One Model that Fits Them All: Psychometrics with Generalized Linear Mixed Effects Models

Wangqian Ju, Susan R VanderPlas, and Heike Hofmann,

Abstract—User experiments are essential for informing researchers what an audience is seeing in a chart. User experiments are generally quite expensive in monetary value and in the time spent getting data. It is crucial that we make the most out of the data we get from participants. Statistically, the best practice for data with repeated measurements is the use of (Generalized) Linear Mixed Effects Models (GLME). These models increase the statistical power, produce more reliable estimates, and provide better interpretability for population-level and individual-level effects. However, in the literature, a two-stage approach for analyzing results from user experiments is commonly used. We compare the two approaches with example data from psychophysics experiments. We present a strategy on how to evolve a two-stage analysis to a single GLME model and showcase diagnostics for each step of that process. We adhere to the best practices of open science and reproducible research by providing open access to all of our code and data.

Index Terms—Computer Society, IEEE, IEEEtran, journal, LaTeX, paper, template.

Contents

1 Introduction 1
1.1 Analyzing Perceptual Experiments 1
1.2 Data description 2
1.3 Notation 3
1.4 Two-stage model 3

2 Example: full hierarchical model for JNDs in bar-charts 3
2.1 The app: Model Buildr 4

3 Results 5

4 Discussion 7

Appendix A: Error Estimates 8

References 8

1 Introduction

Psychophysics research is often concerned with the edge of perception: the line between "same" and "different", also called the just noticeable difference. The term "Just Noticeable Difference" describes the smallest observable change to a stimulus. Kuroda and Hasuo [1] defined the JND in signal X formally as the difference ΔX that is detected (on average) 75% of the time compared to only 50% of the time, assuming the same (constant) stimulus, i.e. $\Delta X = X_{75} - X_{50}$. Graph and visualization perception is a more applied domain, but the same considerations apply: How different do two bars in a bar chart need to be in order for viewers to see them as different? [2] examined this question for bar, pie, and bubble charts, using the method of constant stimuli.

- W. Ju and H. Hofmann are with the Department of Statistics, Iowa State University, Ames, IA, 50010.
  E-mail: see http://www.michaelshell.org/contact.html
- S. VanderPlas is with the Department of Statistics, University of Nebraska Lincoln.

Manuscript received Month DD, YYYY; revised Month DD, YYYY.

While [2] describes an excellent experiment, its findings could be significantly improved using a different approach to statistical modeling. In this paper, we examine different ways to analyze data from psychophysics experiments, with the goal of translating research on linear mixed models into this domain, increasing statistical power, and producing better estimates. Using data from [2], we provide code and visualizations for mixed effects models and compare the results of this analysis to other commonly used methods. Different methods for modeling data from psychophysics experiments have been discussed in other papers [3], but here we take a more expansive view: while we apply these methods to an experiment about JNDs, the methods we discuss are applicable to many experiments in graphics and human perception beyond psychophysics.

1.1 Analyzing Perceptual Experiments

Experimental psychophysics modeling often tries to separate the effects of individual variation from the overall assessment. Such concerns appear in early papers such as [4], where there is a discussion of the best way to estimate variability in the uncertain region between stimulus presence and absence. Even undergraduate perception textbooks address this issue in some fashion, discussing the difference between low and high thresholds for reporting a stimulus compared to different individual perceptual thresholds [5, pg 18].

There are three main options when working with individual-level data while wanting to draw population-level conclusions:

- Naive approach Analyze all individual data using a single summary model, ignoring the additional variability introduced by combining individual-level data.
- Two-stage approach Analyze the data hierarchically, fitting individual-level models and then summarizing individual effects in a second, population-level model.

- Hierarchical approach Analyze the data using a random effects model, where there are (random) individual-level effects and (fixed) population-level effects.

The first option, fitting an aggregate model, is perhaps the most simple at the analysis level. In some cases, the experimenter may design the experiment so that it is balanced, to ensure that participant-level effects average out by having all participants experience all model conditions. This reduces some issues with estimates but does not usually address the correlated errors introduced at the participant level. Aggregate models ignore the correlation between responses from a single participant, which violates the i.i.d. (independent and identically distributed) assumption of almost every statistical model. This has the effect of underestimating the variability in the model, which leads to overestimating the significance and effect size while producing confidence intervals with poor coverage rates. In addition, these models may be subject to effects such as Simpson's paradox [6], where aggregation over main effects changes the signal present in each subcategory.

The second option is a more complicated approach in that it requires the specification of errors and relationships at least two (and sometimes more) levels of model and parameter structures. In the two-stage approach, a model is fit for each participant, and then some aspect of these participant-level models is used as input to a second stage of modeling that summarizes participant-level effects into an aggregate model. Examples include [7], [8], [9], [10], [11], [12], [13], and variations on this approach are taken in [14]. This approach may be intended to isolate participant and item-level effects to produce invariant comparisons [15].

Unfortunately, the relative simplicity of the required code masks the statistical complexity introduced by this approach, especially when additional transformations of the dependent variable are introduced, as is often the case with psychophysics data. Tracking the different error components through two stages of modeling and necessary transformations quickly becomes a difficult and mathematically complex task. In addition, it is also harder to interpret results from these models, as the second model is fit with quantities derived from the first. While this is in some ways related to the error variance issues we just mentioned, it is a much broader problem in that it is hard for readers of papers using these two-stage approaches to grasp the details and meaning of the second (population-level) model stage. As this stage of the model is typically the one that is the most broadly meaningful, the lack of interpretability is a critical flaw in the two-stage approach.

An additional problem that may occur with this approach, especially when fitting generalized linear models, is that some regression models fail to converge. While fitting hundreds of models researchers are almost certain to encounter problems with at least some of the models. These convergence issues can often be resolved with larger sample sizes, i.e. using more data, as is done when fitting a single omnibus model.

Having dispensed with the first and second options listed above, let us consider a third option: fitting a mixed-effects model. In many ways, this option is intended to address the weaknesses of option 1 and option 2 that have previously been identified. A mixed-effects model allows for the estimation of population-level effects, which are considered "fixed" - that is, they represent quantities that are not a function of the sample. Fixed effects exist in contrast to "random" effects - effects which are a function of the sample and can be expected to differ for a different sample. This partition is useful because it allows us to distinguish, for example, effects that are due to human physiology (and shared in common by all participants) from those that are due to individual participants' skill levels or perceptual biases. The inclusion of random effects in the model allows us to include structural terms - for instance, we can add a participant effect, which adds an error term for each individual participant; this has the effect of modeling the relatedness of all trials completed by a single individual. More importantly, however, fitting a single model that includes participant-level effects allows us to combine the advantages of option 2 and option 1: we get better estimates for population-level terms while still accommodating the individual variability that we know exists. This produces more stable estimates for population-level and individual-level terms and also increases the statistical power.

Mixed-effects models are a more general category of models that include models that are more commonly used when analyzing data from perceptual experiments, such as repeated-measures ANOVA; these models are a very simple subset of the broader class of mixed-effects models but have strict requirements about the types of missingness and levels of repetition required. These restrictions often lead to multiple models being fit to different subsets of the experimental data, as in [16], instead of fitting a single overall model. Generalized mixed-effects models allow the experimenter to account for nonlinear, count, and binary/proportion data easily using the same basic modeling framework, and are more tolerant of imbalances in the number of observations per condition.

1.2 Data description

Researchers in [2] aimed to understand how object intensity and separation distance affect the perception of comparison in common visualizations, such as bar chart and pie chart. In this paper, we focus on the bar chart experiments and their corresponding data in [2]. The perception of comparison is determined by measuring the "Just Noticeable Difference" (JND) using the method of Constant Stimuli. There are 28 participants, five equally spaced object intensity levels, and five equally spaced separation distance levels. For each combination of participant, object intensity, and distance, the participant is asked to judge if bar B (comparison bar) is higher than bar A (reference bar) in a bar chart with 10 bars. The height of reference bar A is determined by the intensity level, the distance between bar A and bar B is determined by the distance level, and the height of bar B varies within a small range centered at the height of bar A and equally spaced into 10 levels. For each level of bar B height, each participant evaluates (approximately) 10 trials; these results are aggregated into a probability value. The varying heights of bar B is the Constant Stimuli, and 10 levels of bar B height can result in 10 probability values, which will later be used to fit the first-stage model in the two-stage approach. [2] fit 700 individual logistic regression models to find the JND for each combination of participant, stimulus intensity, and distance.

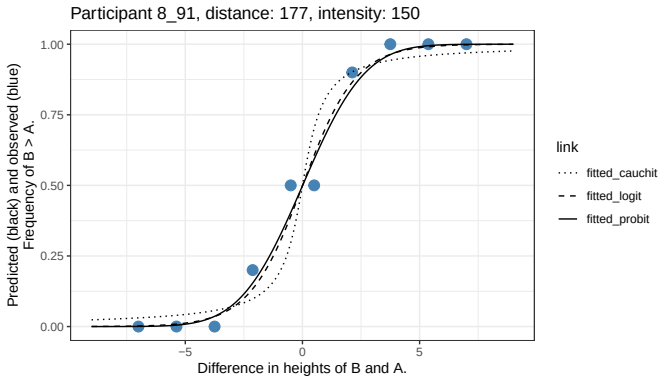

Fig. 1. Logistic regression; linetype to be selected

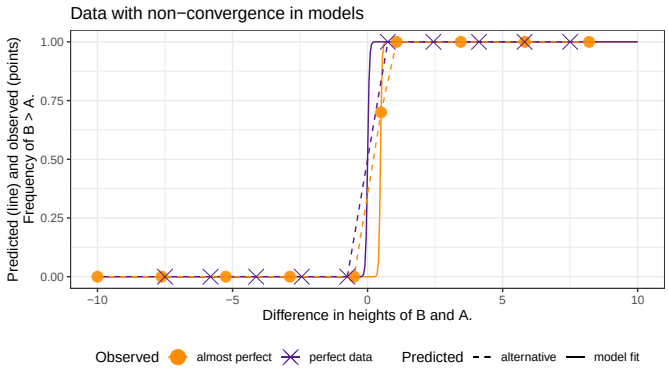

Fig. 2. Problems with Logistic regressions

1.3 Notation

The function that describes the relationship between signal level and prediction performance is called a psychometric function [1]. Different functions can be used as the psychometric function to fit the data. For example, logistic regression uses the sigmoid function, and probit regression uses the cumulative distribution function (CDF) of normal distribution. In some situations, other constraints, such as anchoring points, may necessitate the use of a different psychometric model function, as in [17]. Figure 1 shows the responses made by participant 8-91 of the initial experiment at different signals (difference in heights of B and A) and the fitted psychometric functions. Note that the choice of the psychometric function form should depend on subject matter knowledge regarding signal activation, for example, activation slope, symmetry prior psychometric knowledge regarding the subject of interest. Different choices of the psychometric function form can affect the estimate of just noticeable differences (JND).

For logistic regression

$$\text{logit}P(Y = y) = \mu + \beta x \tag{1}$$

this means:

$$\begin{aligned} \mu + \beta x_{75} &= \text{logit}(0.75) = log(3) \\ \mu + \beta x_{50} &= \text{logit}(0.50) = log(1) = 0 \\ \Delta X = x_{75} - x_{50} &= log(3)/\beta \end{aligned} \tag{2}$$

1.4 Two-stage model

Lu et al fit the data in a two-stage process: first, they fit a logistic regression for each participant's data at each level of distance between bars and intensity level. In the example of the barcharts, this fits two parameters (intercept and slope) for each participant for each of the five levels of distance and five levels of intensity for a total of 700 logistic regressions. From these parameters, a participant-level just noticeable difference is calculated as shown in Equation 2.

In a second step, these just noticeable differences are then combined in a linear model with covariates of distance and intensity. Lu et al show that log-transforming the dependent variable leads to better model performance.

Using this modeling approach the resulting model for just noticeable differences shows significant effects for the distance between bars only, while the height of the reference bar does not factor into the model significantly.

Problems with this approach: Logistic regressions are curious models, in that the convergence of the model fails if the data is "too good". Figure 2 shows two examples of potentially observed data that results in the non-convergence of logistic regression as defined in Equation 1. In both cases, the fitted model chooses estimates for the slope that are very high – alternative models with equally good model fits but far lower slope values are sketched into the figure. Changes in the slope directly affect the estimates for the JND values. In the example, the estimated slope values change from 30.98 and 36.53 in the fitted models to 2.67 (purple) and 2.54 (orange), respectively. This goes in-hand with a ten-fold differences in the associated JND values: the fitted models result in JNDs of 0.04 and 0.03 pixels, respectively, while the alternative fits result in JNDs of 0.41 (purple) and 0.43 pixels (orange).

As a result, the estimated JNDs from the participant-level logistic regressions are more unstable in the case of "too good" data and failed convergence and will thus affect the model fitting at the population level.

Moreover, this two-stage modeling approach introduces extra challenges for variance estimation of the estimated JND in the population-level model. Estimating JND at the participant level has variances, and the variance of each estimated JND will be carried on into the population-level model in a complex mathematical form. When the population-level model is fitted, the variance from the participant-level model will be included as part of the variance of the estimated JND at the population level, which not only increases the variance of the population-level JND estimation but also makes it harder to calculate.

2 Example: full hierarchical model for JNDs in bar-charts

We used the bar chart data from Lu at al. [2] to fit a random effects model as shown below: The data set contains $28 \times 5 \times 5 \times 10 = 7000$ observations since we have 28 participants, 5 levels for separation distance, 5 levels for height or intensity of reference bars (A), and 10 levels for height of comparison bars (B). The height difference in pixels between comparison bar B and reference bar A is the signal to be perceived by the participants.

At the population level, model (Equation 3) fits three coefficients to estimate the average individual's ability to

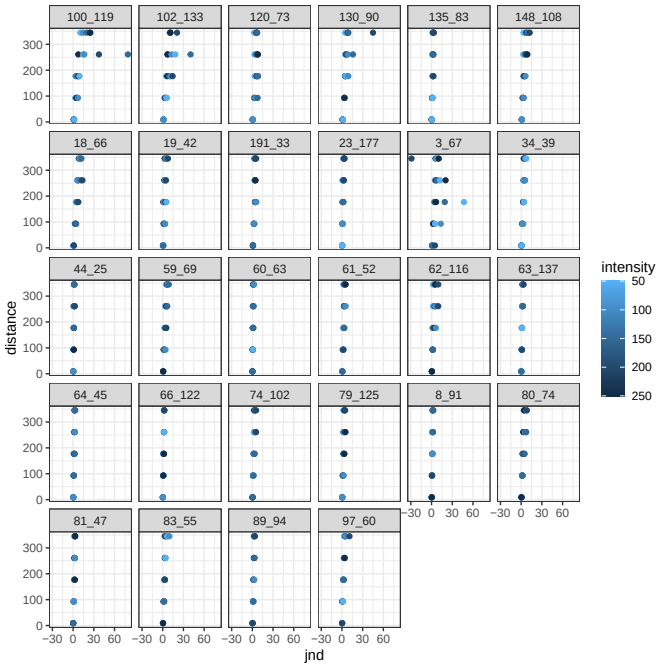

Fig. 3. Just noticeable differences for each participant estimated from each participant's data alone.

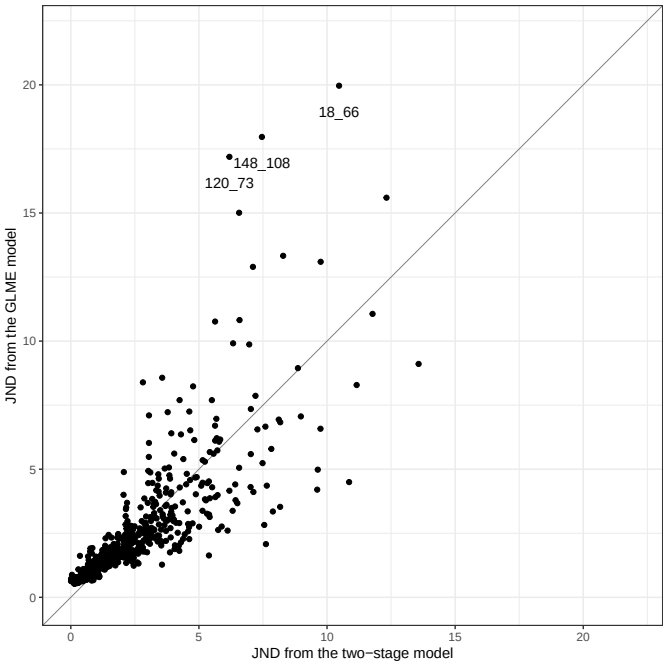

Fig. 4. Scatterplot comparing the JND estimates under Lu's approach and under the approach described here. The correlation is only 0.75, but most of the values are almost identical. In particular, participant 3_67 is problematic for the Lu model. Once this participant's values are excluded the correlation between the estimates jumps to 0.98.

assess the difference between the heights of two bars A and B. This model depends on:

1) the difference in heights between these two bars (encoded as signal S),
2) the horizontal distance between bars A and bars B (encoded as distance D), and
3) the height of bar A (encoded as intensity I). Note that bar A is the reference bar with the fixed height in the experiments.

Both distance and intensity are included using a log transform. This transformation improves model performance significantly. It is appropriate given research on human perception of stimuli [18].

If we extend this model to account for individual effects, we can examine both the population-level trends and the trends for individual participants. We do this by using random effects. For each participant j, $j = 1, ..., 28$, we fit an intercept (u_j) and an effect in signal size S (u_{sj}). We make the usual assumptions for random effects of normality and mutual independence, i.e. $u_j \sim N(0, \sigma_u^2)$ and $u_{sj} \sim N(0, \sigma_s^2)$ i.i.d. with $u_j \perp u_{sj}$ for all j .

$$\begin{aligned} \text{logit } P(Y=1) = \beta_s S &+ \underbrace{\beta_{sd} \log(D) S}_{\text{impact of distance on signal}} \\ &+ \underbrace{\beta_{si} \log(I) S}_{\text{impact of intensity on signal}} \\ &+ \underbrace{u_j + u_{sj} S}_{\text{participants' effects}} \end{aligned} \tag{3}$$

where S is the signal in the study, i.e. the difference in heights between bars B and A, i.e. if S is negative, bar B is shorter than bar A. D and I are the distance between the bars and the height of bar A, respectively. Note that for the overall population, no intercept is fitted, i.e. the point of (relative) subjective equality is set to zero at the population level. Instead, we are interested in how distance and intensity affect a viewer's perception of differences in bars' heights.

This model is fitted in R using the package lme4 [19] and evaluated using package lmerTest [20]. We have also developed an R Shiny app [21] to support users in building GLME models from scratch.

2.1 The app: Model Buildr

The Shiny app Model Buildr assists users in connecting the two-stage modeling approach with the GLME modeling approach. Users can construct their GLME models incrementally, guided by prompts within the app. The process starts with a basic logistic regression for a single participant with one condition level per condition variable. Subsequently, the app extends the logistic regression by incorporating all levels of a chosen condition variable. Users have the option to apply a log transformation to the condition variables and assess the effects of this transformation. These steps aid users in determining the fixed effects to include in the GLME model. Once the fixed effects selection is complete, the app constructs a GLME model by introducing random effects for bias and signal for each participant. The Shiny app Model Buildr can be accessed at https://csafe.shinyapps.io/Model_Buildr/

TABLE 1
Estimates of fixed effects for the model specified by Equation 3.

Term	Estimate	Std. Error	Pr(>|z|)
β_s	2.48	0.059	≤ 0.00001
β_{sd}	-0.34	0.005	≤ 0.00001
β_{si}	-0.04	0.005	≤ 0.00001

TABLE 2
Table 2: Estimated JND according to Equation 4

JND(9, 240)	=	0.7293 pixels
JND(93, 240)	=	1.5351 pixels
JND(177, 240)	=	2.2069 pixels

3 Results

Table 1 gives an overview of the fitted estimates for the model specified by Equation 3 at the population level. All fitted effects are highly significant.

Model to calculate the just noticeable difference at the population level for a distance between bars of d and a height of the reference bar of i. Let n encode the number of bars between the two bars of interest:

$$\begin{aligned} \text{JND}(d,i) &= \frac{\log(3)}{\hat{\beta}_s + \hat{\beta}_{sd}\log(d) + \hat{\beta}_{si}\log(i)} \\ &= \frac{\log(3)}{2.48 - 0.34 \cdot \log(d) - 0.04 \cdot \log(i)} \end{aligned} \tag{4}$$

For intensity of 240 pixels of the reference bar, and distances of $d = 9, 93$, and 177 between bars, the JNDs will result in Table 2.

With the equation for computing the JND above (Equation 4), we are able to plot the estimated population-level JND against the variables of interest. Figure 5 plots JND vs intensity given fixed values of distance. What we can see is that when intensity (height of reference bar A) increases, the estimated JND also increases. The 95% confidence intervals of estimated JND are marked by the ribbon and are calculated using the estimated variance-covariance matrix of fixed effects and the delta method. Note that although the effect of intensity is tiny, it is still significant, and our model with more statistical power is able to capture it. Moreover, increasing the distance between the bars of interest will also increase the estimated JND. The estimated JNDs for each individual are also plotted in Figure 5. Note that the spread of individual lines can be considered as a representation of the standard deviation of the individuals, which is different from the standard errors (represented by the confidence intervals) of the model estimates. The dashed lines in Figure 5 represent estimated JND values that are extrapolated from the model. Weber's Law builds a proportional relationship between the JND and the initial stimuli intensity [22], and the relationship between the estimated interpolated JND and the intensity (height of reference bar A) can be approximated by a proportional relationship. However, the estimated extrapolated JND does not keep a proportional relationship with the intensity as shown in Figure 5. This aligns with some research findings that suggest that Weber's Law fails at low intensities [23].

While the fixed effects in the model specified by Equation 3 allow us to calculate the JND at the population level, the

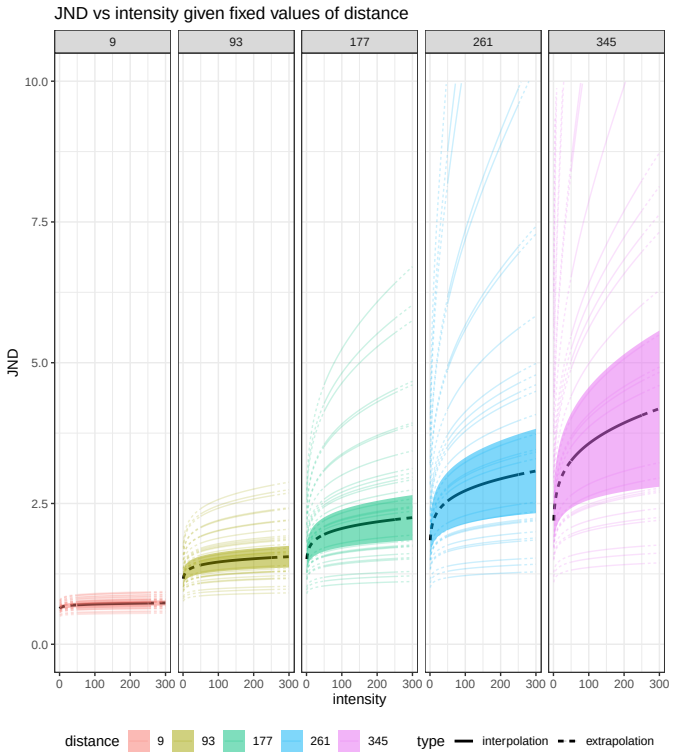

Fig. 5. JND vs intensity given fixed values of distance. The transparent ribbon area shows the 95% confidence intervals for the estimated JNDs. The black line is the estimated population-level JNDs, where the solid part and the dashed part are the interpolation and extrapolation of the model, respectively. Other thin lines are the estimated individual-level JNDs.

fitted random effects allow us to inspect the perceptual skill of each participant. Adding the random effects of a participant is essentially adjusting the population-level estimates based on this participant's performance as shown in Equation 5 and Figure 7.

$$\text{JND}(d,i,j) = \frac{\log(3)}{(\hat{\beta}_s + \hat{u}_{sj}) + \hat{\beta}_{sd}\log(d) + \hat{\beta}_{si}\log(i)} \tag{5}$$

Figure 6 gives an overview of the participant-specific effects. Participants' perceptual skills can be measured directly from the ordering along the slope variable: because the Just Noticeable Difference is inversely proportional to the slope of the estimated probability curve along the signal, the only difference in the JND of two participants j_1 and j_2 is their predicted slope values $\hat{u}_{sj_1}$ and $\hat{u}_{sj_2}$.

The light grey line segments and dots in Figure 6 are intended to provide a reference on how the random effects in the model account for individuals' perceptual skills. All line segments are shown with respect to the overall population average (shown in green). An increase in slope indicates that a participant is able to spot smaller differences between the bars' heights. The dot next to the line segment serves as a point of reference to the theoretical point of equilibrium (0, 0.5). When the line segment is moving away from this point of reference, it means that a participant is exhibiting a subjective bias: when the reference point is on the left of the line segment, a participant has a tendency to respond that B is not larger than A. Three participants (68-122, 191-33, and

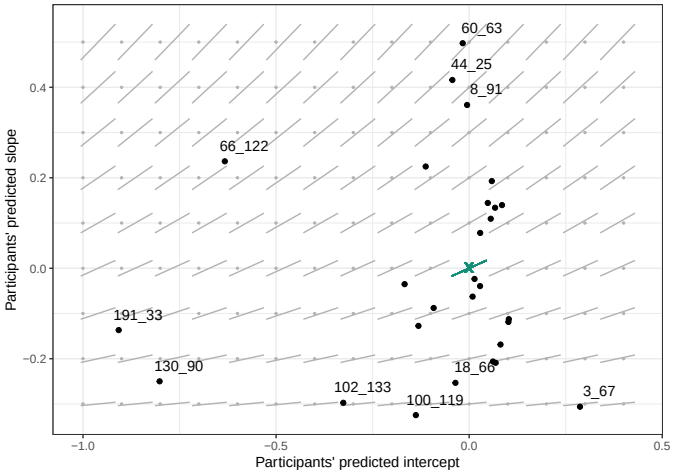

Fig. 6. Overview of participant-based predictions. The green cross indicates the overall population based average. The further away a participant appears on the scatterplot, the more distinctly different their answers are from the population average. The faint grey line segments indicate how the random effects model a participant's perception.

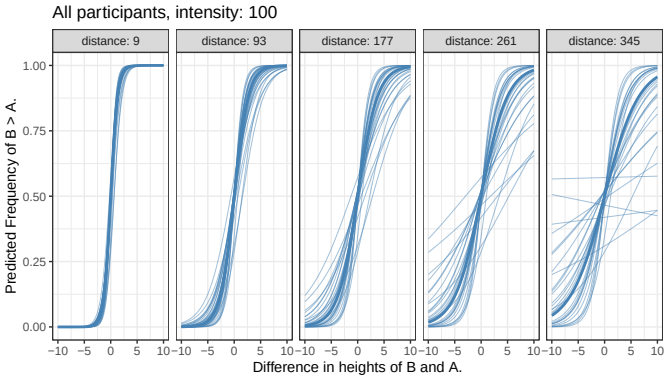

Fig. 7. Predictions for overall population (thick lines) and each of the participants (thin lines)

130-90) show a strong bias in this direction. Participant 3-67 exhibits the strongest bias in the other direction, i.e. has a tendency to respond that B is larger than A.

Responses from individuals labeled in Figure 6 are shown in Figure 8.

Figure 9(a) is the "residuals vs fitted values" plot for the model specified by Equation 3. The grouping structure is seemingly worrying but is actually expected. It is caused by the discrete levels of intensity, distance, and participants. To see how the discrete nature of these variables leads to the structure in the residual plot, we simulate binomial samples for each combination of intensity, distance, participants, and signal values based on the fitted probabilities of model (Equation 3). Then the simulated data is used to fit a model with the same model structure as model (Equation 3), and its residual plot is shown as Figure 9(b). The original data of [2] reveal that some combinations of intensity, distance, participants, and signal values do not have exactly 10 trials, but the simulated data are sampled with fixed 10 trials for each combination. This difference results in the fact that every point in Figure 9(b) can fit in one of the diagonal structures.

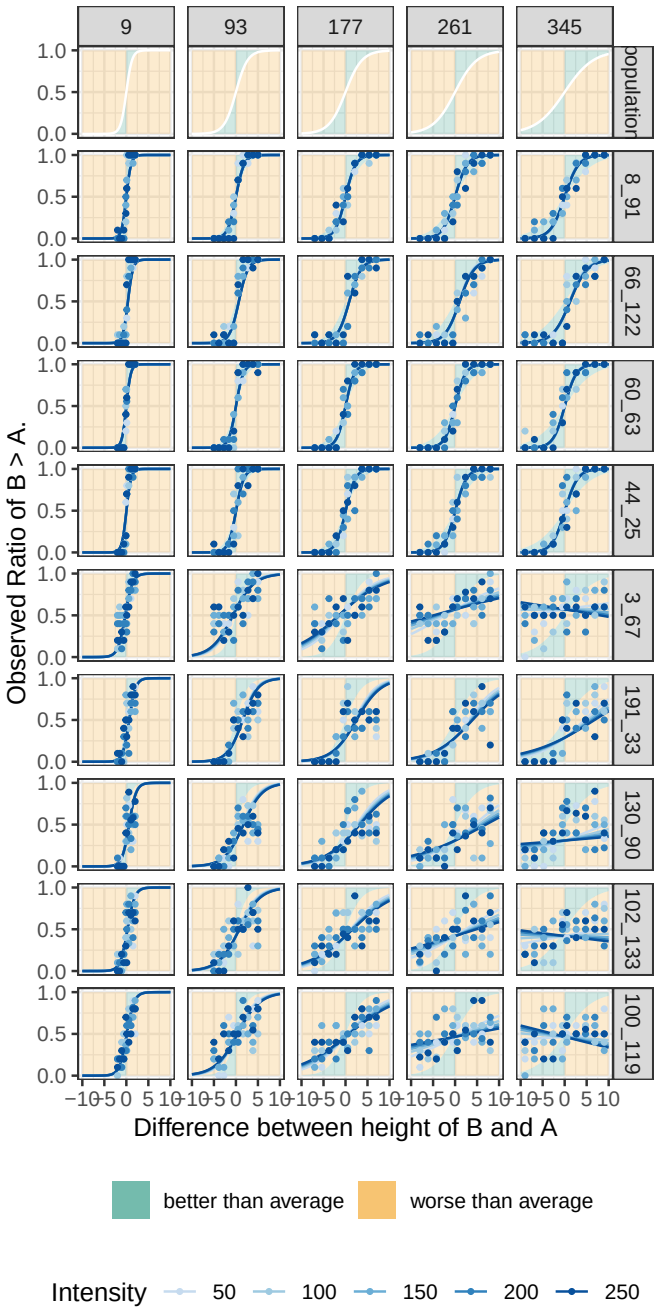

Fig. 8. Overview of participants with the most extreme values in subject-specific skill (slope) and intercept (PSE away from zero).

And we can see from the comparison of the two residual plots that the diagonal structure is caused by the discrete nature of the variables.

In Figure 10, the fitness of the GLME approach and the two-stage approach is compared at the population level. The red dots represent the average probability of predicting that the comparison bar is higher for the 28 participants across all combinations of intensity and distance. And the grey dots represent the probabilities of each individual. The population-level predictions are made by the fixed effects of the GLME model (solid lines) and the second-stage model of the two-stage model (dashed lines).

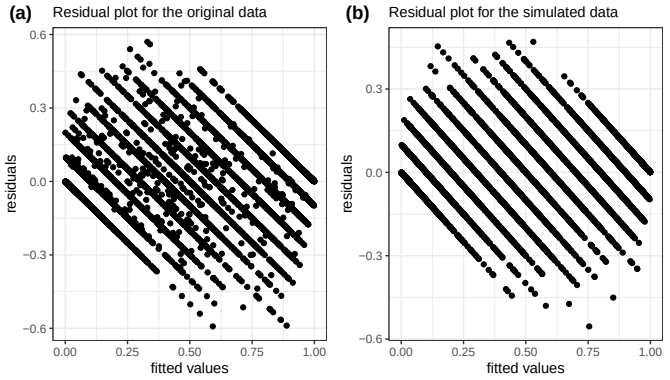

Fig. 9. (a) Residual plot for the original data and our proposed model. (b) Residual plot for the simulated data and our proposed model. The simulated data are binomial samples with fixed 10 trials and fitted probability for each combination of intensity, distance, participants, and signal values. The comparison of the two residual plots show that the diagonal structures presented in the residual plots are actually expected and are caused by the discrete nature of the variables.

What we see from our comparisons between the GLME approach and the two-stage approach is that the two-stage model can better fit the data at the individual level. The reasoning behind this lies in the structural design of these models: the two-stage approach fits a logistic regression model for each combination of distance, intensity, and individual, while the GLME approach pulls information across various individuals when making individual-level predictions. However, Figure 10 shows that the GLME model is more effective in capturing the population-level trend for most combinations of the two variables, which aligns with the core objective of the study and allows for broad generalization.

4 Discussion

In this paper, we demonstrate the advantages of modeling psychophysical data using a generalized linear mixed-effect model (GLMM). The two-stage modeling approach is widely used in the field of psychophysics and is taken by Lu et al. [2]. This approach fits individual-level models first and then builds a single population-level model based on the individual effects. We show that a mixed-effect model provides more benefits than fitting a two-stage model. A mixed-effect model considers individual-level effects as random effects and population-level effects as fixed effects and incorporates both in one single model, which provides more statistical power and better interpretability.

We also developed a shiny app called Model Buildr, which makes it easier for researchers to identify the model, the variables, and the cognitive settings and apply the mixed-effect method to their own topics. This shiny app captures lots of datasets and can be used for many other similar studies.

The data set we used here is from [2]. Using the two-stage approach, Lu et al. [2] fitted a logistic regression to compute a JND for each combination of distance, intensity and participant, and then used the estimated JND as the response variable to build a population-level model. Using the mixed-effect model approach, we can include and analyze both population-level and individual-level effects in just one model.

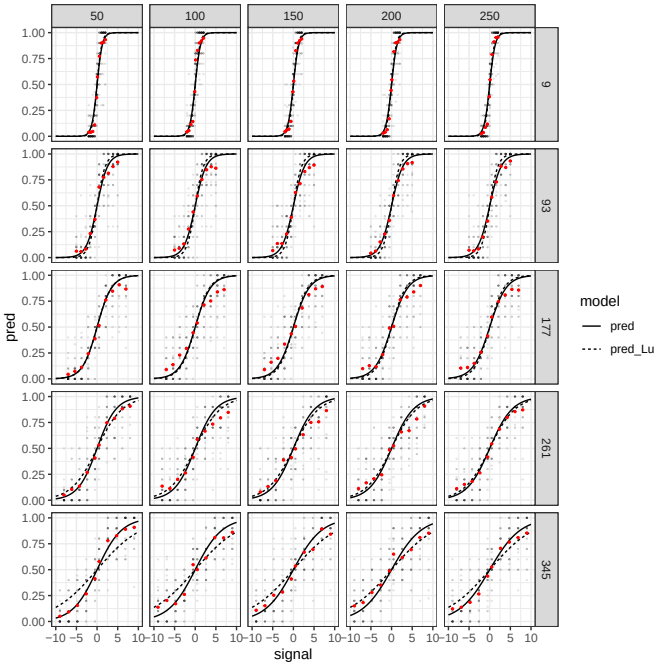

Fig. 10. Comparison of the fitness of the GLME approach and the two-stage approach at the population level. The red dots show the average prediction results of the 28 participants. The fitted population-level curves of the GLME approach and the two-stage approach are presented with solid lines and dashed lines, respectively.

Lu et al. did not find the effect of intensity significant with the two-stage modeling approach. And log-transformation of JND was needed at the population-level model to stabilize the variance. With more statistical power, the mixed-effect model does find the effect of intensity tiny but significant. Log-transformation was applied to variables intensity and distance to align with research on human perception of stimuli [18].

Note that in the original experimental design of [2], the ranges of signal values are the same across different intensity levels for each distance level. But for each intensity level, the ranges of signal values increase as the distance increases. This means that, for a fixed intensity level, more signals are presented to participants as the distance increases; however, for a fixed distance level, the amount of signals in pixels remains the same across different intensity levels. This design conceals the effect of intensity, especially for the participants who are not good at the task. For example, participant "100_119" in Figure 8 was not able to properly spot the difference in height for bar A and B with the current experimental setup when the distance is 345 pixels. The range of signals should have been enlarged to a point where even participants not good at the task can spot the difference. Since the effect of intensity is potentially concealed by the setup, a model with more statistical power is needed in order to find the significant effect of intensity.

Moreover, the mixed-effect model provides better interpretability. For example, if the distance between bar A and bar B increases from 100 pixels to 200 pixels, the task of determining the higher bar becomes more difficult. As a result, the estimated population-level JND increases, and the slope of the predicted probability curve decreases by approximately 0.1554 ($\hat{\beta}_D(\log(200) - \log(100))$). The decrease of the slope

can then be translated into the following statement: on average when bar B is one pixel higher than bar A, the odds of predicting bar B is higher will decrease by 14.39% $(1 - 1/\exp(0.1554))$ if the distance between the two bars increases from 100 pixels to 200 pixels.

The following formula calculates the change of JND:

$$\frac{\log(3)}{\beta} - \frac{\log(3)}{\beta + c} = \frac{\log(3)}{\beta} \frac{c}{\beta + c} \tag{6}$$

We really appreciate the efforts of Lu et al. [2] for making their work open source. These efforts of Lu et al. and other researchers make the discussion about various methods and potential improvements possible. We emphasize the importance of open science since it promotes reproducibility of the work, accessibility for the public, and collaboration for future research. It allows researchers to share their insights, collaborate, and build upon each other's work. Our work is publicly available at: https://github.com/willju-wangqian/one_model_that_fits_them_all The world of open science will be constructed by more and more such efforts and contributions from the community.

Appendix A
Error Estimates

First stage regression:

$$\frac{\log p_{ij\cdot\cdot m}}{1 - \log p_{ij\cdot\cdot m}} = \beta_{0ijm} + \beta_{1ijm} x_{ijklm} + \epsilon_{ijklm}$$

Typically, it is assumed that $\epsilon_{ijklm} \sim N(0, \sigma)$.

As a result of this modeling process, β_{1ijm} is a random variable (because it is a function of the random variables x_{ijklm} and y_{ijklm}). The variance of β_{1ijm} can be calculated using Fisher's information matrix, but that calculation is ancillary to the primary goal here, which is to trace the errors from the first stage regression to the next stage.

The JND calculated for each condition (separation distance i and intensity j) and participant m is a function of β_{1ijm}:

$$JND_{ijm} = \frac{\log 3}{\beta_{1ijm}}.$$

By the delta method, we can calculate the variance of the JND as

$$\mathrm{Var}(JND) = \left(\frac{\log 3}{\beta_{1ijm}^2} \right)^2 \mathrm{Var}\beta_{1ijm}$$

.

Acknowledgments

The authors would like to thank...

References

[1] T. Kuroda and E. Hasuo, "The very first step to start psychophysical experiments," Acoustical Science and Technology, vol. 35, no. 1, pp. 1–9, 2014.

[2] M. Lu, J. Lanir, C. Wang, Y. Yao, W. Zhang, O. Deussen, and H. Huang, "Modeling just noticeable differences in charts," IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis), vol. 28, no. 1, pp. 718–726, 2022.

[3] M. Kay and J. Heer, "Beyond weber's law: A second look at ranking visualizations of correlation," vol. 22, no. 1, pp. 469–478.

[4] M. Treisman and T. R. Watts, "Relation between signal detectability theory and the traditional procedures for measuring sensory thresholds: Estimating d' from results given by the method of constant simuli." Psychological Bulletin, vol. 66, no. 6, pp. 438–454, 1966. [Online]. Available: https://psycnet.apa.org/fulltext/2005-10045-002.pdf

[5] E. B. Goldstein, Sensation and perception, 8th ed. Belmont, Calif.: Thomson Wadsworth, 2010.

[6] A. Alin, "Simpson's paradox," WIREs Computational Statistics, vol. 2, no. 2, pp. 247–250, 2010.

[7] R. A. Rensink and G. Baldridge, "The perception of correlation in scatterplots," vol. 29, no. 3, pp. 1203–1210. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2009.01694.x

[8] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking visualizations of correlation using weber's law," vol. 20, no. 12, pp. 1943–1952.

[9] R. Beecham, J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood, "Map LineUps: Effects of spatial structure on graphical inference," vol. 23, no. 1, pp. 391–400.

[10] W. Nogueira, N. E. Boghdady, F. Langner, E. Gaudrain, and D. Başkent, "Effect of Channel Interaction on Vocal Cue Perception in Cochlear Implant Users," Trends in Hearing, vol. 25, p. 23312165211030166, Jan. 2021, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/23312165211030166

[11] J. D. Hogan, L. M. Fedigan, C. Hiramatsu, S. Kawamura, and A. D. Melin, "Trichromatic perception of flower colour improves resource detection among New World monkeys," Scientific Reports, vol. 8, no. 1, p. 10883, Jul. 2018. [Online]. Available: https://www.nature.com/articles/s41598-018-28997-4

[12] A. Maselli, K. Kilteni, J. López-Moliner, and M. Slater, "The sense of body ownership relaxes temporal constraints for multisensory integration," Scientific Reports, vol. 6, no. 1, p. 30628, Aug. 2016, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/srep30628

[13] T. Koelewijn, E. Gaudrain, T. Tamati, and D. Başkent, "The effects of lexical content, acoustic and linguistic variability, and vocoding on voice cue perception," The Journal of the Acoustical Society of America, vol. 150, no. 3, pp. 1620–1634, Sep. 2021. [Online]. Available: https://doi.org/10.1121/10.0005938

[14] B. M. Hughes, "Just Noticeable Differences in 2D and 3D Bar Charts: A Psychophysical Analysis of Chart Readability," Perceptual and Motor Skills, vol. 92, no. 2, pp. 495–503, Apr. 2001. [Online]. Available: https://doi.org/10.2466/pms.2001.92.2.495

[15] G. Rasch, Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960.

[16] B. Saket, A. Srinivasan, E. D. Ragan, and A. Endert, "Evaluating interactive graphical encodings for data visualization," vol. 24, no. 3, pp. 1316–1330.

[17] J. G. Hollands and B. P. Dyre, "Bias in proportion judgments: The cyclical power model," Psychological Review, vol. 107, no. 3, pp. 500–524, Jul. 2000.

[18] S. Dehaene, "The neural basis of the Weber-Fechner law: a logarithmic mental number line," Trends Cogn Sci, vol. 7, no. 4, pp. 145–147, Apr 2003.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," Journal of Statistical Software, vol. 67, no. 1, pp. 1–48, 2015.

[20] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," Journal of Statistical Software, vol. 82, no. 13, pp. 1–26, 2017.

[21] RStudio, Inc, Easy web applications in R., 2013, uRL: http://www.rstudio.com/shiny/.

[22] E. R. Kandel, Principles of neural science, 5th ed. New York: McGraw-Hill, 2013.

[23] W. F. Norris and C. A. Oliver, System of Diseases of the Eye. J.B. Lippincott, 1900.