Modelling response time and impact of instructional level of support

L Pinos Ullauri, W van den Noortgate, D Debeer

Modelling response time and impact of instructional level of support

L. Pinos Ullauri[1,2], W. Van den Noortgate[1],D. Debeer[1]

[1] University of Leuven, ITEC an imec research group at KU Leuven
[2] IMT Lille-Douai, CERI-SN

luisalberto.pinosullauri@kuleuven.be

Abstract

This study investigates several approaches to modelling the response time a person, which has been given a certain level of support, requires to solve an item. Three different generic models are proposed explaining the involved latent variables and their interactions. The goal of this paper is to present various ways to model the instructional level of support and allow the reader to choose and extend the most suitable model in a particular dataset of interest. For illustrative purposes, the models are implemented within a Bayesian Framework for a specific situation.

Keywords

response time, modelling, level of support, bayesian framework

Résumé

Cette étude examine plusieures approches pour modéliser le temps de réponse dont une personne, qui a reçu un certain niveau de soutien, a besoin pour résoudre un item. Trois modèles génériques différents sont proposés expliquant les variables latentes impliquées et leurs interactions. Le but de cet article est de présenter différentes manières de modéliser le niveau de soutien et de permettre au lecteur de choisir et d'étendre celui qui convient le mieux dans un jeu de données d'intérêt particulier. À des fins illustratives, les modèles sont implémentés dans un cadre Bayésien pour une situation spécifique.

Mots-clés

Temps de réponse, modélisation, niveau de soutien, cadre bayésien

1 Introduction

When a student is working on a test exercise the time it takes for the student to solve that exercise is a response time. This response time is measurable and relates that particular student with that specific exercise. Generally, the terms person and item can be used to describe the student and exercise, respectively. These abstract representations can be applied in various situations. For instance, if an employee working in an assembly line requires a certain amount of time to assemble a part, this part could be described as an item and the employee as the person, while the required time a response time. Similarly, if a student is playing a virtual or real educational escape room, the puzzles within the room can be categorised as items, the students persons, and the time for the puzzles to be solved by the students response time.

Response time is observable and relates a person p with an item i. It has been studied under psychological research and a wide variety of models have been proposed. Response times can serve as outcomes to help explain the underlying factors that are involved in the item solving process [4], which allow researchers understand and possibly improve measurement techniques. A type of response time models has been inspired by Item Response Theory (IRT) [10] [14] , where a similar structure is used [16]. In this framework, response times can be explained by latent parameters related to either items or persons.

A learning environment is a system where participants can interact with exercises, puzzles, challenges or other participants. Its main objective is to foster the conditions the participants' learning. The term e-learning environment as used by [1], encompasses a wide range of applications such as web-based training, Virtual Learning Environments and massively open online courses (MOOCs), among others.

With these technological advancements it is possible to provide assistance to the learners via tips, hints, chatbots or by giving more instructions. This assistance or support can influence not only the probability of correctly answering an item, but also of the time an item demands to be solved by persons. This effect is crucial for assessing the impact and accurately predicting the response time, which can help estimate latent variables related to person and item characteristics. These latent estimates can in turn serve as input for adaptive algorithms or recommender systems in personalised learning. To the best of our knowledge there are no studies regarding the impact of the instructional level of support on response time.

In the next section, the methodology describes the typical behaviour of response time distributions, their transformations through natural logarithms, and their relation with the underlying variables of interest. Three different generic models are proposed to explain the effects of the instructional level of support on log-transformed response times.

In section 3, an analysis with both generated and real-life data is performed. A more visual approach is taken with the generated data navigating through the possibilities of the impact of the level of support in the response time

distribution. On the other hand, with the real-life data a bayesian analysis is performed with the extensions of the generic models to a specific data set.
In section 4, a discussion of the results and limitations is done, laying the ground for future work, later presented in section 5.

2 Methodology

Response time distributions usually follow positively skewed patterns such as Gamma, exGaussian or Weibull distributions. Figure 1 shows a common example of a response time distribution as it decreases its density while the time increases. [9] showed that actual log-transformed response times can be approximated by normal distributions. The logarithmic transformation drastically changes the scale. For instance, if the unit of measure of raw response times is in seconds, then 1 second would become 0 in a log-scale. The equivalent of 18.27 minutes in seconds would be 7, and the conversion of 6.11 hours to seconds would turn to 10.

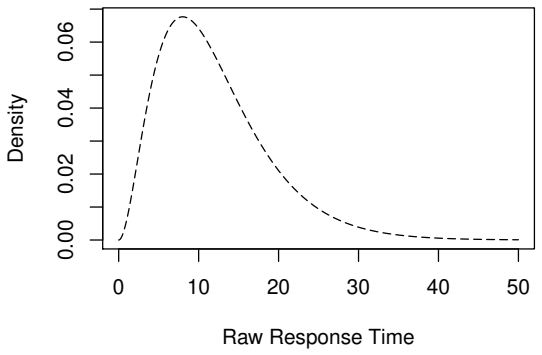

Figure 1: Common example of a raw response time distribution

Response time describes the amount of time a person p requires to solve an item i, which does not imply the accuracy or correctness of the response, but just the time. Nonetheless, the time itself can help measure latent variables or study the relation between them. These variables have particular meanings and interpretations and can generally be classified as time characteristics of items or persons.
From the item side, the time intensity can be interpreted as the time length a particular item requires to be solved, which is not the same as its difficulty. For instance, an item could be both easily solved and time-consuming. Moreover, there a certain cases, such as solving puzzles or mazes, where accuracy does not provide as much information as response time. The main person time characteristic is the speed, which relates to how fast or slow can the person be in solving items.
Following the work by [16], let us define the log response time of a particular item i and person p as a normal distribution with a mean μ_{ip} and an error variance σ^2_{error} as shown in Equation 1. This variance would correspond to the residual differences between the predictions and actual values of log response time.

$$\log T_{ip} \sim \mathcal{N}(\mu_{ip},\, \sigma^2_{error}) \tag{1}$$

In general, the difference between the time intensity and speed can explain the mean μ_{ip}. For those familiar with Item Response Theory, this subtraction is comparable with the 1 Parameter Logistic or Rasch model [14].
Another item characteristic is the time discrimination of an item i. This variable describes how some items are more or less sensitive towards variability of speed than others. In other words, if a person has a constant speed τ, instead of expecting the same linear reduction of τ for all items, it would vary depending on the item. For illustrative purposes, suppose there are two test exercises where the first one involves writing a paragraph and the second one selecting a multiple choice answer. Even if the student is quite fast, solving the first exercise will take at least the time to type or write the words, meanwhile for the second exercise, the student can just select the answer. In this case, the second exercise can be more sensitive to speed in comparison to the first one. The inclusion of time discrimination makes the model analogous with the 2 Parameter Logistic model from IRT.
The new player in these relations of variables is the instructional level of support, which can be interpreted as not belonging to either item or person time characteristics, but rather from the system side. This level of support guides the person in solving an item promoting learning. This new characteristic can be represented in a variety of ways ranging from automatic chatbots in e-learning environments to the manual inclusion of more instructional detail in assembly task training or the provision of hints in educational escape rooms. There may also be other variables valid for specific cases, although special care must be taken to interpret these variables, since these constructs can become intertwined in models.
Considering the support effect as an additional time characteristic, let us define, similarly to the models that stem from Equation 1, the log response time of a particular item i, person p and level of support l as shown in Equation 2.

$$\log T_{ipl} \sim \mathcal{N}(\mu_{ipl},\, \sigma^2_{error}) \tag{2}$$

Maintaining the convention in literature by [11] [16] [8], let us define in this paper:

- λ_i as the time intensity of an item i
- ϕ_i as the time discrimination of an item i towards speed
- τ_p as the speed of a person p
- α_l as the instructional support effect of a level l

This mean μ_{ipl} can be expressed as function of characteristics from the person, item and system side as it can be

seen in Equation 3. This conveys that depending on the estimates and nature of the relation of these latent variables, the mean μ_{ipl} can be steered to the left or right moving the distribution of the log time response. Figure 2 shows an example of how the level of support can help displace the mean. For instance in the figure, the blue dashed distribution could be considered using a medium level of support, so that if there is a change to a higher level of support, the distribution would be the dotted green one. Following the same logic, if the level of support would decrease, the overall mean and distribution would turn into the red one.

$$\mu_{ipl} = f(\lambda_i, \phi_i, \tau_p, \alpha_l) \tag{3}$$

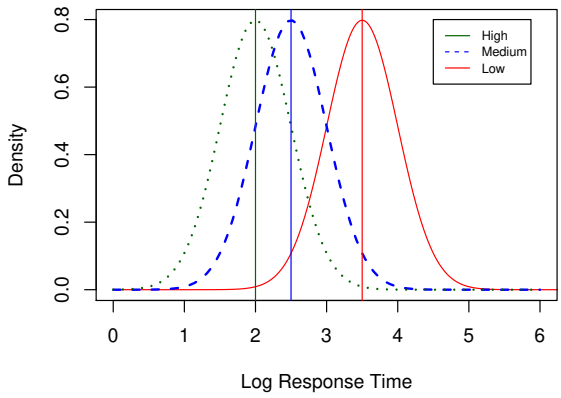

Figure 2: Example of the impact of the level of support

Taking into account the lognormal structure, let us propose three different generic models to express the relation between these latent variables with response time. The first model has a mean that describes a linear relation of the predictor variables with a sole interaction of the item discrimination with the person speed as shown in Equation 4. This sole interaction can be interpreted as the working speed for that particular item [11].This model can serve as a benchmark to compare with the other models. There is a subtraction between the time intensity, the working speed and the level of support decreasing the value of the mean and steering the response time distribution to the left.

The effect of the level of support in this model is the same for all items and persons, which is not the same as its practical impact. For instance, a situation where the time intensity is much larger than the working speed would not be similar to a case where the difference between them is not as large. The level of support could have a bigger impact in the former case, while a modest or perhaps insignificant impact on the latter. Moreover, the effect of the change would not be the same when returning to raw scales due to the logarithmic transformation. Let us imagine an support effect with a magnitude of 0.5. If the log-response time mean is reduced from 3.5 to 3, the effect on raw response time would translate in 13.03 seconds. Meanwhile, that same difference in a log-response time change from 5 to 4.5 would translate to 58.40 seconds.

$$\mu_{ipl} = \lambda_i - \phi_i\tau_p - \alpha_l \tag{4}$$

The second model's mean is shown in Equation 5, which describes a linear relation of the predictor variables with an interaction of the time intensity with the level of support. This interaction can be thought as the item final time-consuming characteristic. For instance, if the system offers high support, then the time intensity would be expected to be lower and with it the overall response time. Similarly, if the support effect is lowest, then the time intensity would be highest and the result from the subtraction would displace the mean to the right. In addition, the multiplicative nature of the relation would make the effect of the level of support even stronger in raw scales.

$$\mu_{ipl} = \lambda_i\alpha_l - \phi_i\tau_p \tag{5}$$

The third and final model's mean describes an overall interaction of the level of support with the subtraction of the time intensity and the working speed, as shown in Equation 6. The level of support acts as an increasing or decreasing factor depending on its level, which can steer the response time distribution to either side. The impact of the support would vary depending on the difference of the item and person variables. This means that if an item would ask much time for a given person, the effect of the support would be larger. Also, if a given person is fast enough for an item, the effect would be smaller.

$$\mu_{ipl} = (\lambda_i - \phi_i\tau_p)\alpha_l \tag{6}$$

Each of these models can be further extended by including additional predictors related to the specific problem context, characteristics of persons or items.

3 Analysis

3.1 Visualisation with Generated Data

In order to have a clearer view of the models, specifically the impact of the level of support, generated data is used to simulate raw response time distributions. A data grid is expanded using multiple combinations of plausible estimates as shown in Table 1. The combinations are a mixture of increasing arithmetic sequences and fixed values, which allow a broader visualisation of the effects of each of these variables towards raw response time. The generated data for model 1 consider a discrimination equal to 1, which is assumed for simplicity in IRT as in the 1P Model [14]. The generated data for model 2 and 3 utilises a fixed time intensity while taking into account varying sensitivities. It is important to remark that these mixture of fixed and varying values is used only for visualisation purposes in order to stress the impact of the level of support.

The first model can be visualised in Figure 3, where it shows a lattice of plots with the possible combinations of the latent variables of interest. These are raw response time density plots with different support effects(0,0.15 and 0.3).

Model	λ	ϕ	τ	α
1	2,2.5,3	1	0.5,1,1.5	0,0.15,0.3
2	2.5	0.5,1,1.5	0.5,1,1.5	0.7,0.85,1
3	2.5	0.5,1,1.5	0.5,1,1.5	1,1.1,1.2

Table 1: Sample values used to generate artificial data

In this case, the discrimination is fixed at 1. It can be seen that when the highest level of support is given(green curves) the distributions concentrate more density near the beginning. Following the same logic, when the lowest level of support is given(blue curves) the distributions flatten decreasing their peaks and spreading probability mass. The vertical lines represent the mean of the distributions. The impact of the support effect can be visualised through the distance between the means, which becomes more significant when the speed decreases and even more so when the time intensity increases.
Similarly to the first model, a lattice of plots representing model 2 is shown in Figure 4. There are three different support effects(0.7,0.85 and 1). Given the multiplicative nature of the interaction, the support effect of 1 shown with the green curve represents the response time with the lowest level of support while the blue curve the highest. It can be appreciated that the distance between the the vertical lines(means) is more considerable when the speed and sensitivities are lowest. It can also be seen that when the speed and discrimination is highest, the overall response time means approach to zero.
A lattice of plots help visualise the behaviour of the third model in Figure 5. There are three different support effects(1,1.1 and 1.2). In this case, the larger values of support effect refer to lower levels of support, since by increasing the overall interaction, the response time grows. The blue curve represents the highest level of support while the green curve the lowest. It can be seen that in this model, the distance between the means is wider when both the discrimination and speed are lower.

3.2 Data Set

For illustrative purposes, the models are implemented with real-life data. The data set was collected within the imec.icon project COSMO, co-partnered by imec itec KU Leuven, among others [7]. The data set consists of:

- 96 participants using VR-supported technology to train for five different assembly tasks with a total of 7161 data points.
- Each task consists from 6 to 12 steps. There are a total of 45 steps for the 5 tasks.
- The number of attempts the participants train at a certain step range from 1 to 4.
- The previous experience in AR/VR is measured with a four-point Likert-type scale and later standardised for modelling.
- The response time of the step is measured in seconds.
- 3 levels of instructional support: Low(L), Medium(M) and High(H) were implemented, and at least one of those levels was always used.

An example of the data set structure is shown in Table 2.

PartID	Step	Support	Attempt	Prev AR/VR	Response time
1	1	H	1	2	22.3
1	2	H	1	2	15.7
1	3	H	1	2	45.3
..	..	..	..	..	..
1	1	M	2	2	24.6
1	2	M	2	2	19.21
..	..	..	..	..	..
96	45	L	3	1	48.1

Table 2: Example of the dataset structure

A brief descriptive analysis of the step log response time is shows that:

- The minimum and maximum values are 1.1 and 6.7 respectively
- The mean is 3.31 and the standard deviation is 0.91

In addition, Figure 6 shows the kernel density plot of the step log response time, which visually suggests it can be approximated by a log normal distribution.

3.3 Model Extension

In this particular data set there are certain characteristics that can be included in the previous models to explain the behaviour of response time. Given the longitudinal nature of data, where several attempts were performed by persons assembling parts, an effect related to learning can be defined. This construct could be interpreted from both the person side, as the learning rate from a person in completing assembly steps, and the item side, as being a time intensity decreasing factor. Therefore, this effect is treated as a generalised fixed effect rather than individual effects specific to persons or items. In addition to the learning effect, an effect related to the previous experience with Augmented and Virtual Reality is taken into account. Following the previous logic, the previous AR and VR experience effect is considered as generalised fixed effect for all persons and items. These new effects are thus represented with γ and ρ for the learning and previous experience characteristics, respectively.
Since for this data set there was always a level of support present in the measurements, a fixed reference is needed to register the change of the level of support. In this case, the High level of instructional support is taken as overall reference for all the models, which means there are two support effects, α_1 and α_2 for Medium and Low level of support respectively. In the case of the first model, given its additive linear relation with the support effect, the High level is fixed at 0. Therefore, the values of α_1 and α_2 are expected to be greater than zero increasing the response time and moving the mean to the right. On the other hand, the

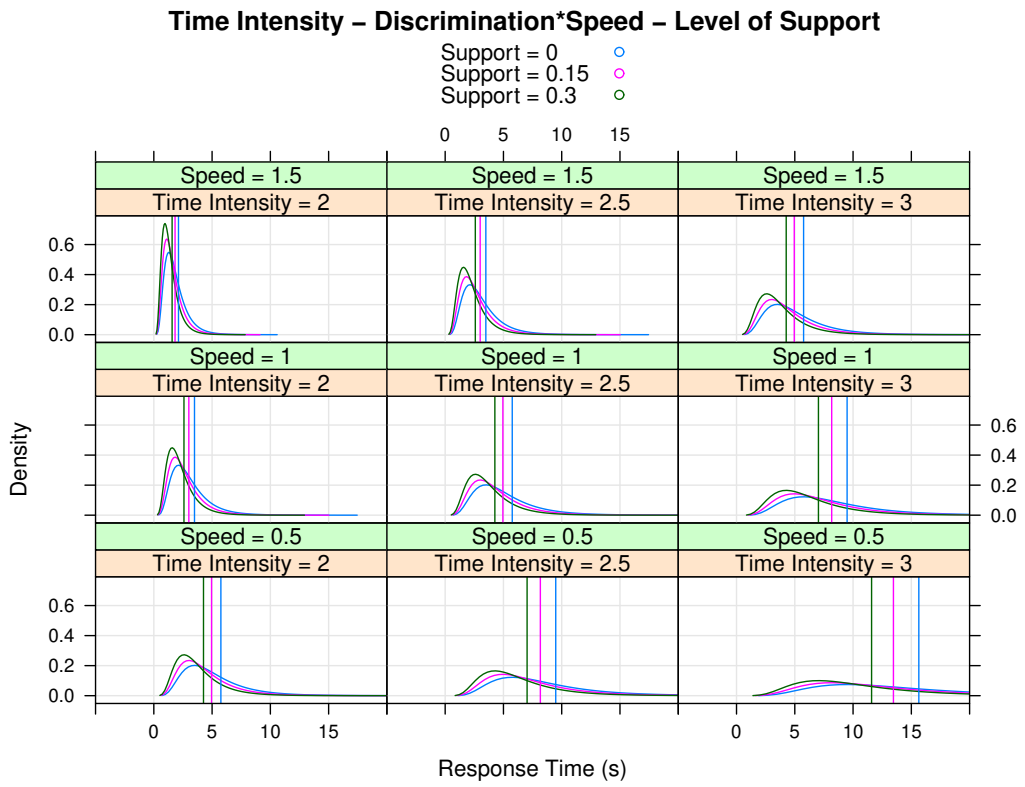

Figure 3: Lattice of Model 1 Response Time Distribution

second and third models, having multiplicative interactions with the support effect, the High level is fixed at 1. This means that the values α_1 and α_2 can be greater or smaller than 1 depending on the relation. For models 2 and 3, if the support effect values are greater than the High support reference of 1, then the response time would increase, which follows the fact the these values represent lower levels of support. Therefore, any support estimates smaller than 1 for models 2 and 3 would not be expected.

3.4 Bayesian Framework

As mentioned previously, the models are estimated through Bayesian methods, in this particular case with Stan by [2] and package RStan by [15], that employs Hamiltonian Monte Carlo to effectively provide a posterior distribution of the log-response time for steps. The number of iterations chosen for these models are 10000 with 2000 burn-in samples with 4 different chains, having a total of post warm-up 36000 samples.

The initialisation of the chains is fixed to stress reproducible results, using preliminary analysis of the models to help define sensible starting points of the variables. Considerable distance is left between the starting points of the different chains, which improves the robustness of the models.

The models suffer from identifiability issues similarly to those of IRT as described in [3]. In general, with the additive identifiability problems for the benchmark model, if a certain constant c is added and subtracted, the mean μ would not be affected. In the same way, for models 2 and 3, if a certain constant c is multiplied and divided, there would not be a displacement of the mean μ. There are several ways to solver these issues, where the chosen one in this work is to set the mean of the speeds to 0 with the priors.

The models use both non-informative and weakly informative priors for the estimation process. Weakly informative priors help regularise and stabilise the chains accumulating probability mass in reasonable regions, meanwhile non-informative priors sparse the probability from $-\infty$ to $+\infty$ specifying no prior knowledge over the measures. As shown in Table 3, a sufficiently wide standard deviation of 100(in log-scale which corresponds to roughly 8.64e+35 years) is set for the support effects, learning and previous experience effects. On the other hand, the other variables($\mu_\lambda, \sigma_\lambda$ and σ_τ) are estimated through the non-informative prior $\mathcal{U}(-\infty, +\infty)$.

Given the large number of parameters and the complex high dimensional space through which the estimation occurs, the discrimination parameter is set to 1 for all items. This paper's goal is to stress the impact of the level of support on response time, nevertheless the readers are encouraged to take into account all of the parameters and use the estimates however they see fit.

3.5 Results

Overall the model parameters reached convergence. A visual case is shown in Figure 7, where the trace plot from Model 1 depicts the different chains that despite starting in different points eventually converge on the estimation of

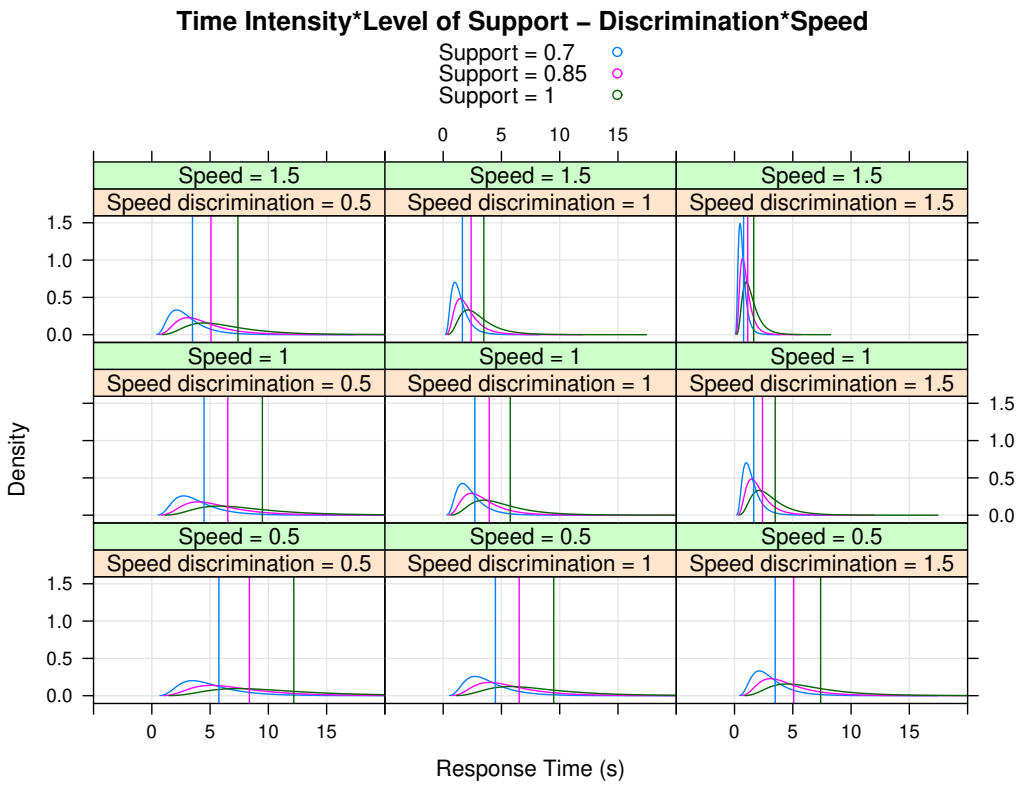

Figure 4: Lattice of Model 2 Response Time Distribution

Priors	Model 1	Model 2	Model 3
Time intensity	$\mathcal{N}(\mu_\lambda,\ \sigma_\lambda)$		
Speed	$\mathcal{N}(0,\ \sigma_\tau)$		
L. Support(Low)	$\mathcal{N}(1,\ 100)$	$\mathcal{N}(1.5,\ 100)$	
L. Support(Medium)	$\mathcal{N}(2,\ 100)$	$\mathcal{N}(2,\ 100)$	
Learning rate	$\mathcal{N}(1,\ 100)$		
Previous Experience	$\mathcal{N}(1,\ 100)$		

Table 3: Weakly informative and non-informative priors chosen for this data set

the support parameters. Moreover, the empirical estimator of geometric ergodicity $\hat{R}$ is equal to 1 in all cases, which further suggests convergence.

Table 4 show the mean estimates of the model parameters. It can be seen that in general the estimates are somewhat similar with the exception of the support effects of the benchmark model with the other models. There is not a considerable difference between the prior mean time intensity with the sample mean (3.31). On the other hand, the sample standard deviation is 0.91 while the residual standard deviation is 0.56, which means part of the variance is explained through that residual, but also probably with the variations of the parameters. The standard deviation of speed is around 0.2 standard deviations from the zero mean.

The support effects show that the difference between the High and Medium level of support is not as considerable as with the low level. This can be seen taking into account the fixed references, in the case of Model 1 the High level was fixed at 0 while for Models 2 and 3 it was fixed at 1. The learning rate has considerable effect on response time. Its effect varies depending on the number of attempts the participant repeats the step. If a participant attempted a particular step for a couple of times, the effect would be a displacement of around 0.26 in log-scales, meanwhile if the participant tried it for a third time the change would be equal to 0.26(2)=0.52. If it happened for a fourth time, the effect would 0.26(3)=0.78. The previous experience has an effect of approximately 0.08, however it cannot be compared with the other parameters since it was standardised prior to the estimation.

Means	μ_λ	σ_λ	σ_τ	α_1	α_2	γ	ρ	σ_{error}
Model 1	3.35	0.63	0.21	0.10	0.65	0.27	0.08	0.56
Model 2	3.37	0.61	0.20	1.03	1.18	0.27	0.08	0.56
Model 3	3.35	0.61	0.20	1.03	1.19	0.26	0.07	0.56

Table 4: Mean Estimate Results

Furthermore, in order to compare the best fitting model for this particular data set, the bayes factor is used. Bayes Factor provides a statistical way to measure the support of a model in favor of another [12]. The bridge sampling algorithm allows to iteratively calculate the bayes factor through samples of the posterior distribution [5], however it requires sufficiently large number of samples in order to create the bridge models [6].

Table 5 show the log bayes factors of the models. The convention for log bayes factor comparison dictates that if a

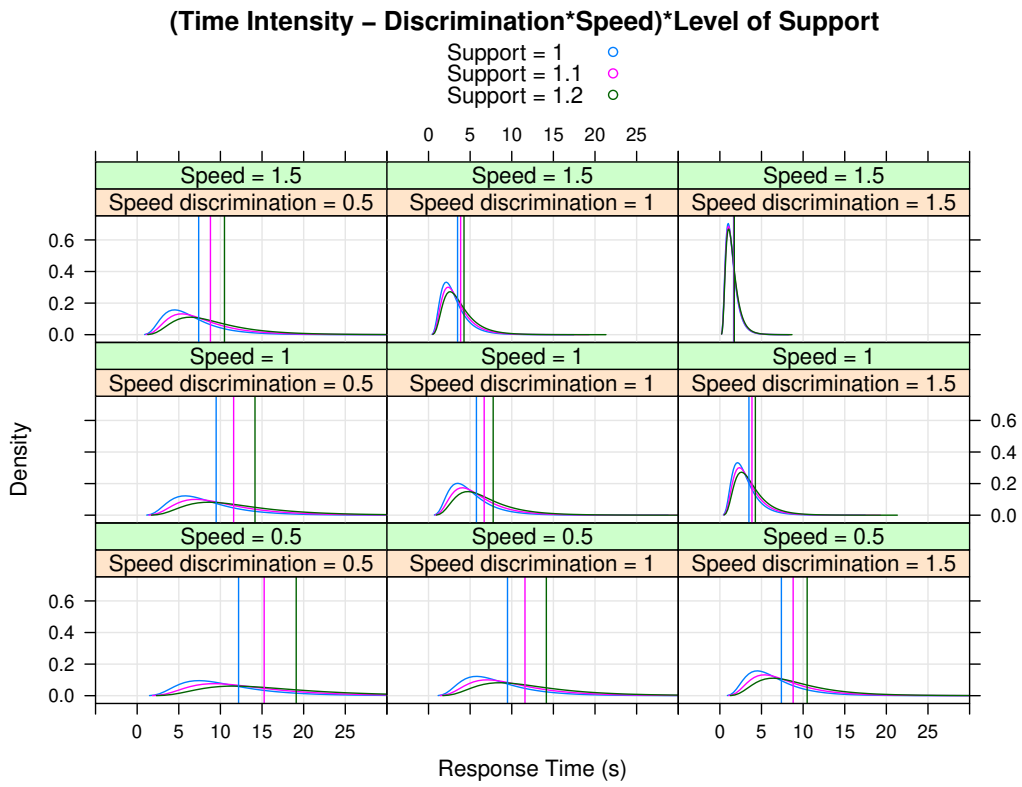

Figure 5: Lattice of Model 3 Response Time Distribution

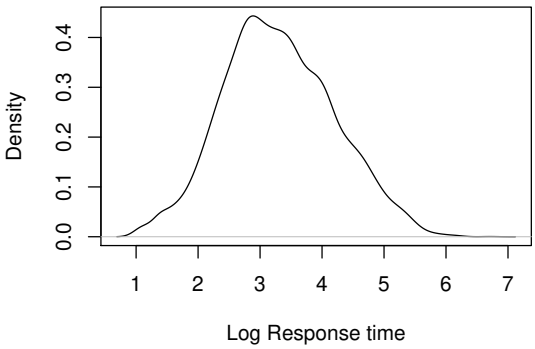

Figure 6: Log Response Time Kernel Density Plot

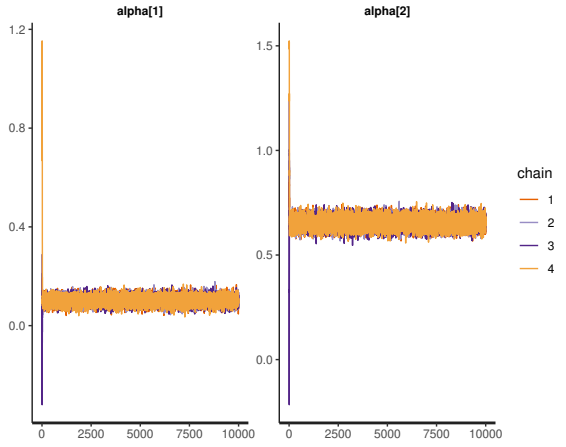

Figure 7: Trace plot support parameter Model 1

Model A has a factor larger than 2 in favor of a Model B, then Model A is preferred. In this case, the benchmark model fares better than the other two possibilities.

4 Discussion and Conclusions

Depending on the complexity and size of the data set, bayesian estimation can become a computationally expensive procedure. Moreover, by including more variables, the parameter space can grow large enough to need an immense amount of iterations to reach convergence. The models are implemented considering an arbitrary fixed value for the discrimination of 1 for all items. This was done for simplicity in calculation as it would add 45 more parameters(one for each item in the data set) considerably increasing the necessary resources to reach convergence.

The results indicate that although all models achieved convergence, the benchmark model fits better than the other alternatives. This behaviour may not necessarily present itself with other data sets from other experiments or remain if more variables were to be included such as an individ-

Bayes Factor	Model 1	Model 2	Model 3
Model 1	-	49.22	40.59
Model 2	-49.22	-	-8.64
Model 3	-40.59	8.64	-

Table 5: Models Bayesian Factor

ual discrimination parameter for each item. Nevertheless, the goal of this paper is to explore and propose models to explain the variability of response time and impact of instructional level of support, and encourage the reader to try, implement and extend the models to other scenarios and choose the best fitting model to the data.
The applications of response time modelling are vast, being the first step towards adaptivity in learning environments. The estimation of the time a person should require to solve a particular item is the key to finding the optimal support, and the algorithm behind this decision can most certainly profit from models that explain the support effect. Works by [13] with the use of Elo-Rating system in learning environments can be adapted to include response times in order to track the growth and current speed of participants as they progress and learn. Moreover, the Elo-Rating system can provide information for personalised item selection, so that items with appropriate time intensities for the persons' current speeds are selected.

5 Future Work

It is important to remark that there are different possibilities to introduce the support effect into the models. Some assumptions that are not considered in this paper(due to the potential of increase of number of parameters, growth in complexity and computational expensiveness), but can be taken into account for future work are:

- Each person has a different speed parameter for each level of support, and that speed dimensions are correlated.
- Each item possesses a different time intensity parameter for each level of support, and that the various time intensity dimensions are correlated.

References

[1] ul H. Anwar, George Magoulas, Jamal Arshad, Asim Majeed, and Diane Sloan. Users' perceptions of e-learning environments and services effectiveness. *Journal of Enterprise Information Management*, 31(1):89–111, 2018. Copyright - © Emerald Publishing Limited 2018; Last updated - 2021-02-19.

[2] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32, 2017.

[3] S. McKay Curtis. Bugs code for item response theory. *Journal of Statistical Software, Code Snippets*, 36(1):1–34, 2010.

[4] Paul De Boeck and Minjeong Jeon. An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10:102, 2019.

[5] Quentin F. Gronau, Alexandra Sarafoglou, Dora Matzke, Alexander Ly, Udo Boehm, Maarten Marsman, David S. Leslie, Jonathan J. Forster, Eric-Jan Wagenmakers, and Helen Steingroever. A tutorial on bridge sampling, 2017.

[6] Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. bridgesampling: An r package for estimating normalizing constants. *Journal of Statistical Software, Articles*, 92(10):1–29, 2020.

[7] imec. Project COSMO cognitive support for manufacturing operations. `https://www.imec-int.com/en/what-we-offer/research-portfolio/cosmo`. Accessed: 2021-04-29.

[8] Konrad Klotzke and Jean-Paul Fox. Modeling dependence structures for response times in a bayesian framework. *Psychometrika*, 84, 05 2019.

[9] Wim J. Van Der Linden, David J. Scrams, and Deborah L. Schnipke. Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23(3):195–210, 1999.

[10] Frederic M Lord. *A theory of test scores*. Psychometric Society, 1952.

[11] Sukaesi Marianti, Jean-Paul Fox, Marianna Avetisyan, Bernard P. Veldkamp, and Jesper Tijmstra. Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39(6):426–451, 2014.

[12] Bruno Nicenboim and Shravan Vasishth. Statistical methods for linguistic research: Foundational ideas—part ii. *Language and Linguistics Compass*, 10(11):591–613, 2016.

[13] Radek Pelánek. Applications of the elo rating system in adaptive educational systems. *Computers & Education*, 98, 04 2016.

[14] Georg Rasch. *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, 1960.

[15] Stan Development Team. RStan: the R interface to Stan, 2020. R package version 2.21.2.

[16] Wim J. van der Linden. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2):181–204, 2006.