Psychological Test and Assessment Modeling, 2025.67:90-126
DOI: https://doi.org/10.2440/001-0020

Applicability of Process Models for the Joint Analysis of Responses and Response Times in Complex Cognitive Tasks

Raimund J. Krämer[1], Jochen Ranger[2], Marco Koch[3], Frank M. Spinath[3], Florian Schmitz[1]

1 Department of Psychology, University Duisburg-Essen
2 Department of Psychology, Martin-Luther-University Halle-Wittenberg
3 Department of Psychology, Saarland University

Abstract:
Process models for the joint analysis of responses and response times have been developed to disentangle different cognitive processes in experimental paradigms. More recently, they have been applied to complex tests of intelligence as well. However, the adequacy of the modelling approaches for such task types has been rarely tested. The present study compared two popular process models, a race model and a diffusion model, with the purely statistical hierarchical model in terms of relative fit to data from typical intelligence tests with varying response formats: a cube rotation test with a binary response format (n = 257), a figural matrix test with a distractor format (n = 229), a figural matrix test with a response construction format (n = 185), and a knowledge test (n = 3142). Compared to the diffusion model, the race and hierarchical models better described the data for all tests but the cube rotation test. Yet neither was able to adequately predict response time quantiles for the matrix construction or knowledge test. Model-based trait estimates displayed only moderate reliability, suggesting limited utility for the assessment of individual differences. This study highlights that process models can be useful for evaluating performance in complex tasks, but emphasizes to carefully consider model assumptions and task requirements.

Keywords: response latency, intelligence testing, assessment, model fit

Correspondence:

Raimund J. Krämer; Universitätsstraße 2, 45141 Essen, S06 S03 B76; Phone: +49 201 1836983; raimund.kraemer@uni-due.de

Introduction

Response Times in Cognitive Testing

"If any broad taxonomic classification of cognitive ability factors were to be formulated, in fact, it might be one based on the distinction between level and speed" (Carroll, 1993, p. 644). Most speed abilities are operationalized through tasks that require few mental processes to be solved correctly (elementary cognitive tasks) and for which response times are therefore the outcome of primary interest (Danthiir et al., 2005). Level abilities on the other hand are measured by the amount of correct responses to test items of varying complexity. In this context complexity does not refer to empirical item difficulty but is defined by task characteristics such as information load, the variety of information and the rate of change of information (Campbell, 1988). Typical tasks of fluid intelligence (gf), for example, can be considered complex due to their compositionality of several individual parts or rules that need to be combined or segmented (Duncan et al., 2017). Speed in the context of complex tasks has been studied within different frameworks such as speed of reasoning (Carroll, 1993), or time-on-task (Goldhammer et al. 2014). But still, the analysis of response times is largely limited to elementary cognitive tasks in ability research. We argue that response times in complex cognitive tasks should be considered for three reasons:

First, from a research perspective, if we are interested in the cognitive processes underlying cognitive abilities, process information is required. The time one needs to solve a task may serve as a source of information to identify or to quantify such processes (De Boeck & Jeon, 2019). Naturally, multiple process components and their flexible assembly may be involved in complex tasks. However, even if response times constitute the only available process information, they can help to identify different test-taking strategies, such as effortful processing versus rapid guessing (Wright, 2016) or cheating (van der Linden, 2011).

Second, when evaluating task performance, speed and level may be confounded due to item- or test-level deadlines or because persons apply strategies that favour speed over accuracy or the other way around (Goldhammer, 2015; Schweizer, 2025). For example, Borter et al. (2023) have shown that the correlation between performance on mental speed tests and reasoning tests with time-constraints decreases when controlling for speededness in reasoning. On the other hand, if no or very generous time-constraints are applied, test takers are free to decide how much time they invest in solving a task. In the context of complex tasks, investing more time does not necessarily increase the probability of a correct answer in the sense of a simple speed-accuracy trade-off. Instead, studies with response times as covariates reveal that the relationship of responses and response times varies in strength and direction, depending on person as well as item characteristics (Goldhammer et al., 2014; Stadler et al., 2020; Krämer et al., 2023; Weber et al., 2025). Consequently, it has been suggested to control for the effects of response times by calibrating item-level deadlines that account for item-time intensity as well as individual differences in speed

(Goldhammer, 2015). However, several challenges pertain to this approach (Schmitz & Wilhelm, 2015): Estimating item time intensity and person speed requires extensive pretesting which is hardly feasible in applied assessment settings. Further, the assumption needs to be made that person speed is temporarily stable and broadly reflected across diverse task types. Finally, controlling for speed in such a manner risks eliminating meaningful variance in cognitive ability, as processing speed is theoretically and empirically linked to higher-order cognitive ability constructs.

Third, while the effect of response times on responses can be statistically or experimentally controlled for in research, this is not easily feasible in an assessment context. However, specifically for the psychological assessment of cognitive abilities the relevance of complex cognitive tasks is undisputed, due to their high predictive validity for a variety of external criteria. For example, fluid intelligence is a good predictor of job performance and training, school grades, school achievement, and different life outcomes (Kyllonen & Kell, 2017). Crystallized intelligence (gc) has been shown to predict some of the aforementioned criteria as well, at times better than gf (Postlethwaite, 2011). Complexity in terms of gc depends heavily on the type of task in question. Reading comprehension, for example, can be considered complex because it requires the integration of phonological, orthographic, and semantic information (Kendeou et al., 2016). And even though the cognitive processes involved in solving gc tasks are presumably fundamentally different from those involved in a reasoning task, response times may still be worth looking upon. For example they could be used to distinguish knowledge retrieval processes from back-up strategies, in case the relevant information is not known (Chen et al., 2018).

In summary, abilities that feature complex tasks have a high relevance in psychological assessment. However, observed performance in such tasks is to some extent related to the time invested in solving the test items. Taking response times into consideration may therefore improve reliability (precision) and validity (predictive power) of the assessed ability estimate. This requires performance models that integrate both responses and response times.

Joint Models for Responses and Response Times

Statistical Models

Statistical models describe observable behaviour through parameters that are expected to reflect the latent value of an unobservable variable (e.g., ability or speed) and are subject to measurement error (Frischkorn & Schubert, 2018). Statistical models can integrate responses and response times either through distinct unrelated models, distinct but related models or through one joint model that accounts simultaneously for both variables (van der Linden, 2009). It has been shown that statistical models with related ability and speed fit mathematical test data better than approaches with unrelated traits (Hohensinn & Kubinger, 2017). Assuming that both traits are indeed

related, the estimation of ability is necessarily improved by additionally incorporating response times (Bolsinova & Tijmstra, 2018). The most prominent statistical model for the joint analysis of responses and response times in ability tests is the hierarchical model by van der Linden (2007). It was developed with the goal of explaining the distributions of both observable variables through separate models with distinct latent person and item parameters. The level-one measurement model for responses is an IRT model, originally the three-parameter normal-ogive model (3PNO). Based on the 3PNO, responses are determined by latent person ability as well as item discrimination, item difficulty and a guessing parameter. However, any model that fits the response (accuracy) data may be chosen instead, for example the 1PL or 2PL (van Rijn & Ali, 2017). The response times in the hierarchical model are assumed to be lognormally distributed, with the reciprocal standard deviation interpreted as discrimination. The expectation of the distribution is the difference score of latent person speed and item time-intensity. The response time model can also be flexibly replaced, for example with a Cox proportional hazards model (H.-A. Kang, 2017). Importantly, person parameters of both measurement models are considered constant across test items and only related on a second level, assuming a multivariate normal distribution in the population. Likewise, values of item parameters are drawn from a multivariate normal distribution, the so-called item domain.

The hierarchical model is a valuable approach for the joint analysis of responses and response times, because separate and easily interchangeable measurement models are used. It has been convincingly demonstrated that the hierarchical model can be applied to complex cognitive tasks (van der Linden, 2007; Glas & van der Linden, 2010; Shaw et al., 2020). It is, however, a purely statistical model with little explanatory value concerning the cognitive processes underlying the observable responses. In the present study, the hierarchical model serves mainly as a benchmark for the relative model fit of the tested cognitive process models.

Process Models

Responses or response times provide little insight into the nature of a cognitive process beyond its outcome and duration. Therefore, assumptions regarding the composition of the task-solving process are usually grounded on experimental research or theoretical considerations related to the nature and processing requirements of the cognitive task. They may concern the number, sequence, and duration of processes involved, or how these processes interact with each other and external variables. Cognitive process models formalize such assumptions through a set of interpretable parameters that are mathematically linked to observable variables (e.g., responses and response times; Frischkorn & Schubert, 2018). This allows the estimation of parameters that reflect the effective ability of a test-taker, adjusted for the influence of working speed or motivational aspects.

A prominent class of process models in laboratory research are continuous sampling models like the diffusion model (Ratcliff, 1978) and the race model (Townsend & Ashby, 1978), which use the information from responses and response times exhaustively. Both

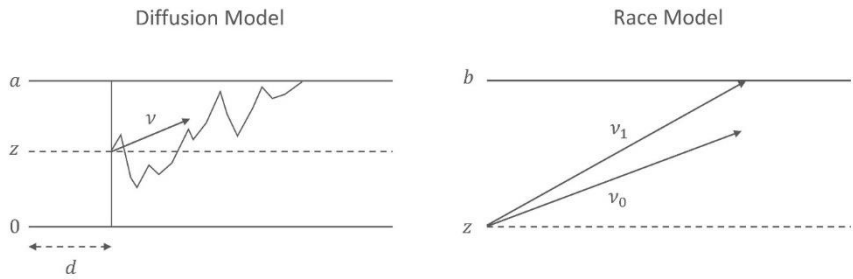
models assume an accumulation process, reflecting the accumulation of response evidence over time, typically in simple alternative-choice tasks. The speed of accumulation is called drift rate v. When a critical threshold is reached, a response is elicited. The nature of the accumulation process is not specified for either model family but is supposed to correspond with task-solving requirements. The diffusion model assumes that the accumulation process is subject to random within-trial fluctuations according to a Wiener process. Within-trial variation has been considered for race models as well (e.g., Usher & McClelland, 2001), but does not seem necessary to replicate typical response time distributions or a speed-accuracy trade-off (Brown & Heathcote, 2005). Instead, both types of models frequently incorporate sources of between-trial variation. Specifically, drift rates and the starting points of the drift process z are allowed to randomly fluctuate across items of a cognitive test (Ratcliff & Rouder, 1998; Brown & Heathcote, 2005).

Figure 1 shows schematic illustrations of the diffusion and the race model versions that have been applied in the present study. The fundamental difference between the two models is that the diffusion model assumes only a single accumulator that represents relative preference of one response option over the other. The race model on the other hand assumes one accumulator and a corresponding response threshold for each response option. The diffusion model predicts that a correct response is elicited when the accumulator reaches the upper threshold, whereas a false response is given when it reaches the lower threshold. The higher the drift rate, the higher is the probability of a correct response and the faster the response is given. The distance between the thresholds is called boundary separation a and is interpreted as response caution. When the boundary separation is narrower, responses are given faster but have a higher probability of being incorrect due to the stochastic nature of the drift process. The classic diffusion model has been developed for binary choice tasks, where each response threshold corresponds with one of two choices. However, the model has been expanded to accommodate alternative-choice tasks with more response options by recoding all responses as either correct or false (e.g., van der Maas et al., 2011).

In case of the race model, several accumulators compete with each other. Typically, they are assumed to represent the acquisition of information or knowledge with respect to one of the response options. The accumulator that reaches its threshold first, determines the response that is given as well as the response time. While usually the number of accumulators in a race model corresponds with the number of response options, an approach with only two accumulators was proposed by Ranger and Kuhn (2014). Here, one accumulator is content related whereas the second accumulator is conceived as a tendency to discontinue the task at hand due to frustration.

Both, diffusion model and race model families allow for the separation of the accumulation process from all non-decision processes, summarized as non-decision time d. What is actually incorporated in non-decision time remains a matter of discussion, but encoding of task requirements and execution of the motor response are typically assumed. Non-decision time is usually considered for diffusion models but has been added to race models as well (e.g., Brown & Heathcote, 2008).

Figure 1
Schematic Illustrations of the Diffusion Model and the Race Model



Note. Both models are continuous sampling models that postulate decision processes originating from starting point *z*. Information and noise are assumed to drive the decision process (response accumulator) until it reaches a response boundary, upon which the corresponding response is elicited. In the diffusion model, the upper and lower boundary correspond with a correct and incorrect response, respectively. The boundary separation a determines the amount of evidence that needs to be gathered before a response is elicited and, thus, corresponds with response caution in terms of a speed-accuracy setting. The drift rate *v* reflects the mean slope of the evidence accumulation process, and thus, corresponds with the speed or efficiency of task processing. The duration of all processes not directly involved in the evidence accumulation process are subsumed by a non-decision parameter *d*, which may comprise stimulus encoding or the execution of the motor response. In the race model, the accumulation process for correct responses (v_1) reflects the gathering of task-related knowledge, while the accumulation process for incorrect responses (v_0) reflects frustration. The first accumulator that reaches the response threshold *b* determines the response as well as the response time. In this version of the race model, the thresholds for correct and incorrect responses are identical. Neither of the two models specified here assumes an initial bias in the starting point z of the decision processes (i.e., no shift of *z* towards any of the of the boundaries). From an assessment perspective, the person components of the diffusion model drift rate and the race model drift rate for the correct response accumulator, are indicators of effective ability.

Process models have been extended in an IRT framework that allows to decompose person and item effects (Tuerlinckx & De Boeck, 2005). This promises to make these models better applicable for ability test where item difficulty typically varies a lot. For instance, the drift rate can be decomposed into a person parameter reflecting the ability or trait and an item parameter reflecting the easiness of the item. Similarly, boundary separation can be decomposed into a person's trait cautiousness and cautiousness triggered by the characteristics of an item.

Applicability of Process Models for Complex Cognitive Tasks

Only for about a decade have process models been used to model performance in complex cognitive tasks. Van der Maas et al. (2011) tested the diffusion model for a chess puzzle test, in which chess players had to find the best move for a given position on the board. The response data alone could be better explained by a 2PL as compared to the diffusion model. However, the drift rates were a slightly stronger predictor of chess *elo* (indicator of chess proficiency based on officially played matches) than the

responses. The same items were analysed by Ranger and Kuhn (2014) with their proportional hazards race model, assuming one information accumulator and a second accumulator that reflects the tendency to discontinue. The race model could better predict *elo* than the standard hierarchical model or a 2PL for responses only. For a figural reasoning test with 12 items, two different tests of model fit showed good fit of the race model for 11 and 8 of the items, respectively. Ranger et al. (2015) proposed a race model with two accumulators following a bivariate Birnbaum Saunders distribution. The model was applied to 15 items from the chess puzzle test and accounted well for the response time distributions in all but one item. The model parameters were equally good predictors of *elo* as compared to joint ability and speed from a 2PL for responses and a latent factor model for log response times. Notably, the accumulator for correct responses was also moderately correlated with performance motivation. For gc (as assessed with a spelling test) van Rijn and Ali (2017) compared different adaptions of the diffusion model. The data were described better by a diffusion model with varying versus constant boundary separation across test items and when intertrial variation for non-decision time was included. Similarly, for a verbal analogies test in a study by I. Kang et al. (2022), a diffusion model with random variability of starting point and drift rate performed best. In case of a matrix reasoning test intertrial variability of either starting point or drift rate improved the fit as well. Finally, Jin et al. (2023) for a proportional hazards race model and I. Kang et al. (2023) for a diffusion model, introduced an approach where residual dependencies between responses and response times are mapped onto the same latent space. Both reported reasonable reproducibility of the chess puzzle data based on model parameters.

Goals and Motivation of this Study

In conclusion, there is evidence that process models can be applied to complex cognitive tasks and that the additional integration of response times may increase the predictive power of latent trait estimates. However, several challenges pertain to the previously presented studies. Firstly, all but three studies used the same data set to evaluate their model (van der Maas & Wagenmakers, 2005), since it contains the *elo* ratings as an external criterion variable. And while undoubtedly complex, chess puzzling is not a typical test of intelligence and strongly dependent on training and experience. Secondly, there is a lack of clarity whether or how process models can be applied to tasks with different response formats. Previous research indicates that the response format of a cognitive test has no influence on the measured trait, however, this does not exclude the possibility of conceptually different problem-solving strategies as manifestations of the same underlying construct (Hohensinn & Kubinger, 2011, 2016). Process models make theoretical assumptions concerning the cognitive processes involved in solving a task that may or may not align with these different strategies. For instance, van Rijn and Ali (2017) did not fit a diffusion model for tasks with multiple choice options or an open response format as they did not consider the approach suitable. In other studies, diffusions models or race models with two accumulators have been applied to such tasks without discussing or testing their suitability. To cope with multiple choice options, van der Maas et al. (2011) suggested a

correction of the diffusion model that decreases drift rate estimates as the number of response options increases. This appears plausible only when test-takers employ primarily response elimination strategies or when response alternatives are equally plausible. Thirdly, both diffusion models and race models have been used for dissimilar test contents without much consideration regarding the cognitive processes involved. For instance, an information accumulation process seems appropriate for gf tests where a set of rules is inferred one-by-one. By contrast, gc tests require the retrieval of information from declarative long-term memory, and it is less certain that this also follows a sequential accumulation of evidence.

Both gf and gc feature complex tasks and represent comprehensive cognitive abilities with a high predictive validity for diverse life and job outcomes. Process model accounts of these abilities are particularly valuable from an assessment perspective, as they provide ability estimates that consider the influence of work pace and, to some extent, motivational factors. However, the applicability of process models for applied assessment purposes requires on the one hand that the models can describe the data sufficiently well, and on the other hand, that model-based traits are estimated reliably. Consequently, the first goal of the present study is to evaluate the adequacy of process models for describing data from different complex tasks. Based on the result, we give practical recommendations as to which type of model or parametrizations might be applied with reference to task characteristics. For this purpose, a diffusion model and a race model are compared to the purely statistical hierarchical model in terms of model fit. These model families have been the most popular approaches for the joint analysis of responses and response times in recent decades. They are applied to three tests of fluid intelligence and one knowledge test as a measure of crystallized intelligence. While most gc measures focus on verbal abilities, it has been argued that tests of declarative knowledge are better suited indicators, because they can be more clearly distinguished from gf and general intelligence g (Schipolowski et al., 2014). Fluid intelligence was assessed with two figural matrix tests and a cube rotation test, both task types representing highly established measures of gf (Kyllonen & Kell, 2017; Lohman, 1996). Moreover, tests with figural task content exhibit the strongest correlations with general fluid intelligence (Wilhelm & Schroeders, 2019). As the assumptions of process models are closely linked to the response format of the presented tasks, tests were selected that vary systematically in this respect. Specifically, one gf test applied a binary response format, one a multiple-choice format, and the last one an open response format. The models are evaluated with respect to fit to the intelligence test data and with respect to the plausibility of the estimated parameters.

The second goal of the present study is the analysis of psychometric properties of latent person parameters from the investigated process models. To this end, we estimate and compare the person parameters and their respective standard errors for each model and intelligence test. Finally, the person parameters for two of the intelligence tests are related to self-reports of personality, working pace, and motivational factors.

Methods

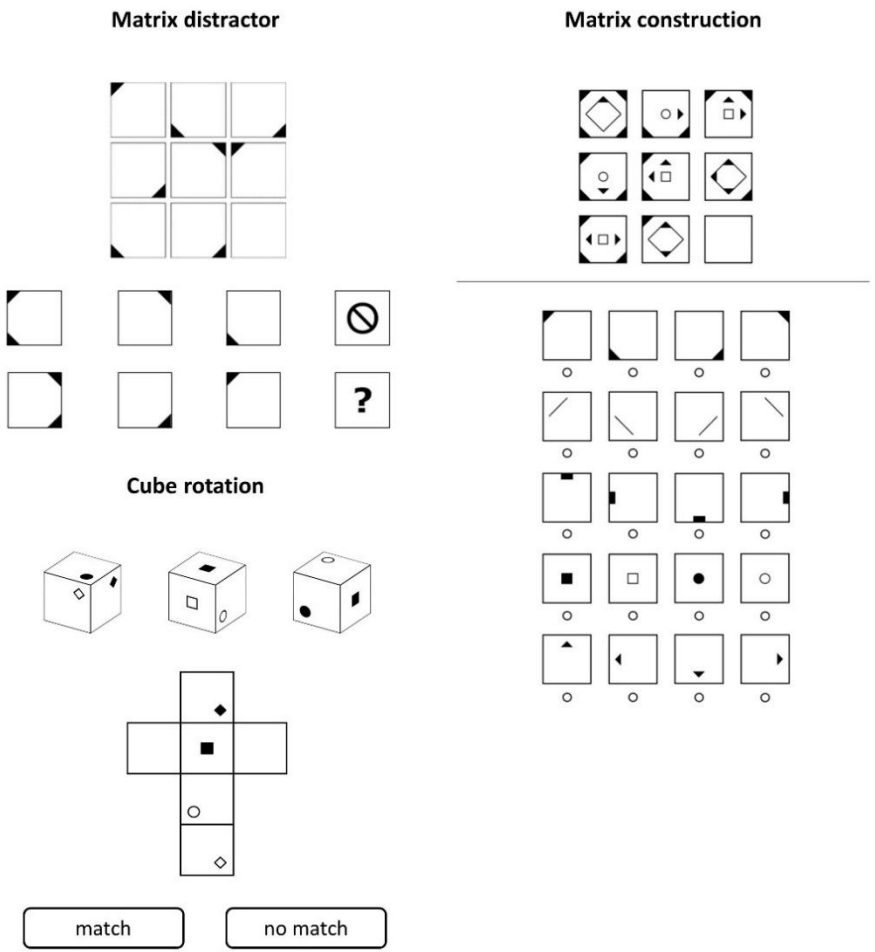
Intelligence Tests

Exemplary items for each of the fluid intelligence tests are shown in Figure 2. Two tests are figural matrix tests featuring items from the open matrices item bank (OMIB, Koch et al., 2022). Test-takers are presented with a 3x3 Raven-type matrix, where each cell displays one or several simple geometric shapes. The shapes are arranged so that each row follows the same set of 6 possible rules, specifically addition, subtraction, disjunctive union, intersection, rotation, and completeness. Up to five rules per item are applied simultaneously. The last cell in the bottom row remains empty and must be filled according to the set of rules inferred from the previous rows. In the response construction format test-takers must assemble the missing cell using 20 construction elements, each consisting of one geometric shape in a specific position. In the more conventional distractor format, test takers must choose from eight different response options: six display combinations of the geometric shapes in specific locations, one indicates that none of the six displayed solutions is correct, and another one indicates that the correct response is unknown to the test-taker. Distractors use the same geometric shapes that are depicted in the matrix but do not apply the rules correctly. All in all, 28 items had to be solved in the response construction format and 40 items in the distractor format.

The third gf test is a cube rotation test based on the cube construction test (Thissen et al., 2018). Test-takers are presented with three 3D images of the same cube from different perspectives. Each image shows three sides of the cube. One pair of images has two overlapping sides, one pair has one overlapping side, and the last pair has no overlap. Each side of the cube shows a simple geometric shape. The shape may be in the middle of the surface or in one of the four corners. Below the three 3D cube images an unfolded cube shows all six sides in one plane. Some sides of the unfolded cube surface may be blank and therefore non-informative. Test-takers are asked to decide whether the unfolded cube surface matches the 3D cube shown above or not. Test-takers had to solve a total of 42 cube rotation items.

Finally, the gc test is a knowledge test with questions from biology, chemistry, physics, and mathematics at the level of high school graduation. Test-takers are asked to choose one out of five response options by checking the adjacent box. An example is: “A peptide bond is a bond between an (A) ester bond, (B) C-C bond with unrestricted rotatability, (C) C-C double bond, (D) amide bond, (E) anhydride bond”. Test-takers were presented with 20 items.

Figure 2
Exemplary Items for the Fluid Intelligence Tests



Procedure

Data for the present study stems from two independent samples. The first data set was available and deemed well-suited to test the models, the second was specifically collected to complement this study. The first data set comprised the matrix construction test and the knowledge test and was collected as part of an online training session for the admission tests for German medical schools. Participation in the training was voluntary and not compensated. While the same items were presented to the whole

sample in case of the knowledge test, the matrix construction test was split into ten equally difficult subsets of 28 items each and randomly assigned to test-takers. For our analyses, only the subset with the largest sample size was analysed. The second data set comprised the matrix distractor test and the cube rotation test and was collected in an unrelated online study conducted at a German university. During this study, participants also filled out a number of questionnaires. Specifically the HEXACO-PI-R-100 (Lee & Ashton, 2018), German adaptions of the Short Boredom Proneness Scale (SBPS; Struk et al., 2017), the fatigue subscale of the Brunel Mood Scale (BRUMS; Terry et al., 2003), the Academic Boredom Scale (ABS-10; Acee et al., 2010), and self-constructed items to assess subjective speed-accuracy trade-off (e.g., "I tried to solve the tasks as carefully as possible, even if it took a little longer to get the result"). Descriptive statistics for the questionnaire data can be found in the supplementary materials (Table S1). The cube rotation test was administered before the matrix distractor test. After item 11 and item 25 of the matrix distractor test, the ABS-10 was presented as a measure of state boredom. Participation in the study was compensated with either course credits for Psychology students or 15 euros. For all tests response time in seconds was recorded from the presentation of an item until a confirmation button was pressed to log the response. Items were presented in a fixed order and, in case of the gf tests, with increasing complexity (i.e., number of combined rules).

Data Preparation and Sample

Two matrix distractor items and one matrix construction item were not displayed correctly and had to be removed. Furthermore, one knowledge item and four cube rotation items were not considered for further analyses because of negative item-total correlations. Data preparation and analyses were conducted independently for each intelligence test, but according to the same protocol. Participants younger than 18 or older than 40 were excluded from the analyses. This is because speed of processing is subject to major age-related changes in adolescence and old age (e.g., Nettelbeck & Burns, 2010). Test-takers who did not respond to all test items were removed as well. Further exclusion criteria were log response times above or below 2.5 times the interquartile range from the median for the respective item. However, due to the low-stakes setting of both study sessions there were many very fast inaccurate responses that were not marked as outliers by the said criterion. Consequently, test-takers that responded below six seconds for at least half of the items were excluded as well. Remaining responses times below six seconds and the respective responses were removed. Six seconds was determined as the minimum amount of time to fully comprehend item requirements (e.g., read the question and response options of a knowledge item). Responses were coded dichotomously as either correct or incorrect.

Of 355 persons solving at least one cube rotation item, $n = 257$ remained after applying the exclusion criteria listed above. Mean age within the sample was $M(SD)$ = 24.36(4.64), 67 % of the subjects were female and 72 % were students. In total 322 test-takers solved at least one matrix distractor item. The final sample size after all

exclusions was $n = 229$ with a mean age of $M(SD) = 24.32(4.54)$, 69 % female subjects and 72 % students. In the case of the matrix construction test 488 test-takers got presented with the subset of items analysed in the present study. After 303 exclusions, mostly due to incomplete responses, the sample size was $n = 185$. The subjects were on average $M(SD) = 20.09(2.16)$ years old, and 85 % were female. Since the admission to medical school requires a high-school degree with excellent grades, the sample can be assumed to be well educated. The same applies to the subjects from the knowledge test. Here the final sample size after 1385 exclusions was $n = 3142$, 76 % female with a mean age of $M(SD) = 20.31(2.44)$.

Statistical Analyses

Details concerning the modelling approaches are provided in the Appendix. All statistical analyses were conducted in R 4.4.0 (R Core Team, 2024). In a first step the three models were fit to responses and response times from the four intelligence tests using marginal maximum likelihood estimation (MML). In the MML approach item parameters are estimated by maximizing the likelihood function that has been integrated over the latent traits. For details concerning MML with respect to the different models see Glas and van der Linden (2010) for the hierarchical model, Ranger et al. (2015) for race models, and Molenaar et al. (2015) for the diffusion model. Said studies also report good recovery of model parameters for sample sizes varying between 96 and 317 test-takers. As the proportion of missing values was below 8 percent for all intelligence tests in the present study, no missing data techniques were applied. In a second step model fit was evaluated based on the Akaike Information Criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC, Schwarz, 1978), and the AIC corrected for small sample size (AICc; Hurvich & Tsai, 1989). Burnham and Anderson (2002) suggest the use of AICc instead of AIC when the ratio of sample size and number of parameters is below 40, which was the case for all data sets and fitted models. Subsequently, the MML item parameters were used to estimate the latent person parameters in a traditional maximum likelihood approach. To assess absolute model fit, person and item parameters were used to simulate response time distributions that were compared to the empirical distributions (conceptually similar to posterior predictive checks in a Bayesian framework, e.g. Klein Entink et al., 2009). As we were primarily interested if the models are suited to recover individual differences, we calculated the correlations of person-wise observed and predicted response time quantiles across items for correct and incorrect responses. Finally, the psychometric properties of the trait estimates were examined. Reliability was evaluated based on standard errors. Further, correlations of the trait estimates with the questionnaire data were investigated for the data sets that comprised these correlates (i.e., the cube rotation and matrix distractor tests).

Results

Table 1 shows the descriptive statistics for the intelligence tests. The difficulty of all tests was high with the proportion of correct responses ranging between $M_{Acc}(SD)$ = .43(.27) for the Matrix distractor test and $M_{Acc}(SD)$ = .70(.15) for the cube rotation test. With the exception of the cube rotation test ($M_{RT}(SD)$ = 27.15(18.28)) the average response times were comparable and only slightly below one minute. Internal consistency for responses as well as response times was acceptable to excellent, even in the case of the knowledge tests with 19 items only.

Table 1
Descriptive Statistics for the Intelligence Tests

Test	*n*	Items	Responses		Response times	
			M_{Acc} (*SD*)	ω_{Acc}	M_{RT} (*SD*)	ω_{RT}
Cube rotation	257	38	.70(.15)	.76	27.15(18.28)	.97
Matrix distractor	229	38	.43(.27)	.95	48.53(25.87)	.96
Matrix construction	185	27	.68(.25)	.92	52.96(11.57)	.87
Knowledge	3142	19	.49(.19)	.70	56.91(16.87)	.82

Note. Response time in seconds. Responses are proportion correct. ω_{Acc} = McDonald's omega based on observed item mean accuracies . ω_{RT} = McDonald's omega based on observed item mean response times.

The results of the formal model comparison are presented in Table 2. The cube rotation data with the binary response format were best described by the diffusion model. For all other tests it was the race model that best fit the data. However, since the race model was also the most complex one, the hierarchical model performed better according to AICc and BIC in case of the matrix distractor test and according to AICc in case of the matrix construction test. The average trial-wise difference between the best- and worst-fitting model exceeded 6.82 points for each information criterion and intelligence test, providing strong evidence that one model fit the data better (Burnham & Anderson, 2004).

Table 2
Information Criteria for the Models with Respect to Data from the Four Intelligence Tests

Test	Information criterion	Hierarchical model	Race model	Diffusion model
Cube rotation	AIC	37666	37823	37366
	AICc	38794	41725	38494
	BIC	38344	38636	38044
	−2LL	−18642	−18683	−18492
	k	191	229	191
Matrix distractor	AIC	43514	43511	43773
	AICc	45496	–[a]	45756
	BIC	44170	44297	44429
	−2LL	−21566	−21526	−21696
	k	191	229	191
Matrix construction	AIC	24382	23698	26429
	AICc	25158	26243	27206
	BIC	24820	24222	26867
	−2LL	−12055	−11686	−13079
	k	136	163	136
Knowledge	AIC	344399	342311	355616
	AICc	344405	342319	355622
	BIC	344980	343007	356197
	−2LL	−172103	−171040	−177712
	k	96	115	96

Note. AIC = Akaike information criterion; AICc = Akaike information criterion corrected for small sample sizes; BIC = Bayesian information criterion. *k* = number of estimated parameters. [a]The number of parameters equals the sample size resulting in a negative value for AICc.

The estimates of item parameters across test-items for all models and tests are presented in Figures S1 to S12. As expected, the item easiness parameter b_j of the hierarchical model decreased with increasing item complexity for all gf tests. However, only in case of the matrix construction test and to a smaller extent in the cube rotation test this was also associated with increasing item time-intensity β_j. For both measurement models the discrimination parameters a_j and α_j increased across items of either gf test, the only exception being constant discrimination for the speed factor in the cube rotation test.

For the race model, fixed item drift rates for incorrect responses α_{0j} increased across the cube rotation and matrix distractor tests. Not surprisingly, the fixed item drift rates for the content related accumulator β_{0j} decreased with increasing complexity of items in all gf tests. In case of the knowledge test, drift rates for both accumulators were very heterogeneous but strongly related. For both matrix tests, a growing impact of the latent trait for the correct response accumulator α_{1j} could be observed across items. The extent of random variation for both drift rates sv_{1j} and sv_{0j} as well as the effect of the latent trait for the incorrect accumulator β_{1j} was of comparable magnitude for all tests.

For the gf tests the fixed item drift rates of the diffusion model β_{0j} decreased across items as well as the boundary separation parameter α_{0j}. Only for the matrix construction test, an increase of response caution with increasing complexity could be observed. The other parameters of the diffusion model remained largely comparable across items in all tests. The response boundary parameter in the case of the knowledge test was very heterogeneous, but no trend emerged here either.

The predictive checks are presented in Table 3. There were only small differences in the reproducibility of the response time distributions between correct and incorrect responses. And while generally predictions were more accurate around the median of the distributions, correlations of observed and predicted 0.1 and 0.9 quantiles did not deviate strongly. However, there were considerable differences between the models and intelligence tests. The diffusion model only made accurate predictions for the response time distributions in the cube rotation test ($.54 \leq r \leq .92$). The hierarchical model and the race model revealed comparable fits, but the race model performed slightly better for all but the matrix construction test. Both models adequately reproduced the response time distributions for the cube rotation and matrix distractor test ($.74 \leq r \leq .98$) but did only moderately so in case of the matrix construction and knowledge test ($.01 \leq r \leq .59$). In particular fast correct responses in the matrix construction test were not adequately predicted by either model.

Table 3

Correlations of the Quantiles of Observed and Predicted Response Times

		Cube rotation			Matrix distractor			Matrix construction			Knowledge		
	Quantiles	HM	RM	DM	HM	RM	DM	HM	RM	DM	HM	RM	DM
Correct Responses	.10	.86	.89	.85	.64	.63	.22	.01	.28	.07	.38	.47	.22
	.30	.93	.96	.90	.77	.80	.42	.33	.39	.11	.45	.57	.26
	.50	.96	.97	.92	.79	.88	.40	.40	.51	.10	.45	.59	.28
	.70	.96	.98	.92	.78	.88	.47	.57	.53	.01	.42	.59	.28
	.90	.91	.94	.86	.69	.83	.45	.51	.43	.01	.34	.52	.25
Incorrect Responses	.10	.74	.84	.69	.82	.91	.44	.36	.46	.24	.45	.52	.37
	.30	.76	.88	.69	.85	.93	.41	.50	.44	.14	.50	.55	.36
	.50	.76	.89	.69	.84	.92	.36	.50	.43	.06	.51	.54	.33
	.70	.81	.90	.67	.82	.89	.28	.44	.35	.06	.49	.48	.31
	.90	.76	.84	.54	.72	.77	.19	.38	.28	-.04	.40	.42	.26

Note. HM = Hierarchical Model, RM = Race Model, DM = Diffusion Model. Responses and response times were predicted based on the trait estimates for the actual test-takers. Quantiles were calculated person-wise across items.

Table 4 shows the mean standard errors of the trait estimates on a standard normal scale. The difference score between the two race model trait estimates is also included, as it should directly relate to the proportion of correct responses in the tests and, consequently, ability in a traditional sense. Generally, the reliability of trait estimates in this study was limited to a minimum of $SE = 0.15$. The estimates for the knowledge test were the least reliable ($0.36 \leq SE \leq 0.71$), corresponding with the smallest number of available items. The differences between gf tests were mostly negligible. Only for the cube rotation test, the ability parameter of the hierarchical model ($SE = 0.52$) and drift rate of the diffusion model ($SE = 0.62$) had particularly large standard errors. Likewise, differences between parameters were comparably small.

Table 4
Mean Standard Errors of Estimated Trait Parameters

Task	Hierarchical model		Race model			Diffusion model	
	θ_{HM}	ω_{HM}	θ_{RM}	ω_{RM}	θ_{RM}-ω_{RM}	θ_{DM}	ω_{DM}
Cube rotation	0.52	0.16	0.23	0.21	0.31	0.62	0.16
Matrix distractor	0.32	0.16	0.36	0.15	0.40	0.27	0.19
Matrix construction	0.32	0.20	0.32	0.28	0.41	0.22	0.34
Knowledge	0.70	0.36	0.59	0.39	0.71	0.58	0.52

Note. Trait estimates are on a standard normal distribution scale. θ_{HM} = hierarchical model ability, ω_{HM} = hierarchical model speed, θ_{RM} = race model accumulator for correct responses, ω_{RM} = race model accumulator for false responses, θ_{DM} = diffusion model drift rate, ω_{DM} = diffusion model response caution.

Finally, we inspected the correlations of trait estimates for the cube rotation and matrix distractor tests with questionnaire data (Table 5). While most correlations were negligible, state boredom and fatigue were weakly related to several trait estimates. In case of the cube rotation test this includes negative correlations with ability in the hierarchical model, drift rate in the diffusion model and the difference score in the race model. For the matrix distractor test, the correlation between state boredom and all trait parameters was consistently negative ($-.35 \leq r \leq -.11$), with the exception of the accumulator of false responses in the race model ($r = .18$). The subjective speed-accuracy trade-off was not reflected in the response caution estimates. For both tests there were several weak correlations between trait estimates and personality factors. Notably, in case of the cube rotation test, there were associations of extraversion with latent speed ($r = -.26$), both traits in the race model ($r = .26$; $r = .24$), and response caution in the diffusion model ($r = -.29$). The same effects were observed to a lesser extent for the matrix distractor test. Here, neuroticism was negatively related to ability in the hierarchical model ($r = -.20$), and the difference score in the race model ($r = -.16$).

Table 5
Correlations of Trait Estimates with Questionnaire Data

Questionnaire	Hierarchical model		Race model			Diffusion model	
	θ_{HM}	ω_{HM}	θ_{RM}	ω_{RM}	θ_{RM}-ω_{RM}	θ_{DM}	ω_{DM}
				Cube Rotation Test			
Trait boredom	−.11	−.09	.08	.09	−.02	−.13	−.08
State boredom	−.10	.04	−.07	.02	−.18	−.16	.02
Subjective SATO	.06	.09	−.09	−.10	.03	.07	.09
Mental fatigue[1]	−.15	.00	−.03	.05	−.20	−.24	−.01
Openness to experience	.07	−.06	.08	.01	.15	.05	−.06
Conscientious-ness	−.15	−.15	.14	.15	−.03	−.07	−.16
Extraversion	−.19	−.26	.26	.24	.05	−.11	−.29
Agreeableness	.03	−.06	.07	.02	.08	.03	−.03
Neuroticism	−.13	−.07	.05	.07	−.07	−.13	−.09
				Matrix Distractor Test			
Trait boredom	.02	−.07	.13	.05	.09	.03	−.05
State boredom	−.27	−.11	−.18	.18	−.35	−.27	−.15
Subjective SATO	.15	.09	.09	−.13	.22	.13	.13
Mental fatigue[2]	−.27	−.14	−.11	.22	−.32	−.23	−.17
Openness to experience	.02	.00	.06	−.01	.08	.01	−.01
Conscientious-ness	−.08	−.11	.14	.09	.07	−.09	−.11
Extraversion	−.13	−.14	.14	.14	.02	−.10	−.16
Agreeableness	.12	.02	.11	−.06	.17	.14	.02
Neuroticism	−.20	−.12	−.03	.14	−.16	−.14	−.08

Note. Mental fatigue was measured after the cube rotation[1] and the matrix distractor[2] tests, respectively. State boredom was aggregated across three measurement points during the matrix distractor task. Subjective SATO = self-reported speed-accuracy trade-off. θ_{HM} = hierarchical model ability, ω_{HM} = hierarchical model speed, θ_{RM} = race model accumulator for correct responses, ω_{RM} = race model accumulator for false responses, θ_{DM} = diffusion model drift rate, ω_{DM} = diffusion model response caution. Correlations are corrected for attenuation due to unreliability of the trait estimates using their mean standard errors.

Discussion

The present study suggests considerable differences between the three models in terms of fit to data from intelligence tests. The diffusion model turned out to possess inferior fit when compared to the race model and the hierarchical model for all but the cube rotation test. This pattern may not necessarily reflect generally incorrect model assumptions but result from the specific modelling choices in this study. For each of the three model families various extensions and alternative approaches of parametrization have been proposed. This study only considered the most basic versions of each model for several reasons. Firstly, the MML approach of parameter estimation is computationally intensive, making the subsequent testing of various alternative parameterizations a serious obstacle. Potentially, a hierarchical model with conditional dependencies (e.g., Bolsinova et al., 2017), a race model with accumulators for each response option or a diffusion model with random inter-trial variation might offer better accounts of the data. However, the higher complexity of these approaches also constitutes a hurdle for their integration in applied diagnostic settings. Secondly, we chose models that could be fitted to all tasks at hand. For instance, the open response format for one of the figural matrix tests does not allow for a race model with more than two accumulators. Finally, all models fitted in this study were highly comparable with respect to complexity in terms of the number of parameters (five or six parameters for each item). Nevertheless, the restriction to specific parameterizations represent a limitation of this study that will be further discussed in the following.

For instance, no random variation of the parameters across trials was implemented in case of the diffusion model. This assumption might be violated in complex tasks where the task requirements are more heterogeneous across trials than in experimental paradigms. In fact, previous research has shown that allowing for random inter-trial variability can increase model fit (I. Kang et al., 2022). Another challenge when fitting the diffusion model may have resulted from constraining the non-decision time parameter to be no smaller than the fastest observed response time for the respective item. This makes sense from a theoretical perspective, since even a fast guess requires processes subsumed by the non-decision time parameter, like executing a motor response. However, in the present study, data was pre-processed in a way that responses faster than six seconds were excluded. This artificial lower boundary might bias the estimation of non-decision time, and consequently of the other parameters.

Taking into consideration that the race model revealed better fit in the absence of a non-decision time parameter, the relevance of non-decision time for modelling complex tasks can be questioned. Non-decision time is assumed to capture the time for encoding and response execution, but both may play a relatively minor role in complex tasks with average response times of 30 seconds and more. Further, the meaning of encoding is only vaguely defined. Given that this also comprises semantic or elaborative encoding in complex tasks, these processes may be hardly separable from those contributing to the information accumulation process. Encoding efficiency has been shown to be related to intelligence (for example Cusack et al., (2009), using a

change detection task), suggesting that encoding reflects cognitive abilities to some extent. Possibly, omitting the non-decision time parameter could help subsume all ability-related processes in the drift rate parameter.

The comparably good fit of the diffusion model to the cube rotation test suggests particularities of this test that makes it suitable for diffusion model analyses. Specifically, this test had a binary response format, whereas all other tests required the binary recoding of multiple response alternatives as correct vs. false. This aggregation of (dissimilar) incorrect response processes may not be warranted in some cases and deteriorate model fit. In the present study, negative drift rates were observed for multiple items in the matrix distractor and knowledge tests. However, when all incorrect response alternatives are associated with the lower boundary, it remains open which processes actually result in the false response. Moreover, in the version using an open response format, a linear negative drift rate that describes consistent accumulation of wrong information is implausible from a theoretical stance. For example, a test-taker might infer four of five rules correctly but still respond incorrectly. Accordingly, the diffusion model was not suited to account for observed response time data from the matrix construction test.

By contrast, the race model for binary (recoded) responses assumes that the two accumulation processes are executed simultaneously. Therefore, a high drift rate for the content related accumulator and an incorrect response do not mutually exclude each other. Importantly, the accumulator associated with incorrect responses has been suggested to capture frustration. Such a process might well follow a linear trajectory if frustration continues to increase until the task is either solved or discontinued. This might even be the case for a knowledge test, where frustration increases with unsuccessful retrieval of information. Furthermore, the accumulation of frustration is not directly related to the number of response options. For open response formats in particular, a race model with one content unrelated accumulator represents a useful approach, as accumulators cannot correspond with different response options. This flexibility concerning the interpretation as well as the number of accumulators makes the race model more universally applicable as the diffusion model, from a theoretical perspective. The hierarchical model served as a reference for model fit in the present study. As expected, it fit the data from all four tests comparably well, while very little theoretical assumptions needed to be made regarding task processes and response format. However, the hierarchical model does not yield much benefit for pragmatic assessment purposes, because its model parameters strongly overlap with classical performance scores.

We tested the capability of the models to reproduce observed data in terms of predictive checks. The response time data from the cube rotation and matrix distractor tests could be replicated better than those of the matrix construction and knowledge tests. In case of the latter this is presumably due to item heterogeneity rather than characteristics of the fitted models. In fact, the internal consistency of the knowledge test was the weakest of the four intelligence tests, likely due to the high content specificity of test items in the present study. A principal component analysis of the item responses

revealed a three-factor structure, potentially reflecting different knowledge domains. However, in case of the hierarchical model the measurement model for response times fit the data even worse as compared to the response model. This suggests that item response times in knowledge tests are only partially accounted for by a single underlying person characteristic. This begs the question to what extent latent trait models are suited for tests of content-specific knowledge. Concerning gc in general, van Rijn and Ali (2017) reported substantial misfit for a scoring rule model applied to a spelling test. However, I. Kang et al. (2022) found that a diffusion model adequately reproduced response time distributions in a verbal analogies test. While process models may generally offer a plausible account for gf tests that require the uptake, inference, and combination of task rules across time, the situation is less clear for gc tests. From a theoretical stance, it is not directly evident why search in declarative memory should follow a linearly increasing function across invested solving time. In fact, one might suspect that knowledge is either relatively early accessible (after minimal retrieval processes) or not. In line with this, Chen et al. (2018) consistently demonstrated for three knowledge tests that the probability of a correct response follows an inverse U-shaped function. If process models should be applied to knowledge tests at all, it might be indicated to approximate the accumulation functions piecewise until a general pattern (e.g., inverse U-shaped) emerges.

The poorest model fit of all tests was obtained for the matrix construction test. The matrix construction test was also the only test where time intensity as well as item specific boundary separation increased with test duration. While such an increase in response caution due to increasing item complexity could be expected on theoretical grounds, it is unlikely that this is the sole cause in the present data. First, no increase in response caution was observed in the other two gf tests with forced choice responses, despite analogically increasing item complexities. Second, current literature suggests that response caution is more likely to generally decrease across trials of a test (Ranger et al., 2023). We suspect that the particularly pronounced increase in response times in the matrix construction test reflects to a large extent the increasingly time-intensive response construction due to a higher number of elements that have to be assembled for more complex items. In turn, the diffusion model may experience difficulty to adequately separate boundary separation and non-decision time, given that the latter parameter should subsume the said motor responses. In case of the race model, the longer response times presumably translate to lower drift rates of both accumulators, which showed a more negative trajectory across items as compared to the other gf tests. This hypothesis could be tested in future research by using a test with an open response format but faster entry of responses (e.g., typing a number). It should be noted, that differences in absolute model fit between tests could be related to the different samples and testing conditions. Specifically, the low-stakes training setting in case of the matrix construction and knowledge tests might have resulted in careless responding in some cases, which would deteriorate overall model fit. The same exclusion criteria were applied to both data sets in order to ensure a certain degree of comparability. Nevertheless, the possibility of systematic differences in test-taking strategies or motivational factors in the two samples cannot be excluded. Varying

sample characteristics, for example concerning the age of test-takers, could be another factor potentially accounting for test differences.

Another goal of this study was to evaluate the psychometric properties of the estimated person parameters for diagnostic purposes. For the gf tests there were no systematic differences between models in terms of mean standard errors of the ability estimates θ. The only exception were higher standard errors for the ability parameter of the hierarchical model in case of the cube rotation test, most likely due to lower item discrimination. For the remaining models and tests the mean standard errors varied around 0.2 to 0.3, which corresponds with confidence intervals of about one standard deviation on a standard normal scale and therefore a considerable amount of uncertainty. As the even higher standard errors for the knowledge test with 19 items suggest, the reliabilities of all estimates are primarily limited by the low number of trials. However, substantially increasing the item numbers for the assessment of a single ability is not feasible in diagnostic settings for reasons of time economy. This imposes a constraint on the applicability of process models for applied diagnostics.

It would be desirable to validate the trait estimates using an external ability marker. In this study, only questionnaire self-reports of motivational states were available. As anticipated, fatigue and boredom were consistently, though weakly, correlated with the model parameters. Interestingly, their relationship with the correct response accumulator was weaker than their relationship with the ability parameter in the hierarchical model. However, the correlations with the estimates for the incorrect response accumulator, that should primarily reflect the effects of frustration and other negative states, were barely stronger. In case of the diffusion model mostly the drift rate estimates were related to low boredom and low fatigue. This might plausibly reflect reduced speed of information accumulation due to fatigue-induced working memory impairment (Ilkowska & Engle, 2010). Similarly, information accumulation may be hindered by interruptive non-decision processes, such as mind-wandering (Boehm et al., 2021), which is an aspect of boredom. It is of note that also the personality trait extraversion was related with generally faster or less cautious responding, potentially due to higher risk propensity (Nicholson et al., 2005). Based on the questionnaire data, it appears that process models only partially distinguish effective ability from motivational factors. However, the relationships between model parameters and affective variables should be interpreted with caution, as they were empirically weak and could be reconciled with theory only indirectly.

Conclusion

This study investigated the relative fit of a hierarchical model, a race model and a diffusion model for complex cognitive tasks. It was shown that not all models are equally suited to describe data from typical tests of intelligence with different response formats. The diffusion model displayed adequate fit only to data from a gf test with a binary response format. While the race model and the hierarchical model seemed more

universally applicable, both failed to predict response times from a knowledge test and a gf test using a response construction format. The reliability of all trait estimates was compromised by the moderate number of trials in the intelligence tests. Finally, some of the model-based trait estimates displayed plausible relations in magnitude and direction to questionnaire data.

The practical implications of this study are twofold: first, regarding the applicability of process models for the joint analysis of responses and response times in complex cognitive tasks, and second, regarding the usability of process model trait estimates for psychological assessment. Generally, even though the tested process models were originally developed for elementary tasks, they can be used to analyse data from tests of fluid intelligence. For the analysis of knowledge tests or other tests of crystallized intelligence we suggest model versions that do not assume linear accumulation of evidence but curvilinear accumulation or those that approximate the accumulation functions piecewise. Further considerations on the choice of model should be grounded in the response format. For a binary response format, diffusion modelling may be adequate, for multiple choice formats a race model with two accumulators or one accumulator per response option might be better suited. Finally, regarding parametrization, we suggest to critically review the necessity of a non-decision time parameter. Furthermore, if sufficient unique data points are available, the models might profit from allowing drift rate parameters, and potentially boundary separation or starting point parameters, to randomly vary across trials.

In applied diagnostic settings, tests of general intelligence, that include a wide variety of task types, are prevalent. For economic reasons the number of trials per task is often limited. This poses a challenge for trait estimates derived as parameters from process models, as the low number of trials can result in compromised reliability. However, for adaptive testing or narrower tests of cognitive ability with few scales of sufficient length, process modelling might be a promising approach. Validity of the trait parameters could not be conclusively addressed in this study. Further studies should include suitable external criteria to this end.

References

Acee, T. W., Kim, H. J., Kim, J.-I., Chu, H.-N. R., Kim, M., Cho, Y., & Wicker, F. W. (2010). Academic boredom in under- and over-challenging situations. *Contemporary Educational Psychology*, *35*(1), 17–27. https://doi.org/10.1016/j.cedpsych.2009.08.002

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Boehm, U., Marsman, M., van der Maas, H. L. J., & Maris, G. (2021). An Attention-Based Diffusion Model for Psychometric Analyses. *Psychometrika*, *86*(4), 938–972. https://doi.org/10.1007/s11336-021-09783-0

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling Conditional Dependence Between Response Time and Accuracy. *Psychometrika*, *82*(4), 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Borter, N., Schlegel, K., & Troche, S. J. (2023). How Speededness of a Reasoning Test and the Complexity of Mental Speed Tasks Influence the Relation between Mental Speed and Reasoning Ability. *Journal of Intelligence*, *11*(5), 89. https://doi.org/10.3390/jintelligence11050089

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*(1), 117–128. https://doi.org/10.1037/0033-295x.112.1.117

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer. https://doi.org/10.1007/b97636

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Campbell, D. J. (1988). Task Complexity: A Review and Analysis. *Academy of Management Review*, *13*(1), 40–52. https://doi.org/10.5465/amr.1988.4306775

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Chen, H., De Boeck, P., Grady, M., Yang, C.-L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence*, *69*, 16–23. https://doi.org/10.1016/j.intell.2018.04.001

Cusack, R., Lehmann, M., Veldsman, M., & Mitchell, D. J. (2009). Encoding strategy and not visual working memory capacity correlates with intelligence. *Psychonomic Bulletin & Review*, *16*(4), 641–647. https://doi.org/10.3758/PBR.16.4.641

Danthiir, V., Roberts, R. D., Schulze, R., & Wilhelm, O. (2005). Mental Speed: On Frameworks, Paradigms, and a Platform for the Future. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 27–46). SAGE Publications. https://doi.org/10.4135/9781452233529.n3

De Boeck, P., & Jeon, M. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, *10*, 102. https://doi.org/10.3389/fpsyg.2019.00102

Duncan, J., Chylinski, D., Mitchell, D. J., & Bhandari, A. (2017). Complexity and compositionality in fluid intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(20), 5295–5299. https://doi.org/10.1073/pnas.1621147114

Frischkorn, G. T., & Schubert, A.-L. (2018). Cognitive Models in Intelligence Research: Advantages and Recommendations for Their Application. *Journal of Intelligence*, *6*(3), 34. https://doi.org/10.3390/jintelligence6030034

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *The British Journal of Mathematical and Statistical Psychology*, *63*(3), 603–626. https://doi.org/10.1348/000711009X481360

Goldhammer, F. (2015). Measuring Ability, Speed, or Both? Challenges, Psychometric Solutions, and What Can Be Gained From Experimental Control. *Measurement : Interdisciplinary Research and Perspectives*, *13*(3-4), 133–164. https://doi.org/10.1080/15366367.2015.1100020

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626. https://doi.org/10.1037/a0034716

Hohensinn, C., & Kubinger, K. D. (2011). Applying Item Response Theory Methods to Examine the Impact of Different Response Formats. *Educational and Psychological Measurement*, *71*(4), 732–746. https://doi.org/10.1177/0013164410390032

Hohensinn, C., & Kubinger, K. D. (2016). On Varying Item Difficulty by Changing the Response Format for a Mathematical Competence Test. *Austrian Journal of Statistics*, *38*(4), 231-239. https://doi.org/10.17713/ajs.v38i4.276

Hohensinn, C., & Kubinger, K. D. (2017). Using Rasch model generalizations for taking testee's speed, in addition to their power, into account. *Psychological Test and Assessment Modeling*, *59*(1), 93–108.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.2307/2336663

Ilkowska, M., & Engle, R. W. (2010). Trait and State Differences in Working Memory Capacity. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory, and executive control* (pp. 295–320). Springer. https://doi.org/10.1007/978-1-4419-1210-7_18

Jin, I. H., Yun, J., Kim, H., & Jeon, M. (2023). A latent space accumulator model for response time: Applications to cognitive assessment data. *Stat*, *12*(1), 1–18. https://doi.org/10.1002/sta4.632

Kang, H.-A. (2017). Penalized partial likelihood inference of proportional hazards latent trait models. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 187–208. https://doi.org/10.1111/bmsp.12080

Kang, I., De Boeck, P., & Partchev, I. (2022). A randomness perspective on intelligence processes. *Intelligence*, *91*, 101632. https://doi.org/10.1016/j.intell.2022.101632

Kang, I., Jeon, M., & Partchev, I. (2023). A Latent Space Diffusion Item Response Theory Model to Explore Conditional Dependence between Responses and Response Times. *Psychometrika*, *88*(3), 830–864. https://doi.org/10.1007/s11336-023-09920-x

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading Comprehension. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 62–69. https://doi.org/10.1177/2372732215624707

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Koch, M., Spinath, F. M., Greiff, S., & Becker, N. (2022). Development and Validation of the Open Matrices Item Bank. *Journal of Intelligence*, *10*(3), 41. https://doi.org/10.3390/jintelligence10030041

Krämer, R. J., Koch, M., Levacher, J., & Schmitz, F. (2023). Testing Replicability and Generalizability of the Time on Task Effect. *Journal of Intelligence*, *11*(5), 82. https://doi.org/10.3390/jintelligence11050082

Kyllonen, P., & Kell, H. (2017). What Is Fluid Intelligence? Can It Be Improved? In M. Rosén, K. Yang Hansen, & U. Wolff (Eds.), *Cognitive Abilities and Educational Outcomes. Methodology of Educational Measurement and Assessment* (pp. 15–37). Springer. https://doi.org/10.1007/978-3-319-43473-5_2

Lee, K., & Ashton, M. C. (2018). Psychometric Properties of the HEXACO-100. *Assessment*, *25*(5), 543–556. https://doi.org/10.1177/1073191116659134

Lohman, D. F. (1996). Spatial ability and g. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: their nature and measurement* (pp. 97–116). Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203774007

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). Fitting Diffusion Item Response Theory Models for Responses and Response Times Using the R Package diffIRT. *Journal of Statistical Software*, *66*, 1–34. https://doi.org/10.18637/jss.v066.i04

Nettelbeck, T., & Burns, N. R. (2010). Processing speed, working memory and reasoning ability from childhood to old age. *Personality and Individual Differences*, *48*(4), 379–384. https://doi.org/10.1016/j.paid.2009.10.032

Nicholson, N., Soane, E., Fenton-O'Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, *8*(2), 157–176. https://doi.org/10.1080/1366987032000123856

Postlethwaite, B. E. (2011). *Fluid ability, crystallized ability, and performance across multiple domains: A meta-analysis* [Doctoral dissertation, University of Iowa]. https://doi.org/10.17077/etd.zopi8wvs

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* (Version 4.4.0) [Computer software].

Ranger, J., & Kuhn, J.-T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *The British Journal of Mathematical and Statistical Psychology*, *67*(3), 388–407. https://doi.org/10.1111/bmsp.12025

Ranger, J., Kuhn, J.-T., & Gaviria, J.-L. (2015). A Race Model for Responses and Response Times in Tests. *Psychometrika*, *80*(3), 791–810. https://doi.org/10.1007/s11336-014-9427-8

Ranger, J., Wolgast, A., Much, S., Mutak, A., Krause, R., & Pohl, S. (2023). Disentangling Different Aspects of Change in Tests with the D-Diffusion Model. *Multivariate Behavioral Research*, *58*(5), 1039–1055. https://doi.org/10.1080/00273171.2023.2171356

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Schipolowski, S., Wilhelm, O., & Schroeders, U. (2014). On the nature of crystallized intelligence: the relationship between verbal ability and factual knowledge. *Intelligence*, *46*, 156–168. https://doi.org/10.1016/j.intell.2014.05.014

Schmitz, F., & Wilhelm, O. (2015). Item-Level Time Limits Are Not a Panacea. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 182–185. https://doi.org/10.1080/15366367.2015.1115300

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Schweizer, K. (2025). Discrimination between types of common systematic variation in data contaminated by method effects using CFA models. *Psychological Test and Assessment Modeling*, *67*(1), 3–22.

Shaw, A., Elizondo, F., & Wadlington, P. L. (2020). Reasoning, fast and slow: How noncognitive factors may alter the ability-speed relationship. *Intelligence*, *83*, 101490. https://doi.org/10.1016/j.intell.2020.101490

Stadler, M., Radkowitsch, A., Schmidmaier, R., Fischer, M., & Fischer, F. (2020). Take your time: Invariance of time-on-task in problem solving tasks across expertise levels. *Psychological Test and Assessment Modeling*, *62*(4), 517–525.

Struk, A. A., Carriere, J. S. A., Cheyne, J. A., & Danckert, J. (2017). A Short Boredom Proneness Scale. *Assessment*, *24*(3), 346–359. https://doi.org/10.1177/1073191115609996

Terry, P., Lane, A., & Fogarty, G. (2003). Construct validity of the Profile of Mood States — Adolescents for use with adults. *Psychology of Sport and Exercise*, *4*(2), 125–139. https://doi.org/10.1016/S1469-0292(01)00035-8

Thissen, A., Koch, M., Becker, N., & Spinath, F. M. (2018). Construct Your Own Response: The cube construction task as a novel format for the assessment of spatial ability. *European Journal of Psychological Assessment*, *34*(5), 304–311. https://doi.org/10.1027/1015-5759/a000342

Townsend, J. T., & Ashby, F. G. (1978). Methods of Modeling Capacity in Simple Processing Systems. In N. J. Castellan Jr. & F. Restle (Eds.), *Cognitive Theory* (pp. 199–239). Erlbaum.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. https://doi.org/10.1037//0033-295x.108.3.550

van der Linden, W. J. (2007). A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual Issues in Response-Time Modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, *53*(3), 334–358.

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356. https://doi.org/10.1037/a0022749

van der Maas, H. L. J., & Wagenmakers, E.-J. (2005). A Psychometric Analysis of Chess Expertise. *The American Journal of Psychology*, *118*(1), 29–60. https://doi.org/10.2307/30039042

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *The British Journal of Mathematical and Statistical Psychology*, *70*(2), 317–345. https://doi.org/10.1111/bmsp.12101

Weber, D., Koch, M., Spinath, F. M., Krieger, F., & Becker, N. (2025). Log File Times as Indicators of Structured Figural Matrix Processing. *Journal of Intelligence*, *13*(6), 63. https://doi.org/10.3390/jintelligence13060063

Wilhelm, O., & Schroeders, U. (2019). Intelligence. In R. J. Sternberg & J. Funke (Eds.), *The Psychology of Human Thought: An Introduction* (pp. 255–275). Heidelberg University Publishing. https://doi.org/10.17885/HEIUP.470.C6677

Wright, D. B. (2016). Treating All Rapid Responses as Errors (TARRE) Improves Estimates of Ability (Slightly). *Psychological Test and Assessment Modeling*, *58*(1), 15–30.

Appendix

We will now briefly present the exact version of each model applied in the present study. For none of the four intelligence tests a 3PL model explained the response data considerably better than a 2PL model, therefore the latter was chosen for the hierarchical model. In the 2PL model, the probability of a correct response of the test-taker *i* with ability θ_i on item *j* is

$$p(x_{ij} = 1|\theta_i, b_j, \alpha_j) = L\left(a_j(\theta_i + b_j)\right). \quad (1)$$

The parameters a_j and b_j represent item discrimination and easiness, L is the logistic distribution function. The second measurement model is a log-normal model

$$f(t_{ij}|\omega_i, \alpha_j, \beta_j) = \frac{1}{\sqrt{2\pi}\sigma_j t_{ij}} \exp\left[-\frac{1}{2\sigma_j^2}\left(\log(t_{ij}) - (\beta_j - \alpha_j\omega_i)\right)^2\right], \quad (2)$$

where the response times t_{ij} of the test-taker are determined by latent person speed ω_i, a parameter β_j reflecting item time-intensity, and the reciprocal standard deviation σ_j^{-1} of the log response times, reflecting item discrimination. The parameter α_j is the regression coefficient for the relationship of item *j* with the latent speed factor, indicating discrimination as well. On the second level of modelling the joint distribution of responses, response times, and latent traits across *J* items of a test is given by

$$f(x, t, \theta, \omega) = \prod_{j=1}^{J} f(x_j|\theta) f(t_j|\omega) f(\theta, \omega). \quad (3)$$

Here, $f(\theta, \omega)$ denotes a multivariate normal distribution, $f(x_j|\theta)$ and $f(t_j|\omega)$ follow from equations (1) and (2).

For the race model, a version with two accumulators was chosen. This means that the accumulator for incorrect responses either represents the accumulation of frustration (see Ranger & Kuhn, 2014) or the accumulation of evidence for several incorrect response options at once. The defective density of response times for correct responses in this race model is

$$f(x_{ij} = 1, t_{ij}|\nu_{1ij}, \nu_{0ij}, s\nu_{1j}, s\nu_{0j}, b) = \frac{1}{t_{ij} s\nu_{1j}} \phi\left(\log\left(\frac{t_{ij}}{b}\right) + \frac{\nu_{1ij}}{s\nu_{1j}}\right) \left[1 - \Phi\left(\frac{log(\frac{t_{ij}}{b}) + \nu_{0ij}}{s\nu_{0j}}\right)\right]. \quad (4)$$

The density for incorrect responses follows from equation (4) by exchanging the parameters of the two accumulators. The parameter Φ is the normal distribution function and ϕ its density. The parameters ν_{1ij} and ν_{0ij} represent the expected values for the drift rates of the correct and incorrect accumulator, which vary randomly across trials according to a lognormal distribution

$$\log(v_{1ij}) \sim N(\beta_{1j}\theta_i - \beta_{0j}, sv_{1j}) \tag{5}$$

$$\log(v_{0ij}) \sim N(\alpha_{1j}\omega_i - \alpha_{0j}, sv_{0j}). \tag{6}$$

The drift rates were separated into person and item components. The parameters β_{0j} and α_{0j} determine the drift rates for test-takers with average traits, therefore representing fixed item effects. The item-specific relationship between the traits and the drift rates is captured in the parameters β_{1j} and α_{1j}. Similar to the hierarchical model we assumed dependencies of responses and response times to be conditional on the multivariate normal distribution of the latent traits with the correlation matrix $R_{\theta,\omega}$. There is no bias in the starting points of the drift. This is because in non-experimental settings featuring complex tasks there is no straightforward interpretation of bias, especially so when one accumulator represents multiple response options. Therefore, the distance either accumulator needs to pass is determined solely by the response threshold b, which has to be fixated on a constant value in the present model.

The joint density of responses and response times in the diffusion model is given by

$$f(x_{ij}, t_{ij}|\theta_i, \omega_i, \alpha_j, \beta_j, d_j) = \frac{\pi}{a_{ij}} \exp\left[a_{ij}v_{ij}\left(x_{ij} - \frac{1}{2}\right) - \frac{v_{ij}^2}{2}(t_{ij} - d_j)\right] \sum_{k=1}^{\infty} k \sin\left(\frac{1}{2}\pi k\right) \exp\left[-\frac{\pi^2 k^2}{2a_{ij}^2}(t_{ij} - d_j)\right], \tag{7}$$

where non-decision time d_j is treated as a fixed item parameter and constraint to be no smaller than the smallest observed response time for the respective item. Boundary separation a_{ij} and drift rate v_{ij} were separated into person and item components

$$\log(a_{ij}) = \alpha_{1j}\omega_i - \alpha_{0j} \tag{8}$$

$$v_{ij} = \beta_{1j}\theta_i - \beta_{0j} \tag{9}$$

The logarithm applies a positive constraint for the boundary separation. Considering the high complexity of the intelligence tests, no positive constraint was applied to the drift rate. This model is also unbiased, meaning there is no shift towards either response option for any item. Conditional independence was assumed for the diffusion model as well, the latent traits following a multivariate normal distribution with $R_{\theta,\omega}$. All in all, five parameters had to be estimated per item in case of the hierarchical model$(a_j, b_j, \alpha_j, \beta_j, \sigma_j)$ and the diffusion model $(\alpha_{1j}, \beta_{1j}, \alpha_{0j}, \beta_{0j}, d_j)$, and six in case of the race model $(\alpha_{1j}, \beta_{1j}, \alpha_{0j}, \beta_{0j}, sv_{1j}, sv_{0j})$.

Supplements

Table S1

Descriptive Statistics and Correlations of Questionnaire Data

Questionnaire	*M*	*SD*	Min	Max	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Trait Boredom	2.93	1.26	1.00	6.88									
2. State Boredom	4.92	1.60	1.00	7.00	.12								
3. Subjective SATO	4.27	1.09	1.40	7.00	−.04	−.27							
4. Mental Fatigue[1]	4.24	1.45	1.00	7.00	.19	.43	−.14						
5. Mental Fatigue[2]	5.00	1.80	1.00	7.00	.12	.72	−.22	.65					
6. Openness to Experience	4.04	0.66	2.13	6.00	−.03	−.12	.21	−.06	−.08				
7. Conscientiousness	4.21	0.87	2.50	6.69	−.07	−.15	.45	−.06	−.16	.22			
8. Extraversion	4.03	0.66	2.13	5.88	−.27	−.04	.09	−.02	−.06	.19	.31		
9. Agreeableness	3.98	0.63	1.75	5.75	−.03	−.19	.22	−.09	−.15	.17	.25	.21	
10. Neuroticism	4.23	0.73	2.25	6.75	.13	.17	.26	.12	.13	.14	.26	.10	.02

Note. Mental fatigue was directly assessed after the respective tests, i.e., after the cube rotation[1] and matrix distractor[2] tests. State boredom was aggregated across three measurement points during the matrix distractor test. The other self-reports were assessed once at the beginning or at the end of the study. Subjective SATO = self-reported speed-accuracy trade-off.

Figure S1

Estimates for the Hierarchical Model Across Items of the Cube Rotation Test

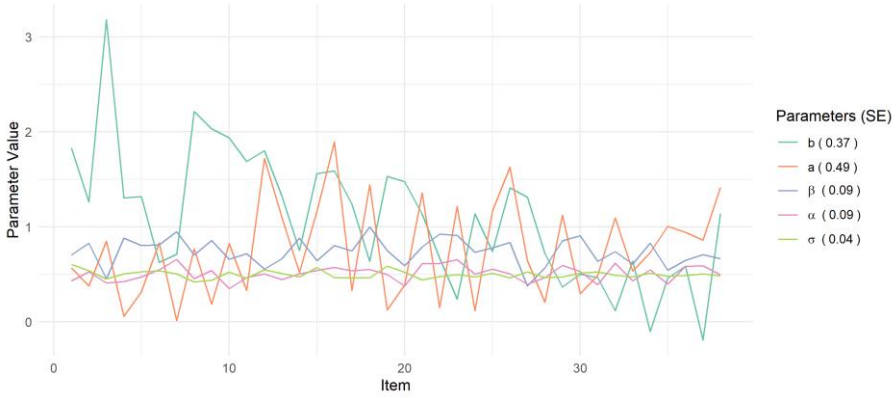


Figure S2

Estimates for the Hierarchical Model Across Items of the Matrix Distractor Test

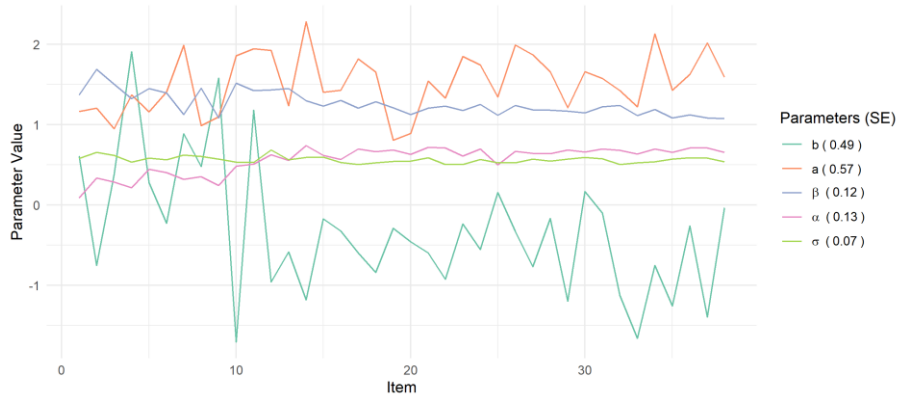


Figure S3

Estimates for the Hierarchical Model Across Items of the Matrix Construction Test

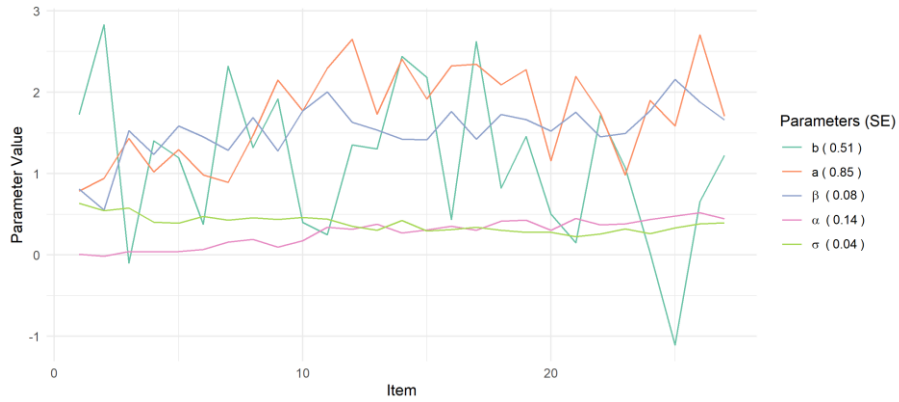


Figure S4

Estimates for the Hierarchical Model Across Items of the Knowledge Test

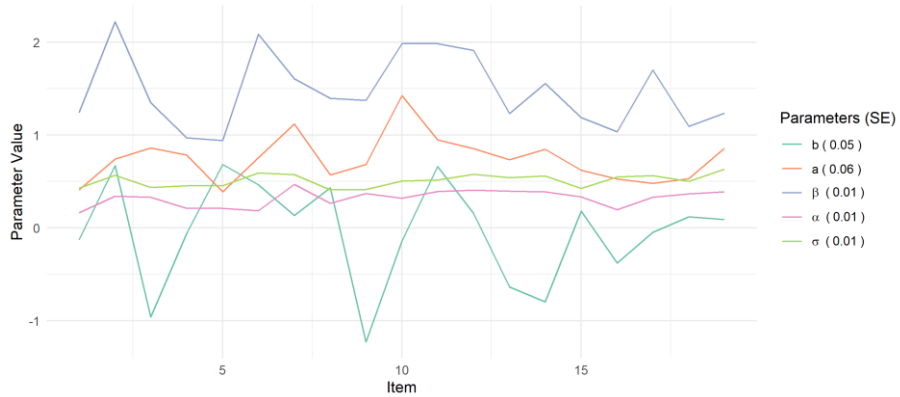


Figure S5

Estimates for the Race Model Across Items of the Cube Rotation Test

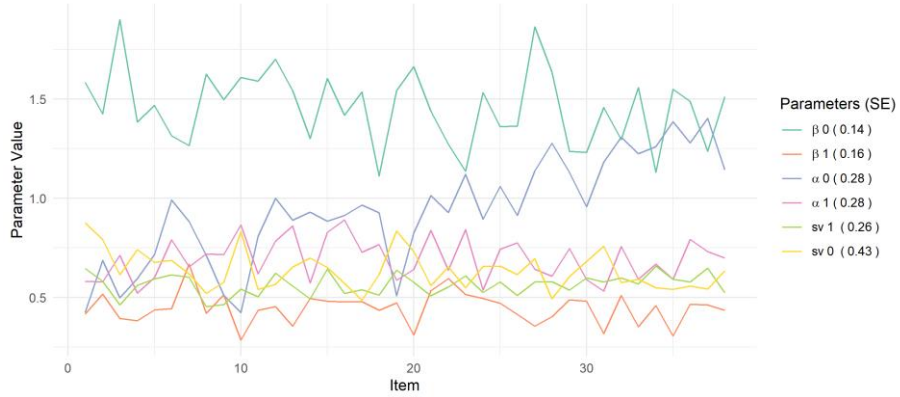


Figure S6

Estimates for the Race Model Across Items of the Matrix Distractor Test

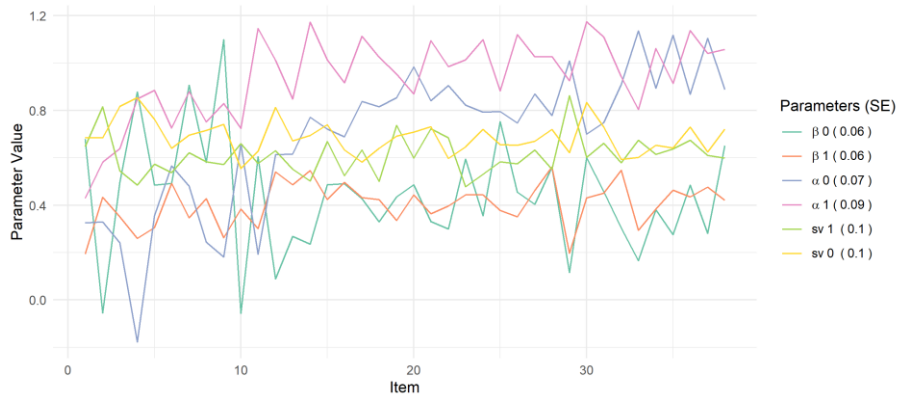


Figure S7

Estimates for the Race Model Across Items of the Matrix Construction Test

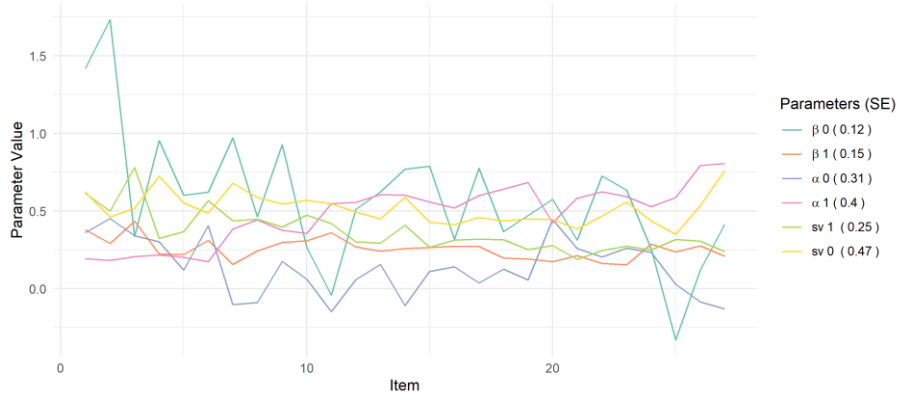


Figure S8

Estimates for the Race Model Across Items of the Knowledge Test

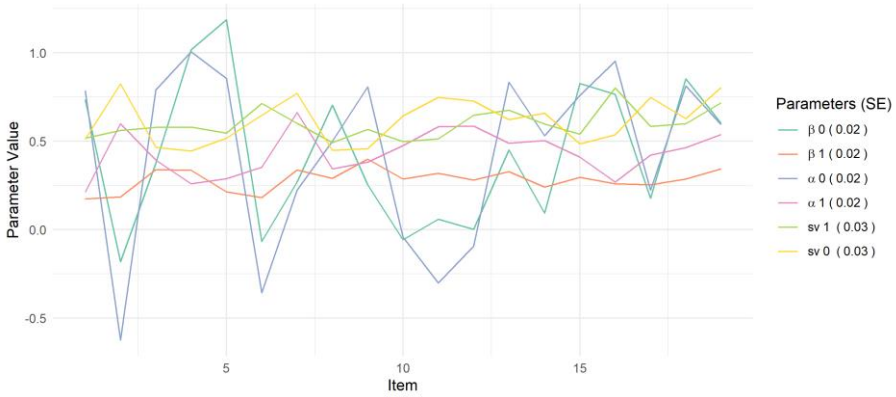


Figure S9

Estimates for the Diffusion Model Across Items of the Cube Rotation Test

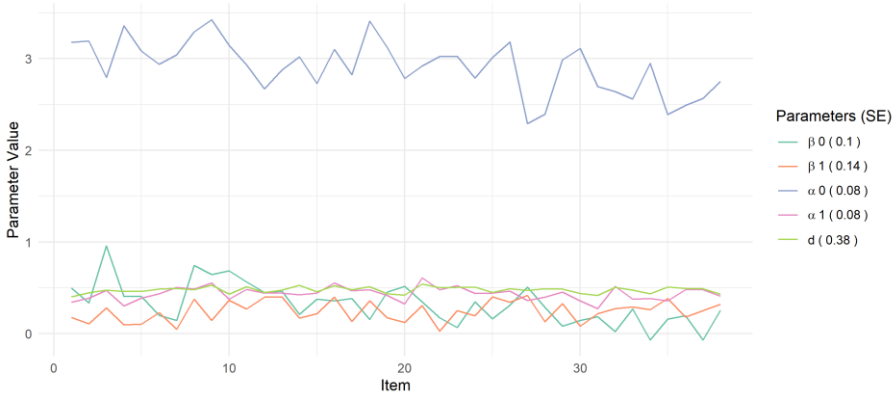


Figure S10

Parameter Estimates for the Diffusion Model Across Items of the Matrix Distractor Test

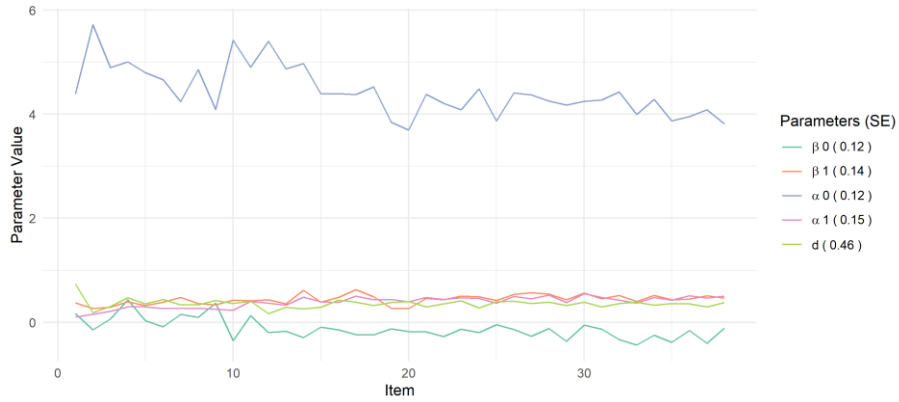


Figure S11

Parameter Estimates for the Diffusion Model Across Items of the Matrix Construction Test

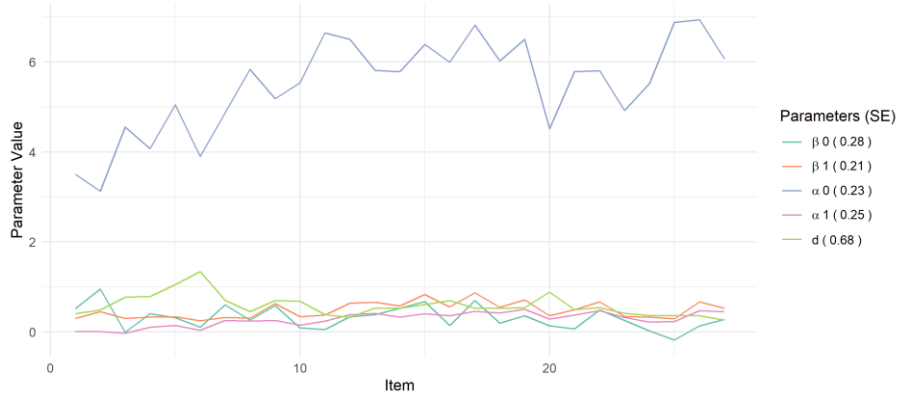


Figure S12

Parameter Estimates for the Diffusion Model Across Items of the Knowledge Test