interestingly, Rasch (1960) proposed to decompose the speed parameter 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.

Cognitive psychology meets psychometrics 12

This model adheres to the properties attributed to simple abilities in the previous paragraph: a) both ability and difficulty are positive, b) drift rate $v = v^p / v^i$ is always positive, so that c) 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 (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 (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. In van der Maas and Maris (2010) we derived such an extension from Bock’s (1972) nominal response model (an IRT model for multiple-choice items with unordered categories). Under the assumption that incorrect alternatives are equally attractive to all subjects, a simple extension of the model to multiple-choice is possible. Here, we suffice by giving the model formula for a correct response, which is transparent in itself:

$$P_* = \frac{\frac{a^p v^p}{e^{a^p v^p}}}{1 + e^{a^i v^i}}$$

$$\text{(7)}$$

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 $1/M$, which is precisely what is desired. $^8$

$^8$ Tuerlinckx and De Boeck’s (2005) derivation of the 2PL model was based on the relation $\logit(P_*) = v = \alpha$
The model in (7) still has the structure of a 2PL model. Given
\[ \theta_k = a_j^\alpha v_j^\alpha; \alpha_j = 1/a_j^\alpha v_j^\alpha; \beta_j = \ln(M_j - 1)a_j^\alpha v_j^\alpha, \]
leads to equation (1), the standard 2PL model, with the constraints that both \( \theta \) and \( \alpha \) are nonnegative. For the linear setup of the 2PL, \( \logit(P_s) = \alpha_j^* \theta_k + \beta_j^* \) (as used in the nominal response model), \( \beta_j^* \) further reduces to \(-\ln(M_j-1)\), and \( \alpha_j^* = \alpha_j = 1/a_j^\alpha v_j^\alpha \).

We propose that model formula (7) 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. For the problem of naming this model, we propose to link it to an earlier model proposed by Ramsay (1989), which was proposed on a different basis, and which he called the Quotient Model (QM):
\[ P_s = \frac{e^{\alpha_j^* \theta_k + \beta_j^*}}{1 + e^{\alpha_j^* \theta_k + \beta_j^*}}, \theta_k \geq 0, \beta_j > 0 \]
(8)

When \( \theta = a^\rho v^\rho, \beta = a^\nu v^\nu, \) and \( K=M-1, \) Ramsay’s QM and the model in (7) are equivalent. Therefore, the model we proposed in Equation (7) 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 \( \theta, \alpha, \) and \( \beta. \) 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 not only depends on the information powers of the subject but also on his or her response caution. These two factors can be separated using response times. 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 \( \alpha \) 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 \( \beta \) (or actually \( \beta^* \)) 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 \( \theta \) parameter; i.e., the scale of \( \theta \) 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-

\[^9\] As Ramsay notes, 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:
\[ P_s = \frac{c e^{\alpha_j^* \theta_k + \beta_j^*}}{(1-c) + c e^{\alpha_j^* \theta_k + \beta_j^*}}, \]
which, given \( c=1/M \) and logarithmic transformations of \( \theta^* \) and \( \beta^* \), is again equivalent to Ramsay’s QM and thus to (7).
diffusion model certain ratio properties: For instance, if $M=2$ a doubling of $\theta$ gives a doubling of the logit score. A similar line of reasoning applies to the $\alpha$ parameter. For example, for the ability of addition, we could think of $1+1$ as an item for which the parameter $\nu$ is nearly zero, which implies that it has very high $\alpha$. For extremely difficult items, on the other hand, $\alpha$ approaches zero.

The item characteristic curves (ICC) of the Q-diffusion model are simple. Since all $a$ and $\nu$ are positive, all $\theta$ and $\alpha$ are nonnegative. This implies that all ICC’s reside in the first quadrant and all ICC’s are monotonically non decreasing. The intercept is simply $1/M$. Slopes are determined by $\alpha_j = 1/\alpha_i \nu_i$. Figure 3 gives examples for $M=2$, $M=4$ and $M=500$ (as an approximation to open questions$^{10}$).

The ICC’s 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 2- and 3PL models, the Q-diffusion model excludes the possibility that subjects score below chance level. Second, if $M_i = M$, then the model implies that ICC’s do not cross. As in the Rasch (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 manuscript does not work within the tradition that views data-fitting flexibility as the prime virtue of a measurement model. Third, all ICC’s start increasing at $\theta=0$ (see figure 3, where $M$ is 2 and 4). In the 1-, 2- and 3PL, the ICC’s of difficult items first remain at zero (or chance) level and begin to increase only when $\theta$ reaches the $\beta$ of the item.

Another way to demonstrate the ICC’s of Q-diffusion model is to compare them with the ICC’s 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 ICC’s 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 question, that the test does not measure a single ability, or both. In many practical testing situations, we do suspect that scores in fact 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 separable abilities. Thus, such a test may be treated with, for instance, a conjunctive multidimensional or multi-component Q-diffusion model (de la Torre, 2009; Embretson, 1984; Hoskens & De Boeck, 2001; Maris, 1995). The construction

---

$^{10}$ 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, only consider 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, not covered in the diffusion part of the decision process.
Figure 3: ICC of a restricted MC2PL for ability testing. All $\theta$ ($\nu_i \alpha$) and $\alpha (1/\nu_i \alpha)$ are by definition positive. Intercepts are equal to $1/M$, i.e. 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.

Figure 4: The quotient model compared with the Rasch (difference) model with guessing. To ease comparison, the QM model 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 ICC’s. They leave the lower asymptote slower than they approach the upper asymptote. A larger difference between the models is indicated by the arrows. DM-G ICC’s 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.
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 seconds per item. For some reason we decide to apply this item bank in a test with a different time limit, say 1 minute 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 β’s don’t change at all. Only the α’s change, as they increase upon relaxing the time limit.

**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 RT’s are required to distinguish between the α and ν part of persons and items. If RT’s are not available, however, we may still fit a partial Q-diffusion model to the accuracy data only (see also Ramsay, 1989).

If RT’s 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 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 (α_k = α^p, ν_k = ν^p), and items are divided into a small number of types, say A and B (ν_j = ν_A or ν_j = ν_B). Moreover, because items are submitted in mixed blocks, boundary separation is equal over item types (a_j = a'). Since 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 = a'. 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 ν_j. We developed a Bayesian fit technique to meet the special requirements of psychometric data.

**Example 1: Mental rotation**

We collected data of 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, the reader may confer Borst, Kievit, Thompson & Kosslyn (2010), who used the same experimental setup.

Responses were dichotomous (correct/incorrect) and RT’s were recorded. We randomly selected 10 out of 280 items of three different rotation angles (50°, 100°, 150°) for the model fitting analysis. We discuss the analysis at two levels that may arise in practice. First, we will analyze the accuracy data only. This allows for comparison against standard IRT models, which do not have implications for the response times; in addition, the analysis can be executed using standard software. Second, we will analyze the full dataset, including response times. This analysis presents more difficult problems; we tackled these problems using a Bayesian analysis.
Estimation without response times. Without RT’s we cannot distinguish between the $a$ and $v$ parameters of persons and items, but analyzing only the accuracy does allow for a comparison to standard IRT models. Therefore we choose the linear setup of the 2PL formulation of Equation 7, in which $\theta_k = a_k v_k^p$, $\alpha_j = 1/a_j v_j^p$, and $\beta_j = \ln(M_j - 1)$. The resulting model may be fitted using 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 ML estimates with standard errors. $M_j$ can be fixed or estimated, and models can be compared using information criteria such as AIC and BIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>-2LL</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental rotation example</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-diffusion</td>
<td>832.2</td>
<td>852.2</td>
<td>880.2</td>
</tr>
<tr>
<td>1PL</td>
<td>835.1</td>
<td>857.1</td>
<td>887.9</td>
</tr>
<tr>
<td>1PL guessing</td>
<td>846.5</td>
<td>866.5</td>
<td>894.5</td>
</tr>
<tr>
<td>2PL</td>
<td>830.5</td>
<td>870.5</td>
<td>926.4</td>
</tr>
<tr>
<td>3PL full</td>
<td>819.2</td>
<td>859.2</td>
<td>915.1</td>
</tr>
<tr>
<td>Chess ability example</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q-diffusion</td>
<td>4263.0</td>
<td>4341.0</td>
<td>4479.2</td>
</tr>
<tr>
<td>1PL</td>
<td>4309.3</td>
<td>4351.3</td>
<td>4425.7</td>
</tr>
<tr>
<td>1PL guessing</td>
<td>4214.8</td>
<td>4294.8</td>
<td>4436.7</td>
</tr>
<tr>
<td>2PL</td>
<td>4178.4</td>
<td>4258.4</td>
<td>4400.3</td>
</tr>
<tr>
<td>3PL full</td>
<td>4164.8</td>
<td>4282.8</td>
<td>4491.9</td>
</tr>
</tbody>
</table>

Results are represented in Table 1, and indicate that the Q-diffusion model outperforms the 1PL, 2PL, as well as the 3PL and the 1PL with guessing. Both AIC and BIC 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 will generally be the same.

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 RT distribution with a lognormal distribution, i.e., $\log(RT_{kj}) \sim \text{normal}([\mu_{kj}, \sigma_{kj}^2])$. Parameters $\mu$ and $\sigma$ can be obtained if the values of mean and variance are known:

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

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

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

\[ E(RT_{ij}) = \frac{a_i^c v_i^j}{2a_i^c v_i^j} \left( 1 - e^{\beta_j} \right) + T_{er_k} \]

\[ \text{var}(RT_{ij}) = \frac{\alpha_p^i \left( \frac{v_i^j}{v_i^p} \right)^2 \left[ 2h_{kj} e^{\beta_j} - e^{2\beta_j} + 1 \right]}{(e^{\beta_j} + 1)^2} \] (10)

\[ h_{ij} = -\frac{v_i^j a_i^c}{v_i^j a_i^j} \]

In a simulation study we established good recovery of the parameters of the full model including \( T_{er} \) (i.e., non-decision time varying by person).

As uninformative item priors for \( a_i^j \) and \( \nu_i^j \), we use uniform distributions from .01 to .5 and from .01 to 100, respectively. As person priors for \( \alpha_p^i \), \( \nu_p^i \) and \( T_{er} \) we choose lognormal distributions with \( \mu \) and \( \sigma \) of 0 and 1, respectively. Note that compared to the other priors, the prior of \( a_i^j \) appears to be relatively narrow. However, from experiences with fitting the Q-diffusion model, we encountered that \( a_i^j \) was always roughly in the range of .25 - .35. We drew 10000 samples from the posterior distribution and discarded the first 5000 as burn-in. Judged from the plots of the MCMC 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

Figure 5: The observed data distribution (bars) and the posterior predictive distribution (solid line) for the mental rotation data.
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 dataset that does allow for such comparison is comprised of subset of data derived from a chess ability 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 dataset 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 dataset are challenging. First, the item format of these chess puzzles was open ended. Since the number of legitimate ‘sensible’ moves in chess is limited, we can interpret the item format as multiple-choice 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 RT’s (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 \( M \) and all other person and item parameters, where \( v_i \) represents the item drift rate for two alternatives:

\[
a_i^p; v_i^* = v_i^*; a_i^j = a_i^j
\]

\[
v_i^* = \frac{a_i^p v_i^* v_j}{a_i^k v_j - \ln(M^j - 1)a_i^j/v_j}
\]

These transformed parameters are then used to compute \( E(RT) \) and \( var(RT) \) according to Equation 9. 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 below.

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 end-game 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 (last 20 items of Choose-a-Move test B) and analyze the data with the full Q-diffusion model according to the Bayesian setup described above.

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 10000 samples from the posterior distribution and discarded the first 5000 as

\(^{11}\) 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, for the case that \( M > 2 \), we implemented the q-diffusion model in the WinBUGS Development Interface (WBDDev; Lunn, 2003) which is a freely available WinBUGS add-on program.
burn-in. Judging from the plots of the MCMC 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 RT’s, person drift rates, person response cautions and non-decision time. Using drift rate instead of test score improves validity slightly. However, it is prudent to note that, using 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 highest with $T_{er}$, a result also found in other applications of diffusion modeling (e.g., Ratcliff, Tapar & McKoon, 2001).

### Table 2: Correlations of the standard test statistics (test score and mean RT), person estimates according to the 1PL and 2PL model, and the Q-diffusion parameters person drift rate ($v$), response caution ($a$), and non-decision time ($T_{er}$), with the Elo ratings, and ages of chess players in van der Maas & Wagenmakers (2004) chess study.

<table>
<thead>
<tr>
<th></th>
<th>Test score</th>
<th>RT</th>
<th>1PL $\theta$</th>
<th>2PL $\theta$</th>
<th>$v$</th>
<th>$a$</th>
<th>$T_{er}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elo Rating</td>
<td>0.68</td>
<td>-0.44</td>
<td>0.67</td>
<td>0.69</td>
<td>0.72</td>
<td>-0.38</td>
<td>-0.17</td>
</tr>
<tr>
<td>Age</td>
<td>-0.35</td>
<td>0.54</td>
<td>-0.35</td>
<td>-0.33</td>
<td>-0.34</td>
<td>0.24</td>
<td>0.60</td>
</tr>
</tbody>
</table>

### Relation To Other IRT Models

Historically, there have been many attempts to incorporate RT’s in IRT models. This is an ongoing research topic, and in addition there are several relatively novel IRT models that handle RT 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.

### 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, since attitudes and personality traits often suggest a bipolar structure; examples are pro/contra attitudes and introvert/extravert personality. In such cases, item and person drift rates can be negative, and therefore the difference function $v = v_p - v_i$ seems reasonable. Since $a$ has to be positive, the ratio function seems appropriate for $a = g(a^p, a^i)$. Hence for bipolar traits we propose:

$$P_r = \frac{\frac{a^p}{a^i} \left( (v^p - v^i) - \ln(M) \right)}{1 + \frac{a^p}{a^i} \left( (v^p - v^i) - \ln(M) \right)}$$

(12)

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 IRM, is that it allows one to model person fit. Strandmark and Linn (1987) present a 2PL model for person fit (see also Reise, 2000, Ferrando, 2007). Interestingly, their 2PL model is similar to the D-diffusion IRM. 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 $P(\text{agree})$ for $\theta<\beta$. In the Q-diffusion model, however, a lower level of response caution is always counterproductive, since it decreases $P_\text{r}$ for all (positive) values of $\theta$.

IRT Models For Response Times

Van der Linden (2009) reviews a number of IRT-type models for the joint analysis of accuracy and RT. 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 $\theta_{jk} = \delta_j/\xi_k$, i.e., 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 same parameters explain accuracy and response times. However, Rasch (1960) did not necessarily believe this to be the case, and proposed to let this be decided empirically. This is exactly the line followed by van der Linden and his collaborators. Van der Linden (2007) proposes 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. Since his model is currently the most promising approach within IRT, we compare our model to the hierarchical approach.\(^\text{12}\)

A key idea in van der Linden’s and most other IRT-RT 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 distinguishes 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 RT’s, 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

\(^\text{12}\) Van der Linden criticizes models that integrate response time and accuracy at one level. An example is Roskam’s (1987) model in which the log of response time is added to the $\theta-\beta$ part of the 1PL model, which increases $P_\text{r}$ 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 logit($P_\text{r}$)=$\theta_k-\tau_k-\beta_j$. An example of a model that integrates accuracy parameters in a model for RT is Thissen’s (1983) model: $\ln(\text{RT}_{jk})=\mu+\tau_k+\xi_j+\rho(\alpha_i\theta_k-\beta_j)+

\varepsilon_{jk}$, where $\mu$ is a general intercept, $\tau_k$ and $\xi_j$ are slowness parameters of person $i$ and item $j$, respectively, $\rho$ determines the influence of the usual 2PL response parameter structure and $\varepsilon$ is a normally distributed error term.

Cognitive psychology meets psychometrics 21
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 model is the ‘fundamental equation of RT modeling’. According to van der Linden, the person speed parameter equals the amount of labor required to solve the item divided by RT. Hence, RT is the ratio of amount of labor and speed. A logarithmic transformation then gives $E(\ln(RT)|k) = \xi_j - \tau_k$. This equation is the basis for the RT 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 model are typical latent variables effecting either accuracy or RT. 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 above. The reason is that, in the Q-diffusion model, these are process parameters, which are of equal importance to accuracy and to RT. Since 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 use them within the same level of modeling.

Of course, an important advantage of the Q-diffusion model is that its extension to RT modeling need not be crafted, as it is the very basis of the model. Given appropriate data, we could fit the joint distribution of accuracy and RT as specified in equation 2. We can also use the equation for the mean RT (equation 5). It is informative to relate the mean RT prediction of the diffusion model to van der Linden’s fundamental equation of RT modeling. As explained in footnote 1, for reasonable high values of $a v$, expected RT equals $a/2v$. Hence:

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

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

$$\xi_j^* = \ln\frac{v_j}{a_j}; \tau_k = \ln\frac{v_k}{a_k}$$

(13)

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. Since van der Linden (2007) applied standard 2PL or 3PL to the response data, the translation of his difficulty and ability parameters is already specified in Equation 13, as the products of item and person drift rates and boundary separations, respectively.

These relations constrain the model at the second level of van der Linden’s 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 (for an example, 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, since 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 provisional solution, we suggest to define $a^p$ in units of information, and to view the time pressure $a^i$ 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 $\theta$) 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, del Pino and De Boeck, 2005, 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 DM-G, or the 'Rasch model with guessing' discussed above. But this model seems to fit worse 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 above, 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 he or she knows recognizes his or her failure); however, other serial or parallel setups are possible. According to San Martin, del Pino and De Boeck, the success of guessing will also depend 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 $\alpha$:

$$P_+ = \frac{e^{(\theta_k - \beta_j)}}{1 + e^{(\theta_k - \beta_j)}} + \left(1 - \frac{e^{(\theta_k - \beta_j)}}{1 + e^{(\theta_k - \beta_j)}}\right) \frac{e^{\alpha \beta_k + \gamma_j}}{1 + e^{\alpha \beta_k + \gamma_j}}$$

(14)

For $\alpha = 0$ this model reduces to the DM-G with guessing parameter equal to the expit of $\gamma$. In an empirical example, they show that this DM-AG better explains the data than the DM and the DM-G. However, a disadvantage of the DM-AG (or 1PL-AG, as San Martin, del Pino and De Boeck call it)
is that the p-process and g-process are not well separated anymore. If ability plays a too large role in guessing, the g-process could be equally successful, or even more successful than the p-process. Also, for very low ability the model predicts below chance responding (property 3, see San Martin, del Pino and De Boeck, 2005).

As Ramsay (1989) remarks, 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 the 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 ($P_g > 1/M$) 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) discuss several guessing scenarios and 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 for instance, Ratcliff, 2006). The DM-AG may serve as a basic model here.

Another model for guessing is introduced by Hessen (2004; 2005). Hessen investigates a subclass of the four parameter logistic IRM, for which both specific objectivity and sufficiency of the total score for the latent trait hold. One special case that Hessen (2004) proposes is:

$$P_g = \frac{\lambda}{1 + e^{\theta_k - \beta_j}}$$

This IRT model has upper asymptotes at 1, but lower asymptotes at $\lambda/e^{\beta_j}$. 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 will therefore call this model the DM-DG, for difficulty based guessing.

The prime attractiveness of Hessen’s 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 not to the DM-AG and the QM. Ramsay (1989) discusses a way for 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 item response theory, 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, & Van Heerden, 2004; Borsboom & Mellenbergh, 2007). 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. Since the processes that lie between item administration and item response are psychological in nature, the only way to remedy this situation is by constructing psychological theories of item response processes, and linking 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 paper may be that it gives proof of concept: It is possible to systematically connect latent variables to item responses through process models so as 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, the current paper proposed a fundamental difference between ability tests on the one hand, and personality and attitude test 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 $\beta$ 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 time intensity and speed parameters, can be translated to the basic diffusion parameters of drift rate and boundary separation. Van der Linden’s approach 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 for instance observed 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 will appear 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 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, as it extends the standard IRT paradigm with predictions on the behavior of RT 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, involving a series of computations for instance, a simple random walk may be too simplistic but 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, van Zandt, & McKoon, 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 model make perfect sense, as explained in Footnote 4, and we would recommend to use the D-diffusion item response model in this case (Equation 12). However we do not believe this to be a counterexample to our thesis of ability being essentially positive. The reason is that 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. Since 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.

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 RT’s, 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, i.e., by explicitly modeling the processes that lie between item administration and item response (Borsboom, Mellenbergh, & Van Heerden, 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


Cognitive psychology meets psychometrics 28


Cognitive psychology meets psychometrics 30

on the assumptions required to apply the 1PL and 2PL model to dichotomously scored multiple choice data.


