




# Modelling the Effect of Instructional Support on Logarithmic-Transformed Response Time: An Exploratory Study

Luis Alberto Pinos Ullauri<sup>1,2,3</sup> , Wim Van Den Noortgate<sup>1,2</sup> , Dries Debeer<sup>1,2</sup> 

[1] imec research group ITEC, KU Leuven, Kortrijk, Belgium. [2] Faculty of Psychology and Educational Sciences, KU Leuven, Kortrijk, Belgium. [3] Centre for Digital Systems, IMT Nord Europe, Douai, France.

---

Methodology, 2024, Vol. 20(2), 100–120, <https://doi.org/10.5964/meth.12943>

Received: 2023-10-03 • Accepted: 2024-03-25 • Published (VoR): 2024-06-28

Handling Editor: Isabel Benitez, University of Granada, Granada, Spain

Corresponding Author: Wim Van Den Noortgate, KU Leuven, imec research group ITEC, Etienne Sabbelaan 51, Kortrijk, 8500, Belgium. E-mail: [wim.vandennoortgate@kuleuven.be](mailto:wim.vandennoortgate@kuleuven.be)

---

## Abstract

Instructional support can be implemented in learning environments to pseudo-modify the difficulty or time intensity of items presented to persons. This support can affect both the response accuracy of persons towards items as well as the time persons require to complete items. This study proposes a framework to model response time in learning environments as a function of instructional support. Moreover, it explores the effect of instructional support on response time in assembly tasks training using Virtual Reality. Three models are fitted with real-life data collected by a project that involves both industry and academic partners from Belgium. A Bayesian approach is followed to implement the models, where the Bayes factor is used to select the best fitting model.

## Keywords

lognormal models, response time, modelling, support, Bayesian estimation

The last decade saw a significant increase in the variety of digital learning environments (DLE). There has also been a growing pressure to incorporate personalised learning into DLEs. Personalised learning has varying definitions and applications (Bernacki et al., 2021). Nonetheless, the general idea focuses on providing tailored challenges according to the specific characteristics of learners. In addition, Järvelä (2006) argues that personalised learning affects multiple dimensions, amongst which there is the encouragement of learning. That is why it is important to consider the challenges' difficulties and learners' skills. If the challenges are too difficult for learners, it may discourage them and



negatively affect their learning process. Similarly, if the content is too easy, the learners may not feel challenged.

A possible strategy to implement personalised learning is to provide learners with challenges that match their current skill level (Debeer et al., 2021). This is feasible if there is a large inventory of possible challenges of varying difficulty. Such is the case for a DLE dedicated to training mathematics skills (Klinkenberg et al., 2011). Nonetheless, there may also be situations where only a limited number of challenges are available or where learners are required to repeat and practice specific challenges (e.g., assembly task training). In the latter, selecting challenges in a personalised manner may not be an option. A different strategy can be to vary the amount of assistance across learners and challenges, leveraging instructional support. Instructional support is related to scaffolding (Quintana et al., 2004). Scaffolding is generally defined by the process where the instructor helps learners in completing challenges that may be too difficult for them without any assistance. Therefore, the difficulty of a challenge can be either increased or decreased with lower or higher levels of support, respectively.

This idea of increasing and decreasing the level of instructional support is implicitly done in escape rooms. In escape rooms, there is a time limit to solve the puzzles and the game masters provide hints to help the players. Educational escape rooms have been studied as interactive tools to promote learning and soft skills (e.g., team-working) (Dietrich, 2018; López-Pernas et al., 2019). The additional hints can be interpreted as personalised scaffolds. The hints may allow slower players to successfully complete the escape room. Without the provided support, slower players may not be able to finish. This may lead them to experience frustration during the game. Support can be presented in a variety of ways such as promoting social interaction between learners (Gopez & Gopez, 2024; Klapp & Jönsson, 2021), providing hints with chatbots in DLEs or instructions in training DLEs (Vanneste et al., 2024).

## Research Questions

A modelling framework that considers the interplay between challenges, learners, and the provided support is essential for adaptability. Let us consider two learners. Learner one, who completed most (or all) challenges, is considered highly skilled. Learner two, who has not yet passed the challenges, is still learning. Support may affect them differently. It may have little to no effect for learner one who already mastered the challenges, whereas for learner two it may prove beneficial or even essential. Therefore, the need for a modelling framework comes from understanding how support affects learning outcomes in different situations. In doing so, we could properly design a personalised scaffolding system. This would allow the personalised adjustment of the level of instructional support to learners during the learning process.

It is informative to look at the response time when a learner is solving a challenge, particularly if the challenge is expected to be completed (e.g., assembly task training)

(Pinos Ullauri et al., 2021). Response time is the time necessary for the learner to complete that challenge. It is observable and depends on both learner and challenge in a particular time point (e.g., the learner could repeat the same challenge, thereby having many response times). Generally, the learners and challenges can be respectively described as persons and items, which are the terms used for the rest of the article. For instance, if a worker in a production line requires a certain amount of time to complete a step from an assembly task, the assembly step could be described as an item, the employee as the person, and the time as the response time. Similarly, if a player is within an escape room, the puzzles can be considered as items, the player as a person, and the time for the puzzles to be solved by the player as the response time.

One of the most popular modelling approaches to response time is the lognormal model proposed by Van der Linden (2006), which is inspired by Item Response Theory (IRT) (Lord, 1952; Rasch, 1960). IRT models focus generally on the estimation of latent parameters based on the correctness of answers. By defining person and item time characteristics, Van der Linden (2006)'s lognormal model can predict response times. However, this framework currently does not consider the effect of support. Support could affect both the person's probability of correctly answering an item as well as the time an item requires to be completed by persons. This problem motivated us to address the following research questions.

**RQ1:** How can the lognormal response time model be extended to account for the effect of instructional support?

**RQ2:** Which extensions of the lognormal model perform best in an empirical dataset?

The general aim of this study is to investigate the impact of support on response time. This would allow us to understand how to consider the effect of different levels of support during situations such as assessments, assembly-task training or educational escape rooms.

## Response Time

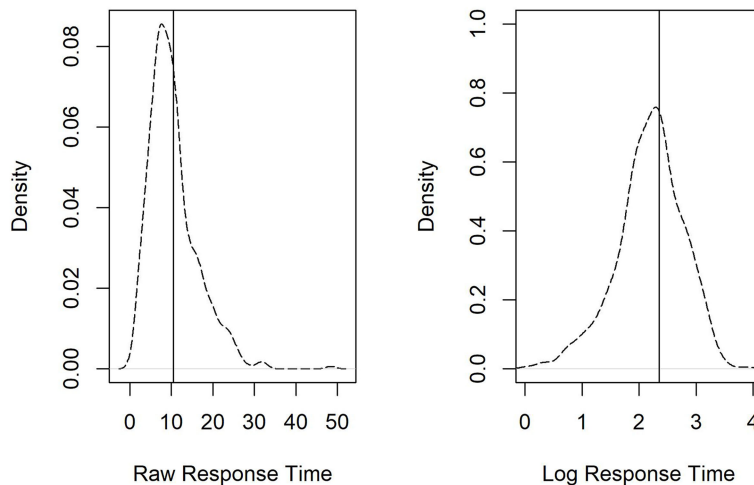
Response time has been studied within the field of psychological and educational research, where considerable modelling approaches have been proposed. The overview by De Boeck and Jeon (2019) describes four different, though possibly overlapping, modelling approaches. These approaches leverage response time to provide insight into the underlying factors involved in the item-person solving process. These underlying factors are described by latent variables related to person or item characteristics. For instance, a student in a test may be faster or slower than others. Likewise, some exercises may in general need more (or less) time to be completed than others. Similarly, there can be puzzles that can be solved in less time if the persons are faster than others.

Response time plays an important role in many applications. It can be used to design tests so that items are calibrated allowing their completion within the established time limits. It can be studied with the speed-accuracy trade-off, which describes situations where persons tend to make more mistakes when they increase their speed in solving items due to time constraints (Luce & Luce, 1986). Aberrant patterns can be analysed in response time (Marianti et al., 2014). It can serve to study speededness and its implications for high-stakes assessments, such as changing the item order within a test for security (Becker et al., 2021, 2022).

Response times cannot be negative and are typically right skewed (especially if the mean response time is close to zero and/or the variance is large). Therefore, response time distributions are often assumed to follow Gamma, exGaussian, Weibull or lognormal distributions. Figure 1 shows two plots. The left plot represents a fictitious, but realistic response time distribution where most of the density is accumulated on the left side. The right plot depicts its transformation through a natural logarithm. Van der Linden (2006) modelled real-life response time data with lognormal and normal distributions, where the lognormal models presented a better fit than normal models. Moreover, the logarithmic transformation completely changes the scale. For instance, if the unit of measure of raw response time is in seconds, then 1 second would become 0 in a logarithmic scale. The equivalent of 2.47 minutes in seconds would be 5, and the conversion of 2.25 hours to seconds would be equal to 9.

**Figure 1**

*Common Example of Raw and Log Response Time Distributions*



## Lognormal Model

Response time  $T_{ip}$  describes the time length a person  $p$  requires to complete an item  $i$ . It is important to remark that the completion of the item does not imply its accuracy (e.g., students can finish exams having wrong answers). In the same way as accuracy gives information on the persons' latent proficiency and the items' difficulties, response time can be used to estimate underlying characteristics from persons and items. In addition, there may be situations where accuracy is not as informative as response time. For instance, when solving puzzles in educational escape rooms or training in assembly tasks, everyone is expected to come to the correct answer. However, there can be much variation in the time needed to finish the items.

In line with [Klotzke and Fox \(2019\)](#) and [Van der Linden \(2006\)](#), let us assume the log response time  $\log T_{ip}$  can be approximated with a normal distribution with an expected value  $\mu_{ip}$  and variance  $\sigma_{error}^2$ , as shown in [Equation 1](#).  $\mu_{ip}$  corresponds to the expected log response time of person  $p$  solving item  $i$ . The variance  $\sigma_{error}^2$  describes the residual differences between the expected and actual values of the log response time.

In general,  $\mu_{ip}$  can be modelled as a function of person and item parameters. In [Equation 1](#),  $\lambda_i$ , which we will call time intensity, describes the time a particular item typically requires. It is important to note that time intensity should not necessarily be interpreted as item difficulty. An item could be both easily solved and time-consuming whereas another item may be difficult to solve without being time-demanding. For instance, an item with high difficulty and low time intensity could be hitting a bull's eye with a dart. Throwing a dart takes a limited amount of time, even for inexperienced players. Yet, hitting the bull's eye is difficult (highly skilled players would have a high probability of success). On the person side,  $\tau_p$  describes the speed of person  $p$ . Finally,  $\phi_i$ , which we call time discrimination, describes the degree of sensitivity of items towards variability of speed. Following [Equation 1](#), if the original response time unit is in seconds, the log time intensity is 2.5, the log time discrimination is equal to 1 and log speed is equal to 0.5, the expected value  $\mu_{ip}$  would result in 2, which corresponds to 7.3 seconds.

$$\log T_{ip} \sim \mathcal{N}(\mu_{ip}, \sigma_{error}^2); \mu_{ip} = \lambda_i - \phi_i \tau_p \quad (1)$$

This model is analogous with the 2 Parameter Logistic (PL) model from IRT. IRT models the probability of a correct response, through a logit transformation, to the difference between an underlying person ability and item difficulty considering a discrimination factor. The response time model in [Equation 1](#), like IRT models, suffers from identifiability issues as discussed in [Curtis \(2010\)](#). If all  $\phi_i$  are multiplied with a constant  $c$ , and all  $\tau_p$  are divided by that same constant, there would not be any changes on the expected response times. Moreover, if the time discrimination is not considered ( $\phi_i=1$  for all  $i$ ) the model still has an additive identification problem. An increase in the time intensity parameter  $\lambda_i$  can be compensated by a decrease in the speed parameter  $\tau_p$ . These issues

lead us to consider the analogous reparameterization of the 1 PL Model, by Van den Noortgate et al. (2003).

In this way, the previous equation can be expressed as Equation 2. In favour of simplicity, the time discrimination  $\phi_i$  is not considered for the remainder of the manuscript. Nonetheless, the time discrimination can be added, in principle, to the models we propose. In Equation 2, the intercept represents the expected log response time of an average person solving an average item. In addition, the latent variables  $\lambda_i$  and  $\tau_p$  represent the item time intensity and person speed deviations from the mean  $\mu_0$ . This means that if a person's speed or an item's time intensity are different from the average, the differences will be shown in  $\tau_p$  or  $\lambda_i$ . By constraining both the time intensity and the speed parameter to have a mean of zero, in Equation 2, the model is identifiable. Moreover, the work in Explanatory IRT (De Boeck & Wilson, 2004) has shown that models such as in Equation 2 allow flexibility regarding the addition of person or item covariates.

$$\log T_{ip} \sim \mathcal{N}(\mu_{ip}, \sigma_{error}^2); \mu_{ip} = \mu_0 + \lambda_i - \tau_p \quad (2)$$

where  $\lambda_i \sim \mathcal{N}(0, \sigma_\lambda)$ ,  $\tau_p \sim \mathcal{N}(0, \sigma_\tau)$ .

The next section describes an empirical application from a project on assembly training that involves both industry and academic partners from Belgium. Furthermore, three extensions of the lognormal model are presented in order to analyse the support effect's relation with the other latent variables. The results are interpreted and the limitations of this work are described further. The last section presents possible alternatives to extend this study and further explore the support effect on logarithmic response time.

## Method

### Application Context

The empirical application is based on the imec.icon project COSMO (COgnitive Support in Manufacturing Operations). One of the aims of the project COSMO was to design and evaluate personalised support for assembly task training in Virtual Reality (VR). For that purpose, students from technical-professional secondary education from two schools in the region of Flanders, Belgium, trained in VR. The students followed specific sequences of picking, connecting and assembling mechanic parts of an agricultural machine (e.g., engines, petrol tanks). They were asked to train on five different assembly tasks. During the training, they needed to complete the tasks and repeat them up to 3 times. They were asked about their past experience with VR. Before starting the experiments, the students were presented with the VR setup in a practice session. In the practice session the controls and actions were carefully explained (e.g., how to teleport, pick parts, connect and hold the pieces).

A random counter-balanced design was set in order to disentangle the effects of the order of the tasks, the repetitions and the levels of support. The order of the tasks was chosen randomly for all students. The number of repetitions was fixed for the tasks so that the first and third tasks were completed only once, the second and fourth task three times and the fifth task two times. The level of support was selected randomly for all students so that all students executed each task at least once. If the task was completed more than once, the support level would be maintained throughout that particular task. These conditions allowed us to mitigate possible confounding effects of the tasks, repetitions or levels of support.

The system provides support via several features as shown in Table 1. There are also three levels of support (Low, Medium and High). Figure 2 depicts six different cases. The left column corresponds to picking steps (e.g., steps where the students need to pick parts such as picking a wrench from the workbench). The right column shows connecting steps (e.g., putting screws in an engine). The three different levels of support and their features are also presented. Let us suppose a student (under High support) has already taken the necessary part and needs to connect it. Thus, the student would not only hear (Auditory information represented by the speaker icon on the lower-left corner of the High support condition) the recording explaining the step, but also see an arch that would appear (revealing the way the student needs to rotate and move the piece to connect it). In addition, a ghosting shape indicates the correct position of the part. In contrast, if another student (under Medium support) needs to pick parts, the support would only portray vertical lights showing the location of the necessary pieces. Moreover, if a student (under Low support) requires to pick something, only the ghosting shape appears. However, there is no indication of the location of the necessary part in the virtual workshop, which comprises several workbenches.

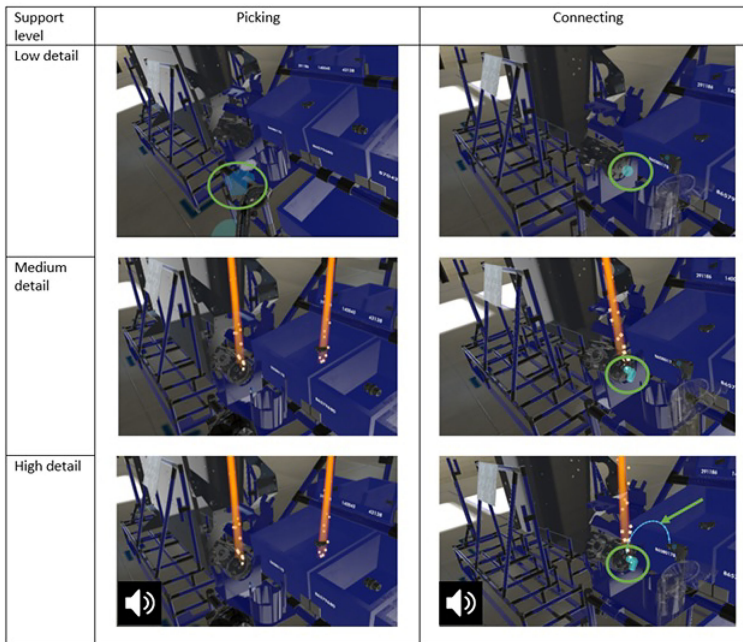
**Table 1**

*Support Features*

Feature	Description	Support Level
Haptic feedback	Vibration in the controllers when the users pick an incorrect part or piece	L, M, H
Visual aids	Vertical lights that pinpoint the location of the parts and shades that reveal the appropriate position of the parts	M, H
Auditory information	Recordings in Dutch regarding the instructions to follow during a step	H

*Note.* L stands for Low, M for Medium and H for High levels of support



**Figure 2***Levels of Support During VR Assembly Task Training*

## Dataset

The data was collected from 86 technical secondary education students using VR-supported technology to train for five different assembly tasks. Each task includes 6 to 12 fixed steps in a fixed sequence. There is a total of 45 steps across all tasks. The number of attempts or repetitions the students train at a certain step range from 1 to 3. The previous experience in VR is measured with a four-point Likert-type scale from 1 to 4. The response time of the step is measured in seconds. For each step, one of three levels of instructional support were given: Low (L), Medium (M) and High (H). A non-support level was not considered, which means that during the training there was always support. The dataset includes a total of 6195 observations at the step-level. For each observation, data were registered regarding the step, the student, the repetition and the support level.

Step response times with duration smaller than 2 seconds were taken out of the dataset. These extreme values indeed do not refer to real response times of the students. When a student was stuck in a step of the assembly sequence, the supervisors activated



the auto-complete feature of the VR setting. The auto-complete feature then rapidly (typically in less than 2 seconds) showed how the step needed to be done. An example of how the data is structured is shown in Table 2.

**Table 2**

*Dataset Structure*

Student ID	Step	Support	Attempt	Prev Exp VR	Response time
1	1	H	1	2	64
1	2	H	1	2	16
1	3	H	1	2	24.2
..	..	..	..	..	..
1	1	M	2	2	91.3
1	2	M	2	2	19.7
..	..	..	..	..	..
86	45	L	3	1	32.4

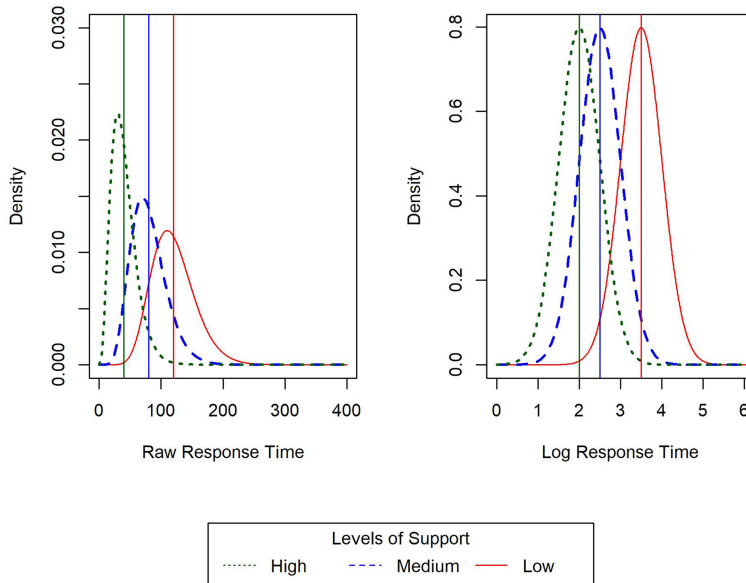
## Modelling Framework

Different approaches can be considered to explore the support effect on response time. First, one could define a separate item characteristic parameter  $\lambda_{il}$  for each item-support level combination, with the index  $l$  referring to the support level. This would however lead to many parameters in the model making it inefficient and difficult to interpret. In addition, it assumes the impact of support is the same for all persons. A second approach is to consider a separate speed parameter for each person-support level pair  $\tau_{pl}$ . In the same way as the previous approach, it would increase the number of parameters, while also assuming the same impact for all items. A third approach could be to include a support-level predictor in the model. The support can be interpreted as not belonging to either item or person characteristics, but rather to the system side (e.g., the DLE). The general intuition behind the support effect is that it reduces the expected time required to solve items. This means that the expected log response time can be steered to the left or right depending on the magnitude and relation of the item, person and support parameters.

Figure 3 shows how the support effect can potentially influence in both raw and log response time distributions. The vertical lines show the location of the means. In the left, the density of the distribution visually changes, whereas on the right plot there is a displacement of the mean. In both plots, the blue dashed distribution could be considered using a medium level of support. If there is a change to a higher level of support, the distribution would become the dotted green one. Similarly, if the level of support would decrease, the overall expected log response time distribution would be represented by the red one.

**Figure 3**

Potential Effect of Support Level on the Raw and Log Response Time



### Model 1

In order to introduce the support effect in the lognormal model, we include a dummy variable as a predictor, for all but one level of support. Their coefficients thus express what the overall effect is on the response time of this level of support, compared to the reference level of support. This means that if there are  $L$  levels of support, then there would be  $L - 1$  parameters. Let us formally define  $\alpha_l$  as the instructional support effect of a level  $l$  compared to the reference. This means the expected log response time  $\mu_{ipl}$  would also depend from the support aside from the person and item parameters. Model 1 is shown in Equation 3, which describes the log response time of a particular item  $i$ , person  $p$  and level of support  $l$ .

$$\log T_{ipl} \sim \mathcal{N}(\mu_{ipl}, \sigma_{error}^2); \mu_{ipl} = \mu_0 + \lambda_i - \tau_p + \alpha_l \quad (3)$$

where  $\lambda_i \sim \mathcal{N}(0, \sigma_\lambda)$ ,  $\tau_p \sim \mathcal{N}(0, \sigma_\tau)$ .  $\alpha_l$  refers to the effect that support level  $l = 0, 1, 2, \dots, L - 1$  has on the expected log response time, compared to the reference support level.

Model 1 assumes the impact of support to be additive (on the logarithmic scale) and independent of either persons or items. Equation 3 describes the generalised linear mixed

model that explains the expected value  $\mu_{ipl}$ . Given its low complexity, this model can serve as a benchmark to compare with the other models. The parameters  $\lambda_i$  and  $\tau_p$  are assumed to be normally distributed, respectively, with means constrained to zero and standard deviations  $\sigma_\lambda$  and  $\sigma_\tau$ . If the reference is the highest level of support, the support effect  $\alpha_i$  is expected to be positive. That is, to increase the expected log response time and steer the logarithmic distribution to the right. Moreover, the support effect is considered as a fixed effect. For each level, except the reference level, the effect is estimated.

Even though the support level effect is constant on the logarithmic scale, its effect on the response time scale is multiplicative. For instance, let us imagine a support level coefficient equal to -1. If the log-response time expected value decreases from 3 to 2, the change on response time scales would result in 12.69 seconds. However, that same difference in a log-response time scale from 4 to 3 would translate to 34.51 seconds.

Furthermore, this model as well as the next ones can be extended to include additional predictors (e.g., a repetition predictor describing the amount of times a person  $p$  has attempted to solve an item  $i$ ).

## Model 2

Instead of assuming a given level of support increases or decreases the log response time by a constant, it may also be realistic that the expected log response time is multiplied by a constant given a certain level of support. Therefore, this model considers a multiplicative approach, shown in Equation 4. In this case, the support effect  $\alpha_i$  multiplies the overall linear combination of the item and person characteristics. This means the support effect acts as either an increasing or decreasing factor on the overall expected  $\mu_{ipl}$ , depending on whether  $\alpha_i$  is smaller or larger than 1. This in turn can steer the response time distribution to either side. Moreover, the impact of the support becomes larger when the expected log time for an average person solving an average time  $\mu_0$  increases.

$$\log T_{ipl} \sim \mathcal{N}(\mu_{ipl}, \sigma_{error}^2); \mu_{ipl} = (\mu_0 + \lambda_i - \tau_p)\alpha_i \quad (4)$$

where  $\lambda_i \sim \mathcal{N}(0, \sigma_\lambda)$ ,  $\tau_p \sim \mathcal{N}(0, \sigma_\tau)$ .

## Model 3

The low complexity of the previous models may be considered an advantage in the sense that they are easier to estimate. Nonetheless, they also assume the support effect to be independent of items and persons. The main rationale behind this model is that the support effect may depend on the item-person combination. If the match is adequate (small difference between item time intensity and person speed), support may have a higher effect. For instance, persons who are expected to complete very rapidly a low time-intensive or average item, may find support distracting, compared to slower

persons who may find it helpful. Similarly, if an item is highly time-intensive, and hence, requires a lot of time to be completed, an average person may not benefit as much from support compared to a fast person, which may indeed be able to finish.

A possible implementation of these notions is presented in Equation 5, which describes a similar linear relation of the latent variables presented in Model 1 with the inclusion of a term that depends on the difference  $\lambda_i - \tau_p$ . Several functional forms (such as  $e^{-(\lambda_i - \tau_p)^2}$  and  $\frac{1}{1 + |\lambda_i - \tau_p|}$ ) were also tested to portray the intended intuition. However, the reciprocal reciprocal quadratic item-person difference produced a better fit to the dataset, compared to the previous functional forms.

$$\log T_{ipl} \sim \mathcal{N}(\mu_{ipl}, \sigma_{error}^2); \mu_{ipl} = \mu_0 + \lambda_i - \tau_p + \alpha_i \frac{1}{1 + (\lambda_i - \tau_p)^2} \quad (5)$$

where  $\lambda_i \sim \mathcal{N}(0, \sigma_\lambda)$ ,  $\tau_p \sim \mathcal{N}(0, \sigma_\tau)$ .

## Model Extension

The previous models are applied to the dataset from the assembly task training study. The models are fitted at the step level considering context-specific characteristics. For instance, students may have trained on a particular assembly task multiple times, therefore an effect related to learning can be defined. This effect could be considered a person characteristic as the learning rate from a student's mastery of an assembly step. Moreover, an effect from the students' previous experience with Virtual Reality is also included in the models.

To account for the multiple attempts, the models are extended including an index  $t$  for the time a student  $p$  may train on a step  $i$  with a support level  $l$ . Furthermore, the following coefficients and variables are considered for the models:

- The variable  $attempt_{ipt}$ , which ranges from 0 to 2, corresponds to the number of repetitions the student  $p$  trained on an assembly step  $i$  on a moment  $t$ . On that account, for the first attempt,  $attempt_{ipt}$  is equal to 0 and for the third attempt has a value of 2.
- The  $a$  coefficient describes the learning effect on log response time. In general, we would expect response time to decrease as the students train on the steps multiple times.
- The variable  $prevexp_p$ , which ranges between 0 and 3, corresponds to the self-assessment of the student  $p$  regarding its previous experience in VR, where 0 relates to having no experience at all and 3 to be fully experienced in VR.
- The  $e$  coefficient which relates to the previous experience in VR effect on log response time. Similarly to  $a$ , we would expect  $e$  to have a negative effect on log response time. In other words, student with more experienced background in VR would be expected to have a lower response time.

Equation 6 describes the extension of Model 1 (Equation 3) with the addition of the learning and previous experience in VR effects.

$$\mu_{iplt} = \mu_0 + \lambda_i - \tau_p + \alpha_l + a(\text{attempt}_{ipt}) + e(\text{prevp}_p) \quad (6)$$

Models 2 and 3 are extended in a similar way as with Model 1.

In addition, by including the coefficients  $a$  and  $e$  the interpretation of  $\mu_0$  changes. In this case,  $\mu_0$  describes the expected log response time for an average student ( $\tau_p = 0$ ), without previous experience in VR ( $\text{prevp}_p = 0$ ), taking an average step ( $\lambda_i = 0$ ) for the first time ( $\text{attempt}_{ipt} = 0$ ) at the reference support level.

Furthermore, the reference support level is the High Level. This means that the support effects  $\alpha_1$  and  $\alpha_2$  represent its change to Medium and Low levels respectively. In the case of the first and third models,  $\alpha_1$  and  $\alpha_2$  are therefore expected to be greater than zero (High level constraint), increasing the response time and moving the mean to the right. On the other hand, for the second model,  $\alpha_1$  and  $\alpha_2$  are expected to be greater than 1 (High level constraint), because for this model, a larger  $\alpha$  corresponds to a higher response time.

## Bayesian Modelling

A Bayesian approach was considered over frequentists statistics given the flexibility the former allows when defining models. It also has the advantage that not only point estimates are obtained, but rather whole posterior probability distributions. There are various software packages that employ Markov Chain Monte Carlo (MCMC) to estimate the posteriors. The software Stan by [Carpenter et al. \(2017\)](#) and package RStan by [Stan Development Team \(2020\)](#) were selected for the analysis (see [Pinos Ullauri, 2024](#) for the R and Stan codes for the proposed models). These employ Hamiltonian Monte Carlo, which in general provides posterior distributions more efficiently compared to the traditional MCMC algorithm ([Radivojević & Akhmatkaya, 2020](#)).

The models are estimated providing posterior distributions of the parameters that explain the log-response time of the steps. Depending on the number of parameters, the models may require more iterations to converge. In this case, the number of iterations chosen are 10000 with 2000 burn-in samples with 4 different chains, having a total of 36,000 post burn-in samples. The chains are randomly initialised with Stan's default initialisation with a fixed seed of 1, in order to allow reproducible results.

Bayesian estimation can profit from the use of previous information through priors. Given that this experiment was performed for the first time, we could not provide informative priors on the latent variables. Nonetheless, the models can still benefit from weakly informative priors. Weakly informative priors assist in stabilising the chains by guiding the algorithm to accumulate probability mass in reasonable regions. [Table 3](#) shows the weakly informative priors for the model parameters. It can be seen that

a sufficiently large standard deviation of 100 (especially for a logarithmic scale) is set for the effects. The standard deviations  $\sigma_\lambda$  and  $\sigma_\tau$  are estimated through the flat prior  $\mathcal{U}(0, 100)$  (since time and standard deviations are positive by definition) whereas the other effects with  $\mathcal{N}(0, 100)$ . Using similar priors for the standard deviations can have similar effects on the results as using constraints (Gelman, 2009).

**Table 3**

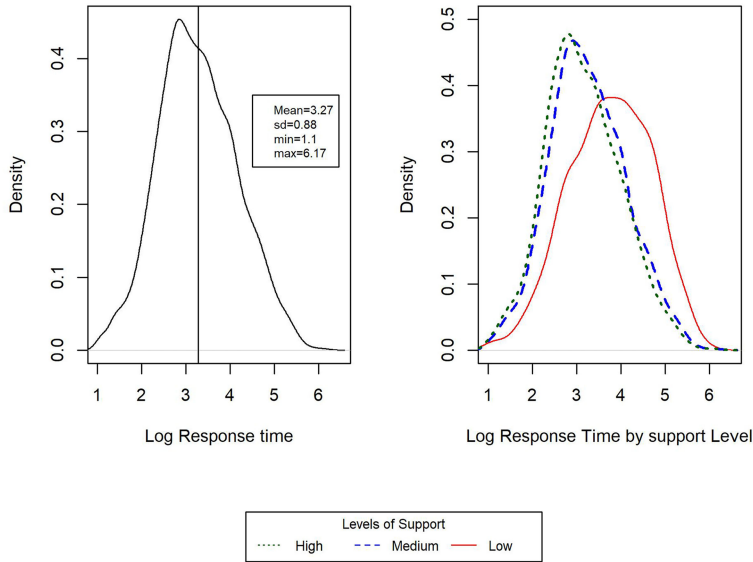
*Weakly Informative Priors for the Latent Parameters*

Parameters	Distributions
Fixed intercept $\mu_0$	$\mathcal{U}(0, 100)$
Time intensity $\lambda_i$	$\mathcal{N}(0, \sigma_\lambda)$
SD Time intensity $\sigma_\lambda$	$\mathcal{U}(0, 100)$
Speed $\tau_p$	$\mathcal{N}(0, \sigma_\tau)$
SD Speed $\sigma_\tau$	$\mathcal{U}(0, 100)$
Support (Low) $\alpha_1$	$\mathcal{N}(0, 100)$
Support (Medium) $\alpha_2$	$\mathcal{N}(0, 100)$
Attempt $a$	$\mathcal{N}(0, 100)$
Previous experience $e$	$\mathcal{N}(0, 100)$

The Bayes factor (BF) is a statistical metric that estimates which model fits the data better compared to another (Nicenboim & Vasisht, 2016). Bridge Sampling can be used to estimate it (Bennett, 1976). This algorithm iteratively estimates the BF by using random samples of its posterior distribution. Although, it requires a sufficient amount of samples from the parameter estimation to converge (Gronau et al., 2020). The convention for log BF describes that if Model X has a log BF larger than 2 compared to Model Y, then Model X is preferred (Kass & Raftery, 1995).

## Results

Figure 4 shows the kernel density plots of the log response time in seconds of all steps (left plot) and separated by support level (right plot). The bandwidth was estimated through the rule of thumb by Silverman (1986) due to its robustness and the lognormal nature of the data. Moreover, a brief descriptive summary of the step log response time is shown in the legend of the left plot. In addition, the Low support distribution can be easily distinguished from the other support levels, whereas the same cannot be said for the Medium and High support, whose densities are similar. This may suggest

**Figure 4***Log Response Time Kernel Density Plots*

the difference between the effects of Medium and High levels of support was smaller compared to the differences with the Low level of support.

Table 4 shows the log BF<sub>s</sub> of the posterior distributions. The model in the rows are compared with those in columns. For instance, Model 1 has a log BF (32.83, higher than 2) against Model 2, but a log BF towards Model 3 (-27.16, lower than 2). This summarises Model 3 (reciprocal quadratic difference between time intensity and speed) as the best fit, and Model 2 the worst. This means that for the data of this experiment, the impact of support on log response time is higher when the speed of students is closer to the time intensity of the steps.

**Table 4***Models Log Bayesian Factor*

Log Bayes Factor	Model 1	Model 2	Model 3
Model 1	-	32.83	-27.16
Model 2	-	-	-59.99
Model 3	-	-	-

*Note.* Only the elements above the upper diagonal are shown. The BF<sub>s</sub> are read from row to column



Given that Model 3 is the best fitting model, its results are discussed. However, the parameter estimates for the other models are quite similar, as shown in Table 5. The empirical estimator of geometric ergodicity  $\hat{R}$  test suggests convergence if  $\hat{R} < 1.05$  (Vehtari et al., 2021). In this case  $\hat{R}$  resulted in 1 for all the models. For Model 3, the expected log response time for an average person and an average item at the first attempt,  $\mu_0$ , is equal to 3.15, which corresponds approximately to 23.58 seconds. The residual variance is 0.54, which can be quite considerable when transformed back into response time. The standard deviation of speed is 0.18, whereas the deviation of time intensity is 0.73. The deviation of time intensity in Model 3 shows greater variability amongst assembly steps compared to the other models (0.63 and 0.61). This greater variability could also compensate the difference in the expected log response time  $\mu_0$ , which is lower compared to Models 1 and 2.

**Table 5**

*Mean Estimate Results*

Parameter	Expected RT	SD I.TI	SD P.Speed	M Supp.	L Supp.	Attempt	Prev Exp	Error Res.
Means	$\mu_0$	$\sigma_\lambda$	$\sigma_\tau$	$\alpha_1$	$\alpha_2$	<b>a</b>	<b>e</b>	$\sigma_{\text{error}}$
Model 1	3.47	0.63	0.18	0.09	0.68	-0.28	-0.06	0.55
Model 2	3.48	0.61	0.17	1.02	1.21	-0.27	-0.05	0.55
Model 3	3.15	0.73	0.18	0.12	0.9	-0.28	-0.05	0.54

The support effects show that the distance between the High and Medium level of support is not as wide as with the distance between the Medium and Low level further supporting the visual evidence in Figure 4. For Models 1 and 3, the effect of Medium level (compared to the High level;  $\alpha_1$ ) is close to zero, whereas the effect of the Low level is considerably larger. For Model 2, the effect of the Medium level is close to one (High level reference), whereas the effect of the Low level is much larger. The attempt has considerable negative effect on response time suggesting learning. If a student repeated once (second attempt) a particular step, the effect would be a displacement of around 0.28 in log-scales, whereas if the student repeated twice (third attempt) the change would result in  $0.28 \times 2 = 0.56$ . The previous experience has a negative effect of 0.05 providing a closely negligible impact on log response time.

## Discussion and Conclusion

We have addressed our research questions by proposing a framework to extend the lognormal response time model considering the effect of instructional support. Moreover, the models are very flexible and can be further extended to include context-specific

characteristics. Model 3 proved to be the best fit to our dataset, which suggests that support can indeed have a larger effect when the difference between person speed and time intensity is lowest. In other words, when the match between person and item is more adequate, the support may have a stronger effect. In cases where an item is very time-consuming for a person, completing this item will take its time, regardless of the level of support that was provided. Similarly, if an item is less time-consuming for a person, the item will be done faster, and there would not be a large effect of support.

The proposed models have certain limitations to consider. First, the models assume a constant effect (same for all items and persons) after each attempt, which may not necessarily be realistic. Nevertheless, the addition of more parameters (e.g., effects for each attempt) can, in principle, increase the computational expense needed to fit the models. Second, the support is modelled on the expected log response time, which may be difficult to interpret. Nonetheless, it also allows a flexible effect on different cases. For instance, let us suppose a person is expected to use 10 seconds for an item. The potential profit of support can never be larger than 10 seconds, whereas if the expected time is 10 minutes, the effect of support can be much larger than 10 seconds. Third, the choice of time unit (e.g., seconds, minutes) for the response time scale may indeed affect the applicability of Model 2 (log multiplicative model). For instance, if the time intensity of an item is 50 seconds, it could also be expressed as 0.83 minutes. When log-transforming both times, 50 seconds would become 1.69, and 0.83 would result in -0.08. In that case, a support effect lower than 1 would decrease the former RT, and increase the latter RT (making it less negative). Fourth, the data of our study visually suggests a lognormal distribution to be a good approximation. However, it may not be the case for other studies, and we would recommend to check the response time density in preliminary analyses. Finally, we proposed three models, but other models are possible. The flexibility of our approach allows using a model that fits the type of data and interest of the researcher. However, a drawback is that the model has to be specified. If not, it could lead to flawed conclusions. Therefore, we propose to use multiple plausible models and compare the model fit.

The theoretical background behind scaffolding suggests that it can be valuable to give support. In doing so, the difficulty of the challenge can better match the learner's ability (Quintana et al., 2004). In this case, the support design allows the item-person response time to vary. This can be naively interpreted as either boosting the person speed or decreasing the time intensity, thereby satisfying scaffolding and promoting learning. The results showed that the Medium and High support effects were closer compared to the Low level of support. This is something important to take into account. Some students may have experienced that the High and Medium levels made the tasks quickly solvable, whereas the Low support tasks still may have been too time-consuming, prolonging their training time. Further initial validation stages of the levels of support would be recommended for new studies. This would allow a more even spread of the support level

effects. This is essential if the aim is to provide adaptive level choosing, considering the wide variation of the learners' speeds.

This methodology can be further combined with the work by Pelánek (2016), who proposes the use of the Elo-Rating system in DLEs. The Elo rating system has been proposed to update after each response the person's ability and/or the item's difficulty level. In this way, a changing ability can be tracked, and this can constitute the base for an adaptive item selection. This algorithm can be adapted to include response times and latent characteristics to track the growth of the learners speeds as they progress and learn (Pinos Ullauri et al., 2021). Moreover, the Elo-Rating system can be extended to include the effect of support, so that items with appropriate time intensities and adequate support can be provided for the persons' current speeds. Furthermore, this framework is not only limited to modelling response time. It could be adapted and extended to model response accuracy or polytomous responses.

## Future Work

Several extensions can be considered regarding the addition of the support effect into the response time models. First, the effect of the level of support on the speed can vary across persons. To that end, separate random person effects can be defined for each level of support, effects that can be correlated. In addition, interaction terms can be included to explore whether the variation in the effects of support can be explained by person background characteristics. Similarly, the effect of support can be allowed to vary randomly across items and/or according to item characteristics. Finally, the models by Klein Entink et al. (2009) and Van der Linden (2007) can be extended to explain the support effect on both response accuracy and response time considering correlation between them.

---

**Funding:** This work was funded by imec (research centre for nanoelectronics and digital technology and digital technology) and VLAIO (Flanders Innovation & Entrepreneurship agency) under the imec.icon project COSMO (COgnitive Support in Manufacturing Operations).

---

**Acknowledgments:** We would like to thank Dr. Benítez and the two reviewers for their comments, which helped improve the quality of the paper.

---

**Competing Interests:** The authors have declared that no competing interests exist.

---

## References

Becker, B., Debeer, D., Weirich, S., & Goldhammer, F. (2021). On the speed sensitivity parameter in the lognormal model for response times and implications for high-stakes measurement

- practice. *Applied Psychological Measurement*, 45(6), 407–422.  
<https://doi.org/10.1177/01466216211008530>
- Becker, B., Rijn, P. V., Molenaar, D., & Debeer, D. (2022). Item order and speededness: Implications for test fairness in higher educational high-stakes testing. *Assessment & Evaluation in Higher Education*, 47(7), 1030–1042. <https://doi.org/10.1080/02602938.2021.1991273>
- Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2), 245–268. [https://doi.org/10.1016/0021-9991\(76\)90078-4](https://doi.org/10.1016/0021-9991(76)90078-4)
- Bernacki, M., Greene, M., & Lobczowski, N. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educational Psychology Review*, 33, 1675–1715. <https://doi.org/10.1007/s10648-021-09615-8>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Curtis, S. M. (2010). Bugs code for item response theory. *Journal of Statistical Software, Code Snippets*, 36(1), 1–34. <https://doi.org/10.18637/jss.v036.c01>
- Debeer, D., Vanbecelaere, S., Van den Noortgate, W., Reynvoet, B., & Depaep, F. (2021). The effect of adaptivity in digital learning technologies: Modelling learning efficiency using data from an educational game. *British Journal of Educational Technology*, 52, 1881–1897.  
<https://doi.org/10.1111/bjet.13103>
- De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10, Article 102. <https://doi.org/10.3389/fpsyg.2019.00102>
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer. <https://doi.org/10.1007/978-1-4757-3990-9>
- Dietrich, N. (2018). Escape classroom: The Leblanc process—An educational “escape game”. *Journal of Chemical Education*, 95, 996–999. <https://doi.org/10.1021/acs.jchemed.7b00690>
- Gelman, A. (2009). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- Gopez, J. M., & Gopez, B. (2024). Instructor scaffolding for interaction and online student engagement among a sample of college students in the Philippines: The mediating role of self-regulation. *European Journal of Psychology of Education*, 39, 1069–1091.  
<https://doi.org/10.1007/s10212-023-00728-y>
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2020). Bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software, Articles*, 92(10), 1–29.  
<https://doi.org/10.18637/jss.v092.i10>
- Järvelä, S. (2006). Personalised learning? New insights into fostering learning capacity. In *Personalising education* (pp. 31–46). OECD Publishing.  
<https://doi.org/10.1787/9789264036604-en>
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.2307/2291091>

- Klapp, A., & Jönsson, A. (2021). Scaffolding or simplifying: Students' perception of support in Swedish compulsory school. *European Journal of Psychology of Education, 36*, 1055–1074. <https://doi.org/10.1007/s10212-020-00513-1>
- Klein Entink, R., Fox, J.-P., & Linden, W. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*, 21–48. <https://doi.org/10.1007/s11336-008-9075-y>
- Klinkenberg, S., Straatemeier, M., & Van der Maas, H. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*(2), 1813–1824. <https://doi.org/10.1016/j.compedu.2011.02.003>
- Klotzke, K., & Fox, J.-P. (2019). Modeling dependence structures for response times in a Bayesian framework. *Psychometrika, 84*, 649–672. <https://doi.org/10.1007/s11336-019-09671-8>
- López-Pernas, S., Gordillo, A., Barra, E., & Quemada, J. (2019). Analyzing learning effectiveness and students' perceptions of an educational escape room in a programming course in higher education. *IEEE Access, 7*, 184221–184234. <https://doi.org/10.1109/ACCESS.2019.2960312>
- Lord, F. M. (1952). *A theory of test scores*. Psychometric Society.
- Luce, R., & Luce, V. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.
- Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451. <https://doi.org/10.3102/1076998614559412>
- Nicenboim, B., & Vasisht, S. (2016). Statistical methods for linguistic research: Foundational ideas –Part II. *Language and Linguistics Compass, 10*(11), 591–613. <https://doi.org/10.1111/lnc3.12207>
- Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education, 98*, 169–179. <https://doi.org/10.1016/j.compedu.2016.03.017>
- Pinos Ullauri, L. A. (2024). *Modelling the effect of instructional support on logarithmic-transformed response time: An exploratory study* [OSF project page containing R and Stan codes for the proposed models]. OSF. [https://osf.io/ktuhc/?view\\_only=0d3dad965ee1428080b38e2f84191b7d](https://osf.io/ktuhc/?view_only=0d3dad965ee1428080b38e2f84191b7d)
- Pinos Ullauri, L. A., Van den Noortgate, W., & Debeer, D. (2021, July.). *Modelling response time and impact of instructional level of support* (pp. 65–72). Rencontres des Jeunes Chercheurs en Intelligence Artificielle (RJCIA'21) Plate-Forme Intelligence Artificielle (PFIA'21), Bordeaux, France. <https://hal.archives-ouvertes.fr/hal-03298738>
- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337–386. [https://doi.org/10.1207/s15327809jls1303\\_4](https://doi.org/10.1207/s15327809jls1303_4)
- Radivojević, T., & Akhmatkaya, E. (2020). Modified Hamiltonian Monte Carlo for Bayesian inference. *Statistics and Computing, 30*, 377–404. <https://doi.org/10.1007/s11222-019-09885-x>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall. <https://doi.org/10.1007/978-1-4899-3324-9>

- Stan Development Team. (2020). *RStan: The R interface to Stan* [R package version 2.21.2]. <http://mc-stan.org/>
- Vanneste, P., Dekeyser, K., Pinos Ullauri, L. A., Debeer, D., Cornillie, F., Depaepe, F., Raes, A., Van den Noortgate, W., & Said-Metwaly, S. (2024). Towards tailored cognitive support in augmented reality assembly work instructions. *Journal of Computer Assisted Learning*, 40(2), 797–811. <https://doi.org/10.1111/jcal.12916>
- Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28, 369–386. <https://doi.org/10.3102/10769986028004369>
- Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204. <https://doi.org/10.3102/10769986031002181>
- Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved  $\hat{R}$  for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2), 667–718. <https://doi.org/10.1214/20-BA1221>



*Methodology* is the official journal of the European Association of Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.