

A Modern Test Theory Approach to Selecting Eye Tracking Stimuli

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

BENJAMIN GRAVES

Dr. Edgar C. Merkle, Dissertation Supervisor

MAY 2024

EYE-RT

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

A Modern Test Theory Approach to Selecting Eye Tracking Stimuli

presented by Benjamin Graves, a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Edgar Merkle

Dr. Phillip Wood

Dr. Jeffrey Johnson

Dr. Wesley Bonifay

Acknowledgements

First and foremost, I would like to thank Dr. Ed Merkle for being my advisor for the last 8 years. I could not have asked for a better mentor. I will always be grateful for the knowledge and advice he has given me and for the patience and understanding he showed when life threw me curve balls. My experience here has been a blast, thanks to you.

I would also like to thank Dr. Phil Wood, Dr. Wes Bonifay, and Dr. Jeff Johnson for being on my dissertation committee and sticking with me through the end. I appreciate the insight you have provided me and for pushing me to broaden my perspective of my research by seeing it through other lenses.

I thank my past advisors, Dr. Todd Wiebers, Dr. Wayne Mitchell, and Dr. Erin Buchanan for helping get me to this point. I wouldn't be here without your encouragement and direction. You inspired me to spread my wings and try out new things I don't think I would have otherwise.

I would like to thank my friends and family for being a wonderful support system. I am thankful for all the encouragement my parents and family provided and for dealing with me living so far away. I am also eternally grateful to the friends I made here. Ellen, Kyle, Lauren, and Ron, thank you for all of the adventures and always being there for me. Last but not least, I would like to thank Poppy. This dissertation would have been done a semester sooner if not for your puppy antics.

Contents

Acknowledgements	ii
Table Captions	vi
Figure Captions	viii
Abstract	x
I Introduction	1
Eye Tracking Data	3
Common Models and Extensions	5
Models for Continuous Outcomes	5
Growth Curve Models	7
Models for Categorical Outcomes	7
Models with Mixed Effects	8
Uncommon and Alternative Analyses	10
Time Series Models	10
Hidden Markov Models	11
Saliency Models	12
Machine Learning Models	13
II Model Development	15
Dichotomous IRT Models	15
IRT as Generalized Linear Mixed Models	18

Common IRT Models and Their Extensions Using GLMMs	19
IRT for Eye Tracking	21
Alternative Link Functions and Data Distributions	22
III Data Application	24
Data	24
Analysis and Model Building	29
1PL Model	29
Fitting the Full Model	30
Refitting 1PL Without Large Parameter Images	35
Fitting 1PL with Half the Images	36
Adding Item Covariates	38
Alternative Response Variables and Link Functions	40
Complementary Log-Log Model	41
Poisson Model	42
IV Simulation	47
Simulation 1	47
Design	47
Results	49
Simulation 2	53
Design	53
Results	54

Simulation 3	55
Design	55
Results	56
Simulation 4	58
Design	58
Results	60
Simulation Discussion	65
V General Discussion	67
Dissertation Summary	68
Limitations	69
Future Directions	71
Conclusion	74
References	75
Vita	82

Table captions

- Table 1.* Example of data used in the application.
- Table 2.* 1PL Parameter estimates for selected items.
- Table 3.* Means and standard deviations of the 1PL and CLLM Brier scores and the Poisson model MSE for the simulation condition with 125 observations. The Brier scores and MSE are calculated from a hold out sample of 20% of the total number of observations.
- Table 4.* Means and standard deviations of the 1PL and CLLM Brier scores and the Poisson model MSE for the simulation condition with 250 observations. The Brier scores and MSE are calculated from a hold out sample of 20% of the total number of observations.
- Table 5.* Means and standard deviations of the Brier scores for decreasing image test sets being compared to a model that predicts a constant value. The first column is the initial number of people in the data set, the second column is the set of images the constants are compared to, and the other columns are the average Brier scores for a model that always predicts a response will be correct and one that always predicts a 84% chance of being correct.
- Table 6.* Means and standard deviations from the revised simulation 3. The table is broken down by overall generated sample size, image thinning method, and model. The columns for the 1PL model and CLLM is the average Brier score across the 100 replications and the Poisson column is the average MSE across all reps.

- Table 7.* Means and standard deviations of Brier scores from simulation 4. The table is broken down by overall generated sample size, image thinning method, and model. Brier scores are much higher than in the previous 1PL and CLLM simulations.
- Table 8.* Means and standard deviations for AIC values from simulation 4.
- Table 9.* Proportion of replications that follow the trend of average AIC values from simulation 4.
- Table 10.* Average difference in AIC values from simulation 4.

Figure captions

- Figure 1.* Selection of images that are never incorrectly recalled. A heat map of fixations is superimposed over the images to show where participants concentrate fixations. Participants are drawn to aspects of the images that make them distinct from other images and more easily remembered.
- Figure 2.* The scatter plot presents item parameter estimates from the base 1PL model plotted against the across category hit rates. There is positive trend between the two and, in general, categories are well distributed, though mountains, houses, and badlands are the most difficult.
- Figure 3.* The scatter plot presents item parameter estimates from the 1PL model without items with large parameter estimates. Removing these items has little effect on other item estimates.
- Figure 4.* Item information plots by image category. The curves shows the range of theta values that each item provides the most information for. Many of these curves overlap quite a bit showing that some items are redundant in the range of theta the best predict.
- Figure 5.* The scatter plot presents item parameter estimates from the 1PL model with about half the number of items.
- Figure 6.* Item information for the model with half the number of items. The remaining items still cover a similar range of theta values as the larger model, but is able to do so with fewer items. There are still a few dense areas of some categories where more items could potentially be removed.

- Figure 7.* Comparison of parameter estimates from 1PL model with halved items and additional covariates and the images corresponding hit rates.
- Figure 8.* Comparison of parameter estimates from 1PL model and CLLM using the halved item set and other covariates.
- Figure 9.* Comparison of parameter estimates from CLLM model with halved items and additional covariates and the hit rates for corresponding images.
- Figure 10.* Item information for images in CLLM for each category.
- Figure 11.* Scatter plot of item parameter estimates from Poisson regression predicting fixation count plotted again item hit rates. There is a slight negative correlation between parameter estimates and hit rate for the items. However, there are a few images in the airport category that break this trend.

Abstract

Researchers who conduct eye-tracking studies often consider stimuli to be interchangeable without much consideration for item effects. For example, the saliency of distractor objects in an image plays a role in the difficulty of spotting a specific target object. This, in addition to using a large number of images, use up the resources of both the researcher and participant. The goal of this project is to explore the application of item response models to eye tracking data in order to reduce the number of images used in a study while keeping similar amounts of information. Specifically, I use data from an image memorability study by Bylinskii et al. (2015) to fit item response models formulated in a GLMM framework. Associated memorability scores are used as a standard of comparison for parameter estimates in the item response models. This method provides a way to select images of varying difficulty and to thin out images that overlap in the information they provide. Alternative link functions are explored for use with eye tracking data that is not dichotomous and simulations are conducted to assess various thinning methods as well as their stability. Overall, models tend to retain their predictive ability as the number of images are reduced. These findings suggest that researchers can decrease the number of images used in a study, given that they are high quality and cover a range of difficulty levels. This decrease then saves researchers and participants time and resources.

Keywords. Eye Tracking, Item Response Theory, Generalized Linear Mixed Models, Stimulus Selection

Chapter I

Introduction

Improvements in design and decreasing costs of eye tracking devices have led to an increase in their use for research and practical applications. For research, eye movements are often used as an indication of cognitive processes such as memory, learning, attention, perception, decision making, and language processing. In applied areas they are often used for marketing research, hands-free peripherals, virtual reality, and lie detection. When used as measurement devices, eye trackers provide a wide variety of data types for where and how long a person is looking at something. The data can be combined with other ancillary measures of cognition, such as response latencies. While the use of eye tracking offers a rich source of data, one potential oversight in experimental design is appropriate stimulus selection. Often times stimuli are chosen for convenience or researcher preference. The aim of this dissertation is to propose a method of stimulus standardization for use in eye tracking studies.

Eye tracking studies are often expensive in terms of time and money. Stimulus standardization is useful because it can help to reduce these costs. Most studies uses multiple presentations of tens to hundreds of stimuli, which lead to long, multi-hour sessions for the participant. While having this large amount of data is good for finding effects and fitting models, it creates a situation that leads to excess noise from participants who start to get tired or lose interest. Stimulus selection can also be a time consuming process on the end of the researcher. Stimuli chosen must not only fit the scope of the study, but also be able to induce the phenomenon of interest. For example, it can be difficult to induce scanning behaviors for naturalistic images because familiarity with the scenes or objects lead to rapid, holistic processing. In this situation, the participant gathers just enough information about the scene that

they need to complete the task at hand. Which, in turn, reduces scanning behaviors and fine processing of the scene.

It is better to have fewer high quality stimuli than to have lots of stimuli containing a variety of low and high quality images ([Lapedriza, Pirsivash, Bylinskii, & Torralba, 2013](#)). Developing or applying a method to standardize stimuli can save the time and resources of researchers and participants by cutting down on the number of stimuli used in studies, lowering measurement error, and producing better fitting models.

Applying a method of stimulus standardization to eye tracking images also aids in experiment replication by creating guidelines for stimulus selection and providing a language for reporting the items used in a study. For example, [Holmqvist et al. \(2023\)](#) discuss various common features of eye tracking studies and provide guidelines for what features should be reported in a manuscript in order to aid other researchers attempting to study or replicate similar phenomenon. However, they only focus aspects of experimental procedure, device specifications, and study environment. [Mastergeorge, Kahathuduwa, and Blume \(2021\)](#) discuss the importance of having a standardized set of images for studies attempting to use eye tracking to identify young children or infants at risk for autism spectrum disorder (ASD). In this meta-analysis, the authors discuss the need for researchers to report laboratory or environmental conditions, hardware specifications, and provide the items used in the study in order to aid replication. In the case of ASD, poorly designed eye tracking studies tend to over-estimate differences between ASD and control groups ([Frazier et al., 2017](#)). While the authors do not propose a method to standardize stimuli and only provide a classification of the types of images used in these studies, they do suggest creating a publicly available bank of standardized items for use in eye tracking studies.

In the following pages, I will discuss the various forms of eye tracking data and the types of models that have been applied to this data in the literature. I will then cover the development of item response models for use in this dissertation in chapter two.

In chapter three, an application of these models to eye tracking data is provided. Chapter four contains a simulation to compare the performance of models with a varying number of items. Finally, in chapter five, I will conclude with a discussion of the results and future directions.

Eye Tracking Data

Eye tracking data come in many discrete and continuous forms. Base level output typically consists of gaze points, which are snapshots of where the eyes are currently looking at a fixed point in time. The sampling rate of the snapshots vary depending on the eye-tracker being used but typically falls between 25Hz and 2000Hz. The raw gaze points are not of much use themselves, so many companies will pre-process the data into fixation points and saccades for the user to download. In general, a fixation occurs when a gaze point is held for a certain amount of time and a saccade is the transition between two fixations. However, because eyes make micro adjustments during a fixation, gaze points aren't always in the exact same position during sampling and must be classified into what is considered a fixation and what is considered a saccade. Most of these parsing procedures are some sort of proprietary algorithm for determining what classifies as a fixation. These are often hidden in a black-box and might not produce the output the researcher desires. So, some software will output the raw samples to be parsed into fixations using a method of the researcher's choice.

There are several options for classifying fixations from raw data. These options range from simple noise reduction and gazepoint correction to fully classifying the fixations. For example, [Zhang and Hornof \(2011\)](#) proposes a method of correcting fixations that might be inaccurate by using a mode-of-disparities error correction. Bias in fixation point locations can occur because of the devices themselves or from individual differences in participants. This bias causes fixations to be slightly off

target of the true fixation point. The mode-of-disparities error correction shifts all of the fixation points to fit around a designated (0,0) position.

A person's eyes are constantly moving when fixating on a certain spot. So, individual gaze points must be classified into a single fixation. To make this decision, the raw data is often passed through a filter to classify the gaze points as fixations. [Mack, Belfanti, and Schwarz \(2017\)](#) compile a series of decision trees to help researchers choose an appropriate filtering method and the settings to use based on eye-tracker sampling rates, saccade velocities, and noise levels. Other classification options that have been proposed are the use of two-means clustering ([Hessels, Niehorster, Kemner, & Hooge, 2017](#)) or an algorithmic approach that takes into account individual differences in participants ([van Renswoude et al., 2018](#); [Mould, Foster, Amano, & Oakley, 2012](#)).

Once gaze points have been parsed into fixations, they can be used calculate metrics such as the number of fixations, the fixation duration, the latency till first fixation, and ratios of fixations between target areas. Fixation locations can also be acquired either as Cartesian coordinate output from the software or through the application of a grid based system. If all that matters is the pattern of fixation locations (scan paths), a grid can be placed over an image where each box is labeled with a letter. Then a string of letters can be compiled by starting in the box with the first fixation and following the path of fixations the participant made while making note of the letters when a transition occurs between boxes. These boxes are called of areas of interest (AOI). This can provide binary, on target/off target, or polytomous data for use as a dependent measure. The AOIs also allow for across-area comparisons of the other metrics mentioned above. For example, these comparisons might be simply counting the number of fixations inside and outside of an AOI or assessing the number of transitions between multiple AOIs. Other experimental measures can be tied into eye tracking studies such as reaction time and behavioral responses to tasks.

Data of this size and granularity often come with noise and missing observations. These issues are often due to hardware failures or variability in participants and stimuli. While the best way to reduce these issues is to use a good eye tracking device and to make sure the participant is in a comfortable environment, other methods have been suggested to alleviate these problems. Noise can be reduced by correcting deviations in calibration (Zhang & Hornof, 2011), choosing an appropriate method for fixation parsing (Mack et al., 2017; Hessels et al., 2017; van Renswoude et al., 2018), carefully selecting AOI regions (Hessels, Kemner, van den Boomen, & Hooge, 2016), and various modeling methods. Further discussion on the modeling methods is presented below. However, most of these methods focus on correcting issues with hardware or with dependencies within participants, and they place less emphasis on stimuli. This opens doors to a multitude of potential research opportunities on stimulus selection for eye tracking studies.

Common Models and Extensions

The sections below describe common models used to analyze eye tracking data. These models primarily fall within the generalized linear model framework, such as regression and ANOVA, and are useful for analyzing both continuous and categorical eye tracking outcomes. They can also be extended to growth models and generalized linear mixed models.

Models for Continuous Outcomes

Traditionally, in experimental psychology, eye tracking data has been analyzed and modeled using analysis of variance (ANOVA) and *t*-tests (see Kaakinen (2021) for overview of recent studies and issues in educational psychology). These models are limited to using continuous or vaguely continuous measures as dependent variables, and categorical variables as independent variables. This is often used for comparison between experimental groups, between regions of an image (using AOIs), between

images themselves, or some combination of the three. While these methods are easy to implement and provide valid results with appropriate experimental design and data, they suffer from limitations that more modern models are capable of handling. For example, often times these analyses require some sort of data aggregation either across time, images, participants, or all of the above. This removes the dynamical patterns that could be in the data in favor of an averaged numerical comparison. Another potential issue is the inability of the models to handle spatial data. In this instance, any location information must be categorized and treated as an independent variable. A researcher could compare specific continuous metrics like fixation lengths or number of fixations between different AOIs, but they could not use the AOIs as a dependent variable without going into some sort of logistic regression framework. Similar to the issue with information reduction, fixation locations can only be classified as being on target or off instead of being denoted using coordinates. This binary outcomes removes the ability to do any sort of scan path comparisons or detection of overall patterns in fixations outside of transitions between areas.

Underneath most of these issues is the dependence of observations that occurs in multiple levels of the design. Eye tracking studies often contain a repeated measures aspect, usually in multiple presentations of stimuli, so it would be expected that gaze point observations would be correlated with each other. Dependence also arises within the actual fixation locations and scan paths. Salient portions of images are going to attract and hold attention more than other areas, leading to those fixations being correlated with each other. One potential way to handle this is to remove trials in which the participant was already fixating on the target before a critical time point (Brown-Schmidt, 2009; Heller, Grodner, & Tanenhaus, 2008). In the following paragraphs, model-based solutions to these issues as well as other alternative modeling forms will be discussed.

Growth Curve Models

One potential model that allows for treating time and other variables as continuous is a growth curve model. These models are useful for modeling trajectories and other changes over time. For example, [Magnuson, Dixon, Tanenhaus, and Aslin \(2007\)](#) used growth curve models to assess changes in the proportion of fixations to a target image of a vocalized word between different experimental conditions. The model was able to closely approximate the sigmoidal shape of the proportions as attention changed across time. Similarly, [Indrarathne, Ratajczak, and Kormos \(2018\)](#) used growth curves to model changes in total fixation duration to reading prompts across repeated exposures and different instructional groups. Growth models have also been proposed for popular experimental designs such as the visual world paradigm (see [Zhan \(2018\)](#) for an overview) in order to analyze individual differences in linguistic processing ([Mirman, Dixon, & Magnuson, 2008](#)). These models can handle some random effects, but still require aggregation across either persons or items.

Models for Categorical Outcomes

There are alternative models that allow for the use of categorical outcomes as dependent variables. One alternative is a log-linear model ([Knoeferle, Crocker, Scheepers, & Pickering, 2005](#)) which uses counts or proportions of fixations in each AOI. This model does not require parametric assumptions and can handle issues with homogeneity and independence, but it limits the researcher to only using categorical predictors. So, if a continuous variable such as fixation length is of interest, it must be binned into categorical chunks. Alternatively, [Arai, Van Gompel, and Scheepers \(2007\)](#) converted fixation counts into log-ratios, allowing for a comparison between the proportion of fixations on target and off target across experimental conditions while keeping within a traditional ANOVA framework. This method takes a categorical dependent variable and makes it somewhat continuous, but still requires

the independent variables to be categorical. However, these models still have the same problems as ANOVAs, such as requiring data to be aggregated and ignoring observation dependence.

Models with Mixed Effects

The models above help solve certain situations that the ANOVA framework cannot handle, but they are unable to address all the issues. One solution that can help with the majority of the issues is the use of a hierarchical, mixed, or multilevel model (MLM). [Barr \(2008a\)](#) proposes using multilevel logistic regression when the dependent variable is whether or not the subject is fixating on an AOI. This can then be extended to the use of multilevel linear regressions if the dependent variable is continuous ([Schoemann, Schulte-Mecklenbeck, Renkewitz, & Scherbaum, 2019](#)), allowing for flexibility in what eye tracking variables are included in the model. For example, in the case of a multilevel logistic regression, fixation location can be dichotomized into being on target or off target and time can now be added as a continuous independent variable to account for how fixations change across time. These models allow for more complex functional forms, in order to accommodate non-linear effects. Multilevel models also account for the dependence between observations through the inclusion of random effects, preventing the need for data aggregation and improving power. [Barr \(2008a\)](#) reanalyzed an older data set with two different MLMs, multilevel logistic regression and weighted empirical logit regression (see also [Barr \(2008b\)](#) & [Mozuraitis, Chambers, and Daneman \(2015\)](#)), instead of ANOVA. Both of the new models performed similarly and were able to account for a time effect that was not able to be assessed in the ANOVA.

[Cho, Brown-Schmidt, and Lee \(2018\)](#) expands on the work of [Barr \(2008a\)](#) with the addition of an autoregressive (AR) structure to the generalized linear mixed model (GLMM). They show that responses in a binary time series of eye tracking

data are dependent on the previous response. For example, the probability of a participant looking at the target object or AOI is higher if the participant was already looking at it at the previous time point than if they were not. Variability in the intercept and AR parameters across persons and items suggest a mixed effects model may be more appropriate than just an AR model. However, the use of this specific model removes the ability to assess a time effect, due to the requirement that trends in data be removed. However, the model does allow for the testing of other fixed effects while controlling for previous time points, dependencies in items, and dependencies in persons.

A natural extension of the model above is the ability to handle polytomous response data. In the case of eye tracking data, this could be a situation where multiple AOIs are of interest as the dependent variable. This can be a difficult task because it requires software that can handle multinomial distributions. [Cho, Brown-Schmidt, De Boeck, and Shen \(2020\)](#) proposes the use of a dynamic IRTree model. The dynamic portion of this model is used to assess the trend across time, typically using a GLM or GLMM such as the AR GLMM used in [Cho et al. \(2018\)](#). The IRTree portion, a tree based item response model ([De Boeck, Partchev, et al., 2012](#)), implies a format that each response is a branch on a tree and that choices on lower branches are conditional on the choice at higher branches. In the case of an eye tracking study, potential responses for the upper branch of the tree could be off target vs. on a target and the lower branch responses could be target one vs. target two. Combining these models creates one that allows for each response option to be processed differently while modeling heterogeneity and change processes.

The GLMM models proposed above can fall into an item response framework. These models provide insight into how items are performing within the scope of the experimental paradigm and are not restricted to the binary outcomes that traditional IRT models are. Items that provide information about similar ranges of the latent

trait or those that perform poorly could then be trimmed out of the study. This is helpful because it shortens the measure and reduces redundancy while providing the same information about participants. This creates a more streamlined experience for the participants as well as potentially better model parameter estimates. Another bonus would be item standardization. If well performing items were made available to other researchers, comparisons could more easily be made across different studies and replication studies would be much easier to conduct.

Uncommon and Alternative Analyses

Most of the models above fall within a GLM or GLMM framework and are typical in the field of psychology, but these are not the sole method of analyzing eye tracking data. Just like other fields make use of eye-trackers, there are other models and analytic methods that can handle eye tracking data. The following paragraphs discuss some of these alternative methods and the situations for their use.

Time Series Models

As seen above, time series models offer another potential way to analyze eye tracking data. [McMurray, Klein-Packard, and Tomblin \(2019\)](#) analyze the proportion of fixations on the target image between and within conditions using a bootstrapped difference of time series (see [Oleson, Cavanaugh, McMurray, and Brown \(2017\)](#)). The model allows for the comparison of two groups to determine if their response curves deviate at each point in time. A 4-parameter logistic function is fit to each participant to predict the proportion of fixations across time. These curves smooth each participants' data to be more consistent with the underlying behavioral theory. Bootstrapping is then used to estimate variability in the subject specific parameters and group specific curves, which accounts for within-subject and between-subject variation. A t-test is then conducted across the series of time points between each of the conditions' functions. A new alpha value is calculated based on autocorrelation of

test statistics and then used to identify regions of time where there are significant differences between groups. However, this model cannot account for item effects. So, depending on the variables of interest, it might be better to consider one of the mixed models above.

Hidden Markov Models

Hidden Markov Models (HMM) are useful when outcomes need to be predicted based on eye movements. A HMM takes a sequence of observed events and uses them to learn about a series of discrete unobserved or hidden states. The Markov portion of this model uses knowledge of a single previous state to make predictions about what the current state is or, when current state is known, a future state can be predicted. Knowledge of this state, or a series of states, can then be used to predict the observation or a series of observations.

[Kärrsgård and Lindholm \(2003\)](#) used a HMM to create a predictive text program used in eye-typing software and devices. The model takes information from the string of letters a person has been fixating on and tries to predict what the full word will be. In this case, the fixation points are known and the letter is the hidden state. [Kim, Singh, Thiessen, and Fisher \(2020\)](#) used HMM to aid in the analysis of eye tracking data with moving stimuli, primarily to help with discerning where gaze was fixated when one moving object would obscure another. HMM has also been used to infer the type of visual task that is being carried out using patterns in eye tracking metrics ([Haji-Abolhassani & Clark, 2013](#)). It is also possible to expand these models into a hierarchical framework if states are nested inside other states. For example, [Shi, Wedel, and Pieters \(2013\)](#) propose a three level model of information acquisition from a web page. The lower level accounts for basic eye movements, the middle layer is for the current process being used for information acquisition, and the top layer is for higher order top-down processes that help decide how information should be gathered.

Saliency Models

While they may not be directly applicable to participant data, saliency models could offer insight into scanning behavior or be incorporated into other models as a form of image standardization. Saliency models are more about information within the image itself than about information from the participant. They attempt to model the prominence of specific areas of natural images through the use of variables such as contrast, color, and orientation. They then assign a value to the object or portion of the image based upon how much it stands out from the background around it (Itti & Koch, 2000, 2001). Saliency models are designed to be representative of a bottom-up process for image processing and could be useful in predicting where and in what order people will fixate.

The amount of information that image saliency can provide has been a topic of debate (Jovancevic, Sullivan, & Hayhoe, 2006; Turano, Geruschat, & Baker, 2003) and varies by task and how the model is specified (Navalpakkam & Itti, 2005). For example, if the task is primarily driven by a top-down process, such as searching for an object, saliency models do not perform well in predicting fixations. While correlations between fixation locations and saliency models can be small, they are still greater than chance and are strongest for fixations just after stimulus presentation (Parkhurst, Law, & Niebur, 2002). Saliency models are also better at predicting fixation locations and scanpaths than random models and models that account for bias towards the center of an image (Foulsham & Underwood, 2008).

Saliency models can be extended to include more information about the image than just base color and contrast. Torralba, Oliva, Castelano, and Henderson (2006) propose a contextual guidance model that incorporates contextual information, such as object relations in real environments, into a saliency model in order to provide predictions about where a participant will fixate during search tasks. This could include information such as a person being located near the horizon line of an image

or a painting being located on a vertical surface. This model is based on Bayes theorem and, in this case, implies that the probability of a target object being present in a certain location is the product of the image saliency, top-down knowledge, context-based priors, and prior probability the object would even be in the scene. However, because use of previous knowledge takes a longer time to process, the authors reduce the model to predict the probability of a fixation based only on image saliency and contextual relations (because these occur rapidly when a new scene is presented). Once again, these models correlate with participant fixations during search tasks better than random models and better than saliency models alone.

Machine Learning Models

Finally, like most experimental paradigms, machine learning models and algorithms are being applied to eye tracking data. [Aracena, Basterrech, Snáel, and Velásquez \(2015\)](#) applied a neural network to a data set of pupil dilation measurements and gaze points in order to predict emotions elicited by images. Neural networks are a series of algorithms that use information from nodes in an input layer and pass it through a series of weighted nodes in hidden layers to generate output. In this study, the network was able to classify whether the participant was viewing an image that was rated as having a negative emotional connection or whether the image was positive or neutral in emotion.

Similarly, [Krol and Krol \(2017\)](#) used a neural network to predict the type of strategy a person was using during a decision making task via gaze dispersion and pupil deviation. Another neural network application comes from [Wang, Su, and Ji \(2019\)](#), who propose a novel model they call Dynamic Gaze Transition Network. This model is used to provide a generalized method of gaze point estimation. It uses parts of some of the models above (such as HMM) to develop an algorithm for gaze point prediction that incorporates bottom-up metrics with top-down eye dynamics like

fixations, saccades, and smooth pursuits.

Dalrymple, Jiang, Zhao, and Elison (2019) use a deep learning model and support vector machine to classify infants into different age groups based on the way they visually scan naturalistic images. The deep learning model is used to extract the features (saliency, objects, semantics) and differences in scanning behavior between the age groups, while the support vector machine takes those features and uses them to classify into age groups. While these types of models are very good at prediction and classification, they are “black-box” procedures with uninterpretable parameter estimates. However, if classification is all that matters, they could provide better results than traditional methods.

As can be seen from the model discussion above, many models of eye tracking data are foundations of item response methods or are adjacent to them. So, it would not be a far stretch to apply item response models to eye tracking data in order to provide a method for stimulus selection. The goal of this dissertation is to apply a series of item response models to eye tracking data and then compare parameter estimates from these models to memorability scores from the original study. In the sections below, I first discuss a selection of item response models, presented in a GLMM framework, that are to be applied to eye tracking data. I then provide an application example, using these models, to refine image selection of an eye tracking data set from an image memorability experiment. After that, a series of simulations are conducted to determine how robust results from the application are, test alternative thinning methods, and assess how the models behave when they are misspecified. In the final chapter, I provide an overall discussion of the results, future directions, and concluding remarks.

Chapter II

Model Development

Eye tracking is used as a way to measure the latent, unobservable processes that occur as a person is carrying out a visual or cognitive task. Accurate measurement of these latent processes requires valid and precise measurement devices and a way to link those measures to the latent process. Item response theory (IRT) is a collection of models and methodologies capable of mapping a relationship between latent characteristics and a manifestation of those characteristics. One common example is relating mathematical problems on an assessment to a person's math ability.

However, the often dichotomous response types and large number of stimuli in eye tracking studies make item response models a good candidate for use in relating the stimuli being used in the study to the cognitive processes they are supposed to be measuring or eliciting. While IRT models can get complex and handle various response types, I focus on basic models and their extensions that allow for the accommodation of other covariates found in the data set.

Dichotomous IRT Models

Dichotomous IRT models give a probability for an endorsed or “correct” response for each item in an assessment given the person's location on the latent trait. A general form of this model for the p th person and i th item can be written as follows (Hambleton & Swaminathan, 1985; Rizopoulos, 2007):

$$P(Y_{ip} = 1|\theta_p) = c_i + (1 - c_i)g\{a_i(\theta_p - b_i)\} \quad (1)$$

where Y_{ip} is the dichotomous response from person p for item i , θ_p is the latent ability of person p , and the item parameters are a_i , b_i , and c_i . The notation $g\{\cdot\}$

denotes a link function to transform the linear response into a dichotomous one. Usually this is a logit or probit function which correspond to the standard logistic or normal distributions. The logit link is typically the default option in software, but these two can be equated with the addition of a scaling parameter, $D = 1.702$, to the logit link function in the following manner: $Da_i(\theta_p - b_i)$.

The item parameters provide information about the characteristics of each of the items in the assessment. Parameter b_i is the item difficulty and corresponds to the point on the latent ability scale where the probability of a correct response is $\frac{1+c_i}{2}$ or 50% in models where $c_i = 0$. Larger values of b_i suggest that an item is “more difficult”, or that a person has to be higher on the latent trait to get the item correct. The item discrimination parameter, a_i , corresponds to the slope of the curve and provides a way of telling how well an item differentiates between persons. An item with a high discrimination is better of differentiating between persons of various ability levels than an item with a low discrimination value. Finally, the parameter c_i is a guessing or pseudo-guessing parameter and is the lower asymptote of the curve. This represents the probability that a person will get an item correct based on chance alone.

Freeing or placing constraints onto item parameters of the general form allows one to trade model complexity for ease of estimation. Estimating all of the item parameters gives the formulation for a 3-parameter logistic (3PL) IRT model. Equation 2 presents the 3PL model using an inverse logit link function.

$$P(Y_{ip} = 1|\theta_p) = c_i + (1 - c_i) \frac{\exp\{a_i(\theta_p - b_i)\}}{1 + \exp\{a_i(\theta_p - b_i)\}} \quad (2)$$

While the addition of pseudo-guessing parameter is useful for assessments with multiple choice options, the model requires a very large sample size in order to achieve stability in the parameter estimates. For measures containing around 20 items, it is recommended to have a sample size of at least 1000 observations

(De Ayala, 2013; Yen, 1987). Of course, when increasing the number of items the number of observations must be increased as well.

Constraining $c_i = 0$ simplifies the model and leads to the 2-parameter logistic (2PL) model. This model, in Equation 3, estimates discrimination and difficulty parameters for each item. While it is less flexible than the 3PL, model fit is typically just as good and has better parameter stability (Yen, 1981).

$$P(Y_{ip} = 1|\theta_p) = \frac{\exp\{a_i(\theta_p - b_i)\}}{1 + \exp\{a_i(\theta_p - b_i)\}} \quad (3)$$

The model can be further reduced by placing constraints on the discrimination parameter. Setting $a_i = a$ is the 1-parameter logistic model (1PL). This model, in Equation 4, assumes that the discrimination parameter is constant and equal for all items.

$$P(Y_{ip} = 1|\theta_p) = \frac{\exp\{a(\theta_p - b_i)\}}{1 + \exp\{a(\theta_p - b_i)\}} \quad (4)$$

The most restrictive model is the Rasch model. Similar to the 1PL model above, this model puts a constraint on the discrimination parameter. However, instead of estimating a parameter for a , this model assumes that it is constant and equal to one (i.e., $a = 1$) for all of the items. The Rasch model is presented in Equation 5.

$$P(Y_{ip} = 1|\theta_p) = \frac{\exp\{(\theta_p - b_i)\}}{1 + \exp\{(\theta_p - b_i)\}} \quad (5)$$

The 1PL and Rasch models are mathematically equivalent and mainly differ on philosophical backgrounds. The 1PL model has more flexibility in the discrimination parameter and is mainly used when modeling the data itself is of primary interest. Here the end model is chosen based on the items that lead to the best fitting model and the value of the end discrimination parameter is allowed to be anything. Rasch models are used when construction of the instrument is the goal. Items are chosen

that best fit the mold of the Rasch model (those that fit well with a discrimination parameter of one) and the final questionnaire is decided based on these items.

IRT as Generalized Linear Mixed Models

Item response models can also be framed in terms of generalized linear mixed models (GLMM). [De Boeck et al. \(2011\)](#) presents an outline for various models using this formulation and discusses the potential broad applications to dichotomous and ordered-category data that can be decomposed into binary data. In general, these models follow the format of the simpler models above (Equation 4 & 5) but are slightly rearranged. The GLMM consists of a random component, linking component, and linear component. The Rasch model is presented as a random intercept model below.

$$Y_{ip} \sim \text{Bernoulli}(\pi_{ip}) \quad (6)$$

$$\pi_{ip} = g(\eta_{ip}) \quad (7)$$

$$\eta_{ip} = \theta_p X_{i0} + \sum_{k=1}^K \beta_i X_{ik} \quad (8)$$

Equation 6 is the random component, where the response, Y_{pi} , for person p and item i is distributed as a Bernoulli distribution with probability π . The linking component is once again referenced by $g(\cdot)$ in Equation 7. Similar to the above section, this link function is usually a logit or probit link for dichotomous data. Notation starts to diverge in Equation 8 though. Here, $\theta_p \sim N(0, \sigma_\theta^2)$ and is the random intercept for person. This is akin to the distribution of the latent trait above. The item parameters are represented by β_i . The sign is positive in this formulation, so the item parameters should be interpreted as the easiness of the item instead of the difficulty, that is, items with larger positive values do not require as much of the latent trait to get correct than items with smaller or negative values. The matrix \mathbf{X} is

$(I + 1) \times I$, where the first column is a vector of ones followed by an identity matrix denoting images. Column indexing starts at zero instead of one. In Equation 8, $X_{i0} = 1$ and $X_{ik} = 1$ if $i = k$ ($k = 1, \dots, K$) and 0 otherwise, where K is an index of item covariates. Using this modeling format, typically, item covariates are treated as fixed effects and latent traits are treated as random effects.

Common IRT Models and Their Extensions Using GLMMs

Several basic models in the IRT framework can be viewed as GLMMs and are covered in [De Boeck et al. \(2011\)](#) as well. These models can be estimated using the R package **lme4** and are limited to the class of 1PL and Rasch models. The 2PL model from above cannot be classified as a GLMM because of the introduction of a product from the discrimination parameter and neither can the 3PL because it is a mixture model. However, other GLMM software might be able to handle the 2PL and 3PL cases. This appears limiting at first, but a large number of models can be built off of the Rasch and 1PL model to account for item and person covariates that are not often included in traditional IRT models. These covariates can be indicator covariates, item or person property covariates, or item or person partition covariates. Partition covariates are design factors that group items or persons together either hierarchically or crossed. Property covariates are any other covariates above item or person indicators. A brief description of models using each of these types of covariates is presented below.

Use of only item indicator covariates leads to the traditional Rasch or 1PL model. In terms of the GLMM, the items are treated as fixed effects and persons are treated as a random effect. As mentioned above, the item effects here are interpreted as the easiness of the item. Person parameters can also be extracted from the model via the conditional modes of the random effect, which is akin to maximum a posteriori (MAP) predictions in a traditional IRT context.

Sometimes the items themselves might not be of interest, but the specific category they belong to is. Taking eye tracking stimuli as an example, images could belong to various categories like abstract, ecological, forest, living room, bridge, city, etc. Covariates for these item categories can then be added to the model indicating whether an image belonged to a specific category. The parameter estimates then explain item easiness in terms of their group memberships instead of in terms of the individual items. This type of model can also account for homoscedastic or heteroscedastic error in the items with the addition of a random intercept or random slope effect for item, respectively. These models are considered linear logistic test models.

The last model variant is a multidimensional 1PL. In these models, items are nested into or crossed with other higher level factors. These levels are typically determined by design factors. A good eye tracking example would be using items that are grouped based on an experimental manipulation. For example, if some participants see upright images and others see inverted images. The factors are used as random slopes of the person parameter. However, it is important to make sure that the model is identified in this situation. So, selecting the parameterization that provides the answer to the question being asked is paramount.

Person based covariates can also be added into the models in a similar fashion and also fall under the categories of indicator covariates, property covariates, and partition covariates. However, these are slightly different models. For example, including the person variable as a fixed effect gives rise to a joint maximum likelihood version of the 1PL model. The addition of other fixed person based covariates, such as gender or number of fixations, is a latent regression 1PL. If these are nested or crossed grouping variables being treated as random effects, then a traditional multilevel model is being fit.

IRT for Eye Tracking

The goal of this dissertation is to first fit basic IRT models, like the Rasch and 1PL models, to eye tracking data using typical modeling methods and the GLMM framework. The models will then be extended to include other item or person covariates as discussed above. To the best of my knowledge, these methods have not been applied to validate and standardize eye tracking stimuli. However, as discussed above, GLMMs are starting to be applied to account for dependencies between observations for binary and polytomous data (Cho et al., 2018, 2020). Eye tracking data can be viewed as a form of time series data, and the addition of an auto-regressive component accounts for dependencies between temporally related fixation points. These models are primarily used to reduce bias in parameter estimates for assessing treatment effects or other experimental conditions. While they can be used to account for item dependence, they are not directly used to assess item effects.

Nuthmann, Einhäuser, and Schütz (2017) use GLMMs to predict where a participant will fixate using image saliency and central-bias. Also, in a study closely related to this one, Koutsogiorgi and Michaelides (2022) use linear mixed models (LMMs) to test for individual differences in gaze behavior and response latencies dependent on if an item is worded. These changes in wording use a positive adjective (e.g., “I think I am a decisive person”), an antonym of that word (e.g., “I think I am an indecisive person”), or include “not” before one of the adjectives (e.g., “I think I am not a decisive person” or “not an indecisive person”). While overall item scores did not differ with wording, some gaze behavior and response times did, suggesting that some items were more difficult to comprehend than others.

Alternative Link Functions and Data Distributions

Using a GLMM also allows for flexibility in outcome variable. Not being limited to a dichotomous variable means other forms of eye tracking data could be used as an outcome. For example, a researcher could be interested in finding stimuli that induce scanning behaviors like increased fixation counts. In this case, they can set up the model to predict the number of fixations on an image using a Poisson regression with log link instead of a traditional logit link. Then, they can use the model to select images that are more likely to increase scanning behavior, or that are associated with a variety of scanning behaviors. The formulation of this model is presented below:

$$Y_{pi} \sim \text{Poisson}(\lambda_{ip}) \quad (9)$$

$$\log(\lambda_{ip}) = \theta_p X_{i0} + \sum_{k=1}^K \beta_k X_{ik} \quad (10)$$

Also, the use of packages such as **lme4** allow for alternative link functions, such as the *log*-link in Equation 10, often not found in other IRT programs. Here the response for person p and item i , Y_{ip} , is distributed as a Poisson random variable with parameter λ_{ip} . The *log*-link here takes a count variable and transforms it into linear space where θ_p is the random intercept for person, β_k is the item covariate parameters, and X_{ik} cell of the design matrix for item i and item covariate k .

Alternative link functions can yield models that have all of the psychometric properties of common IRT models while being more parsimonious and providing as good or better fit. For example, the traditional logit or probit models are symmetric link functions and assume symmetric error distributions. [Shim, Bonifay, and Wiedermann \(2022\)](#) show that this symmetry can be relaxed with the use of a one-parameter complementary log-log model (CLLM). This model provides the benefits of a 3PL model while facilitating the estimation process and accommodating smaller sample sizes.

$$Y_{pi} \sim \text{Bernoulli}(\pi_{ip}) \quad (11)$$

$$\ln[-\ln(1 - \pi_{ip})] = \theta_p X_{i0} + \sum_{k=1}^K \beta_i X_{ik} \quad (12)$$

In the following chapters, I will first apply item response models to real eye tracking data and then compare the results of these models to validation data. Then, using simulated data, I will model predictive ability using various sample sizes, decreasing image number, and testing alternative thinning methods. Finally, we close with a discussion of the outcomes, limitations of the models, and future directions.

Chapter III

Data Application

In this chapter, I apply the GLMM based IRT models to a real-world eye tracking data set. The goal of this section is to, first, see how many images can be added to the models while still being computationally feasible. While the models can theoretically handle any number of images, estimation becomes very time consuming and prone to convergence issues as more images are added. The number of images from the full model are then trimmed down through the removal of poor performing images and images that overlap in information they provide. Full model performance is compared to the models with fewer items and models with additional item covariates added. The data is then applied to models using alternative link functions to explore alternative estimation methods for binary outcomes and assess item performance using a non-binary outcome.

Data

The data for the model applications are found in [Bylinskii, Isola, Bainbridge, Torralba, and Oliva \(2015\)](#). In this paper, the authors propose a model to predict naturalistic image memorability using image context and observer behavior. The eye tracking portion of the study uses 630 different target images from 21 different image categories and has 67 participants. Participants only saw a selection of all images, creating a partially-crossed data set. During the study, participants are shown an image and then presented with a forced-choice response prompt to indicate if they remember seeing the image before or not. Target images are repeated 3 times throughout a session with about 50-60 images between each of the specific target presentations. Filler images are also interspersed throughout the session. The authors use the eye tracking data to train an ensemble classifier that first differentiates images

based on fixation maps and then evaluates which fixations were associated with (un)successful image recall. The idea is that a person will successfully recall an image if they fixate in certain locations more than other locations, and knowing these fixations will provide better recall prediction than just base hit rates alone.

Prior to the eye tracking study and model development, the images are given a memorability score based on hit rate or false alarm rate. This creates a list of image memorability within and between each category. The scores are calculated from two studies conducted over Amazon Mechanical Turk (AMT1 and AMT2), and one in a laboratory setting carried out prior to the eye tracking study. The main difference in the studies is how the images were presented to participants. Participants in AMT1 only saw target and filler images from the same category. So, if the participant was in the amusement park group, they only saw images of amusement parks. Participants in the AMT2 and in lab studies were presented images across all the categories. A similar paradigm is used for the eye tracking study.

Memorability scores for this study are calculated as hit rate (**HR**) and false alarm rate (**FAR**). Hit rate is the percentage of time a previously shown image is recalled and false alarm rate is the percentage of time a novel image is “recalled”. More specifically, a “hit” is defined as a correct recall of a previously presented image and a “miss” is when this image is not recognized. A “correct rejection” is when the participant does not recognize an image on the first presentation and a “false alarm” is when they say they do recognize an image on the first presentation. Responses in each of these categories are summed up for each image and used in Equations 13 and 14.

$$\mathbf{HR} = \frac{hits(I)}{hits(I) + misses(I)} \times 100 \quad (13)$$

$$\mathbf{FAR} = \frac{false\ alarms(I)}{false\ alarms(I) + correct\ rejection(I)} \times 100 \quad (14)$$

Averages for HR and FAR were also calculated for all of the images within a category. To provide an example, images in the amusement park category have a higher hit rate than other categories and are more easily remembered than images from other categories. There are also images within this category that are more easily remembered than other images within the same category. Often, these images contain some other information apart from the main context of the image. For example, an image might also include a person, a face, or an animal.

For this dissertation, memorability scores are used to validate the results from the models fit. In the case of an item response model, images with higher memorability scores (i.e. high hit rate, low false alarm rate) should have parameter estimates suggesting that the image is “easier” in difficulty. The same would be true for comparing image categories. For example, amusement park images are more memorable than badlands images, which are more memorable than bridge images. And within those categories, images with distinct features are more memorable than those with generic features. See figure 7 of [Bylinskii et al. \(2015\)](#) for graphical example.

The data from this study can be found online and includes files for memorability scores (<http://figrim.mit.edu/>) and the eye tracking data (http://figrim.mit.edu/index_eyetracking.html). The code to format the MATLAB files can be found on the OSF project page for this dissertation (<https://osf.io/yvf3c/>). The eye tracking data consists of image file names, image categories, the number of times the image had been presented, fixation number, fixation location coordinates, and memory recall variable. The memory recall variable is a factor with four options based on whether they had previously seen the image and their response (hit, false alarm, miss, correct rejection). A few example observations can be found in Table 1.

In addition to the provided variables, I transform the memory recall variable into a binary outcome for correct (hit, correct rejection) and incorrect (false alarm, miss)

responses. This dichotomous response variable serves as the outcome in the more traditional IRT models, where each image file is treated as an item. Total fixation counts are also calculated for each image presentation. The total fixation counts and recall test phase are added into the model as covariates.

ParID	Pres	Image	Category	Recall	Fixation	X	Y	BinRes
17	enc	sun_akeicfsgblregoeok.jpg	cockpit	4	3	446.50	606.50	1
36	enc	sun_btkjvzkfmmmaqhn.jpg	skyscraper	4	3	327.80	347.20	1
20	rec2	sun_bjvbeslhmgqsfmx.jpg	badlands	1	5	584.90	521.90	1
38	enc	sun_bzfcixnmhtzpmwjj.jpg	highway	4	5	557.40	583.70	1
34	rec2	sun_bdwgezrbolobluib.jpg	highway	1	2	457.00	666.00	1
36	rec2	sun_bxlbhoirpbrfqrhu.jpg	skyscraper	1	2	549.40	504.50	1
37	enc	sun_bbjuskmlqfslipfw.jpg	skyscraper	4	4	466.70	681.50	1
57	rec	sun_aarzwglnlqrfnzvf.jpg	amusement_park	3	3	172.60	592.40	0
5	rec	sun_bisqsfcezuskvyu.jpg	bridge	1	1	513.20	509.20	1
23	rec	sun_brknwnzwrmdobiph.jpg	dining_room	1	6	266.30	297.40	1

Table 1

Example of data used in the application.

Analysis and Model Building

Analyses are carried out using R (R Core Team, 2020) and the package **lme4** (Bates, Mächler, Bolker, & Walker, 2015). This is a popular package for estimating GLMMs in R and uses similar syntax to the base `lm()` function for linear regressions. The basic arguments for setting up a GLMM with a logit link is formatted as follows:

```
modex <- glmer(response ~ item + (1 | person),  
              data = mydata,  
              family = binomial)
```

In the `glmer()` function, the model is specified in the first argument. The dichotomous response variable is `response`, `item` is the fixed effect for the items in the measure, and `(1|person)` is the random intercept for person. The intercept for the model can be removed with the addition of `-1` before the fixed effect and random slopes for the person variable can be added after the `1` in the random effect portion (e.g., `-1 + item` and `(1 + covariate | person)`, respectively). Examples using the image data and models discussed previously are presented below.

1PL Model

In the following sections, I illustrate how a model with a large number of images can be reduced do while still providing a similar amount of information. First, a model with as many image parameters as computationally possible is fit in the first section. This model then serves as a base for image reduction in the next sections. Smaller models are fit first by removing images that the model predicts are always scored correctly or incorrectly and then by thinning images out by removing those that have similar parameter estimates or overlapping item information. In each section, item parameters are discussed and compared to the hit rates from the original data, Brier scores are used to compare the models based on predictive ability,

and item information is used to thin out the number of images.

Fitting the Full Model. The data set used for the applications consists of 630 images partially crossed with people. While GLMMs are capable of handling a large number of variables, they might not be able to estimate a parameter for each item with such a limited number of observations or in a timely manner. These estimation issues are also compounded when images have perfect correct or incorrect response patterns. So, in order to set up a baseline model to build from, I first estimate a model that converges in a reasonable amount of time and includes as many images as possible.

Of the 630 images, ten images have zero incorrect responses across all presentation points. These items provide no value in differentiating participants across ability and cause issues with estimation and they are removed from the data set. This leaves 620 images for use in the 1PL model, which is still an extremely large number of parameters to try and estimate in a reasonable amount of time. To aid with this, a series of increasingly complex models are fit in an attempt to include as many categories and images into the baseline model as possible. Categories are added in alphabetical order during data import, so the first category of images is subset out of the full data set and then used to fit a 1PL model using the code below. For the initial model, only participant responses from the first recall phase are used. In order to help with estimation, the `bobyqa` optimizer from the `minqa` package is used (Bates, Mullen, Nash, & Varadhan, 2014). The intercept term is also removed from the model so that the parameter estimates are the easiness of each item, as opposed to a change in easiness compared to a reference item.

```
matchlme4 <- glmer(BinRes ~ -1 + Image + (1 | ParID),  
                  data = matching,  
                  family = 'binomial',  
                  glmerControl(optimizer = 'bobyqa'))
```

If this model converges, the next category of images is added to the subset data and the model is fit again with those images included. If not, the category is removed from the data subset and the next is added. This is repeated with all categories to find the largest number of items that could be conveniently estimated. The final model contains 236 images from 8 categories from 3761 rows in the data set. The categories include airport terminals, amusement parks, badlands, bathrooms, bridges, cockpits, houses, and mountains.

Parameter estimates from the 1PL model are found in Table 2. Once again, these can be interpreted as the easiness of the items, so lower parameter estimates mean an image is more difficult to remember and higher ones mean the image is easier to remember. Several of the items have noticeably large parameter estimates. On a logit scale, a parameter estimate in excess of plus or minus four suggest near certainty that an item will be scored as incorrect or correct. There are 8 items in the 1PL model having parameter estimates around 17, suggesting that they are almost always scored as correct. Upon inspecting the data, there is not a case of the items being missed in the first recall stage but there are a few cases in the encoding or second recall stage, which is why they were not removed in the initial pass removing items with 100% accuracy. It is best to not include these items because they offer little to no information about persons.

At first glance, most of these images would be very memorable because most include humans featured prominently. Inspection of the fixation maps supports this argument for some of these images. However, there are other images in the respective categories with just as distinctive humans that do not have parameters outside of expected bounds. As for the other images, there are distinct aspects that could make the images more memorable, such as assistance bars in one of the bathrooms or branding in the airport terminals. Figure 1 presents these images with their fixation maps overlaid.

Figure 1. Selection of images that are never incorrectly recalled. A heat map of fixations is superimposed over the images to show where participants concentrate fixations. Participants are drawn to aspects of the images that make them distinct from other images and more easily remembered.



Image	Estimates	Category	HR	FAR
sun_aahaqbedkotbsthx.jpg	0.930	Amusement Park	0.78	0.11
sun_aanjcbzdvbtsdy.jpg	1.385	Bathroom	0.55	0.22
sun_aarzwginlqrfnzvf.jpg	-0.155	Amusement Park	0.39	0.02
sun_aauqnwburbymyzyc.jpg	3.060	Bathroom	0.84	0.01
sun_aavgrdvhbtrmxhpy.jpg	1.419	Amusement Park	0.68	0.17
sun_abagndtkcaikrgux.jpg	1.453	Bridge	0.54	0.01
sun_abchldffyzrykkwt.jpg	0.896	Cockpit	0.69	0.25
sun_abgtntbkjhtwkihs.jpg	2.192	Amusement Park	0.88	0.01
sun_abjkjkpuknapfhzy.jpg	1.587	Amusement Park	0.91	0.00
sun_abrlnvzmqklirrnl.jpg	2.949	Airport Terminal	0.88	0.11
sun_abvemrbhejpzgeqi.jpg	2.303	Amusement Park	0.90	0.04
sun_abysfcvgcetptqxa.jpg	0.826	Amusement Park	0.78	0.07
sun_acsgbuurtwghzjel.jpg	1.785	Bathroom	0.55	0.07
sun_actnsdjpqgtzjwcw.jpg	1.847	Airport Terminal	0.77	0.04
sun_adbkymqhrojsynn.jpg	1.417	Cockpit	0.88	0.14
sun_adbmwyujnswnooci.jpg	2.668	Cockpit	0.95	0.11
sun_addtuvmgwajlldow.jpg	1.729	Cockpit	0.70	0.32
sun_adqtdtzyjgyhaa.jpg	0.221	Bathroom	0.61	0.16
sun_aebnrieefokejpz.jpg	2.240	Airport Terminal	0.79	0.04
sun_aekkynucosyodgdj.jpg	1.673	Badlands	0.92	0.07

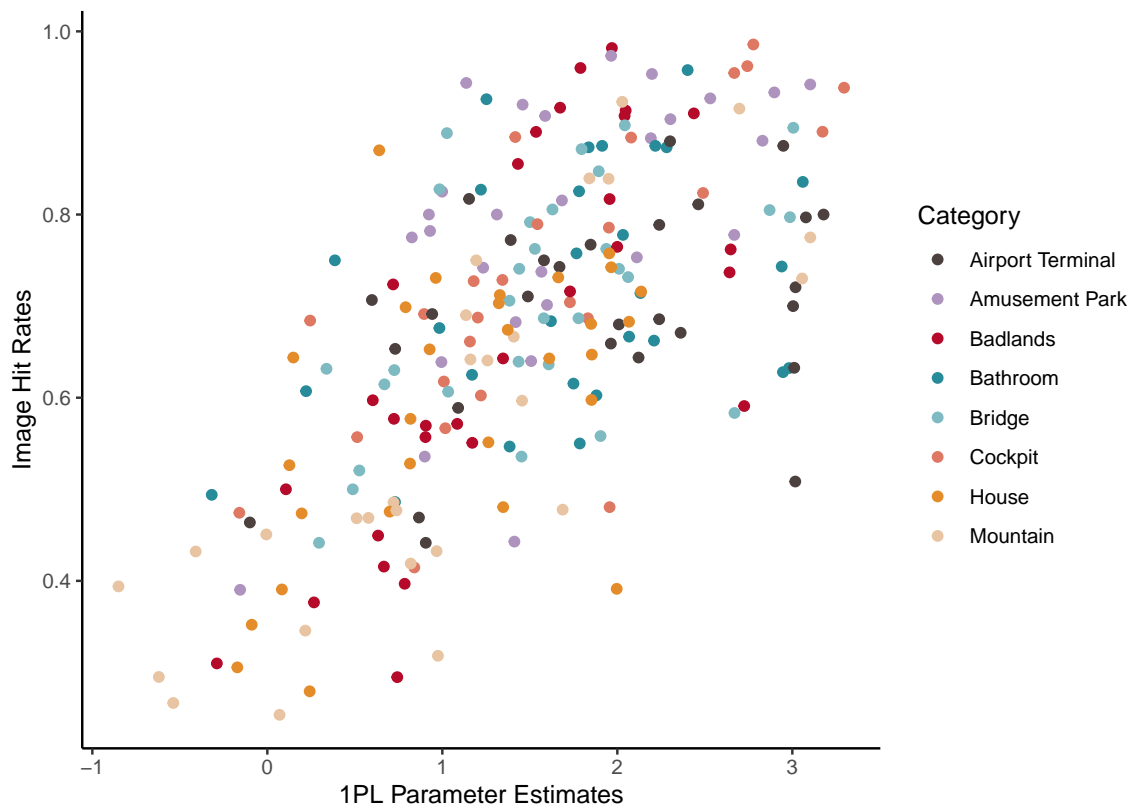
Table 2

1PL Parameter estimates for selected items.

Figure 2 presents a scatter plot of the parameter estimates, not including the excessively large ones, against the across category hit rates from [Bylinskii et al. \(2015\)](#). Four of the images did not have hit rate data, so they are removed from the

graph and further analyses involving hit rates. There is a positive correlation between the model item parameters and the across category hit rates of $r = 0.65$.

Figure 2. The scatter plot presents item parameter estimates from the base 1PL model plotted against the across category hit rates. There is positive trend between the two and, in general, categories are well distributed, though mountains, houses, and badlands are the most difficult.



We can also compare the models by predictive ability with the use of Brier scores. These scores are calculated with the following formula:

$$BrierScore = \frac{\sum_{i,p}(P_{ip} - Y_{ip})^2}{N} \quad (15)$$

This is essentially the mean squared error where P_{ip} is the predicted probability of

a correct response for person p and item i , Y_{ip} is the observed response, and N is the total number of observations. The Brier score when using hit rate values is 0.205, while the 1PL model has a Brier score of 0.14. So, the full 1PL model is better at correctly predicting if a person will remember an image than using hit rate alone.

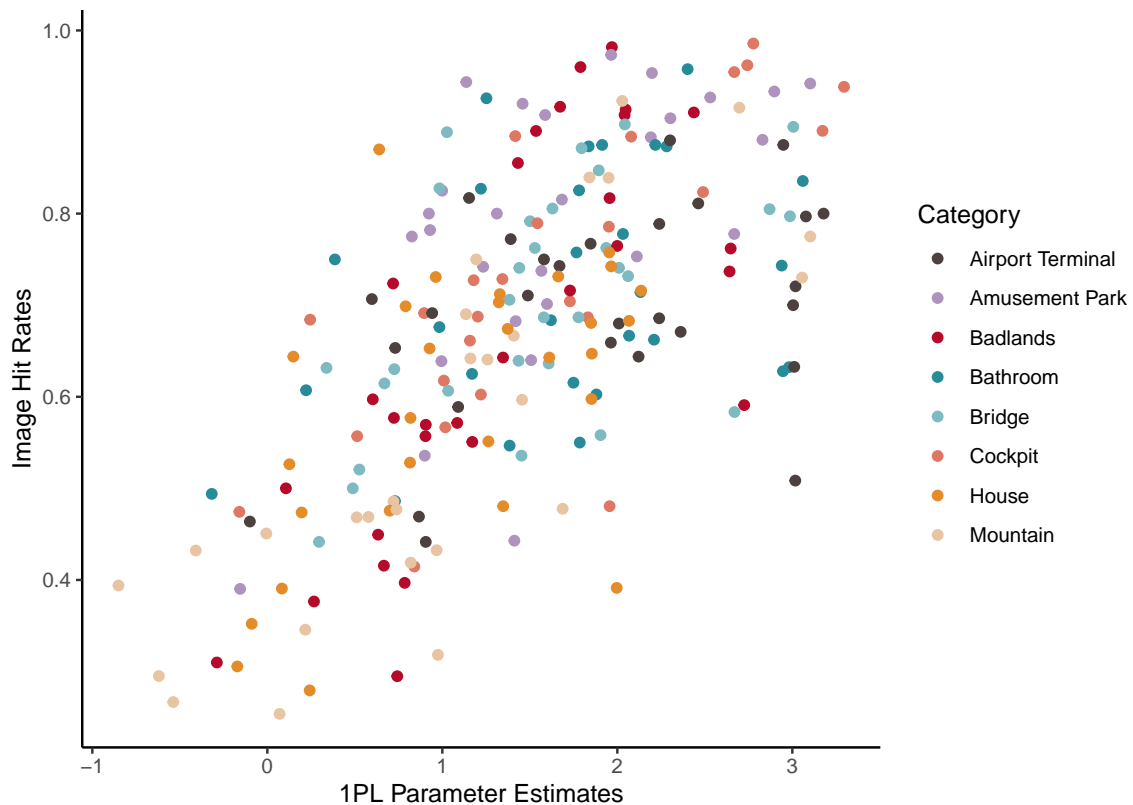
Refitting 1PL Without Large Parameter Images. The 8 images with very large parameter estimates are removed from the data set and the 1PL model is refit with the remaining items. The new model has nearly identical parameter estimates to the previous one and a Brier score of 0.145. This increase in misclassification rate is probably because the removed items are the easiest to predict. However, the reduction in items might be worth the slight loss in prediction. Figure 3 shows the refitted parameter estimates plotted against image hit rate.

Many of the images have parameter estimates that are close to the same values, suggesting that these images are measuring people similarly. This model still contains 228 images, however, selecting images to retain is now less apparent. One method of selection would be removing images that are inconsistent with their hit rates, such as those with high parameter estimates but lower hit rates. Alternatively, item information can be used to thin out images that may overlap in information with each other. Item information quantifies the amount of information an item contributes to locating a person's position on the latent trait (Birnbaum, 1968). Equation 16 shows the item information function for a 1PL model, where $a = 1$ for a Rasch model and $P_i(\theta)$ is the probability of a correct response for item i .

$$I_i(\theta) = a^2 P_i(\theta)(1 - P_i(\theta)) \quad (16)$$

Item information plots for the refit 1PL model are found in Figure 4. These display the range of theta that each item provides the most information for. Many of the images within each category overlap with each other, providing redundant information for various regions of the latent trait. Images that overlap the most can

Figure 3. The scatter plot presents item parameter estimates from the 1PL model without items with large parameter estimates. Removing these items has little effect on other item estimates.

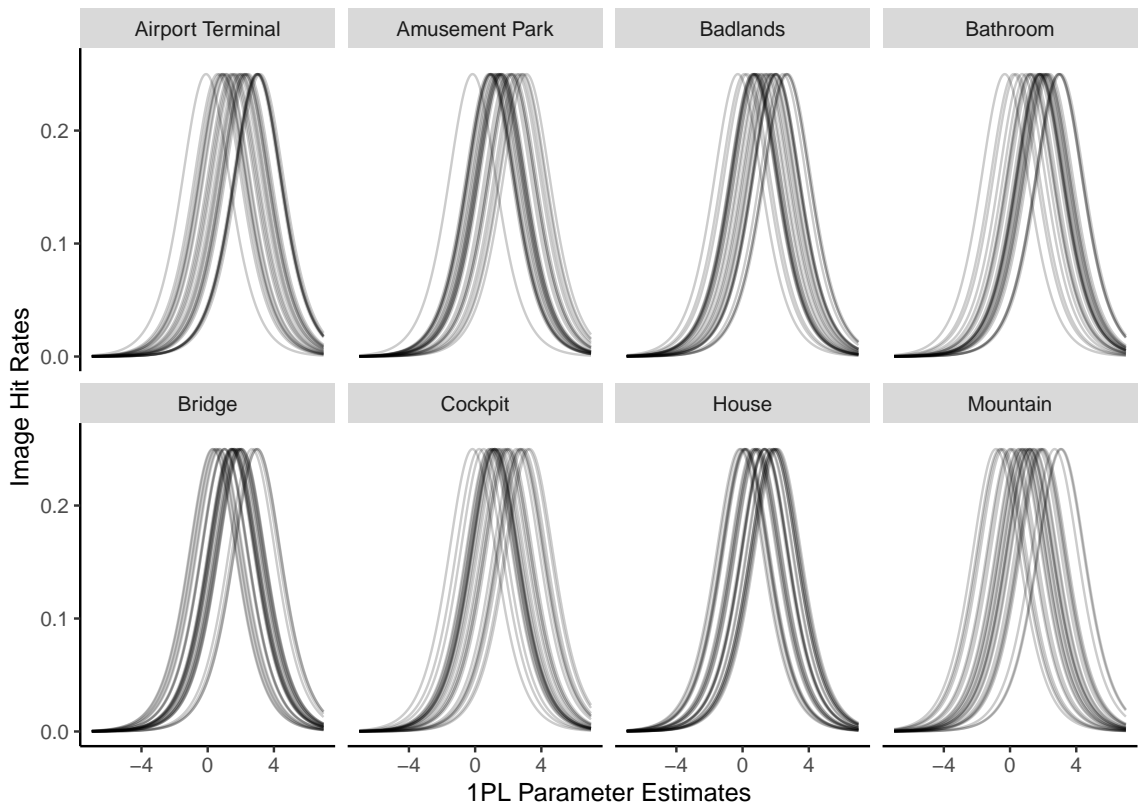


be selectively removed. However, there are still many images in the model and it is worthwhile to thin out every other image in each of the categories.

Fitting 1PL with Half the Images. For the next model, the number of items are cut in half. This is done by first ordering the images by category and then by parameter estimates from the previous model. Then, every other item is removed from the model. Once again, this model yields parameter estimates very similar to the previous ones and only increases the Brier score to 0.147, which is still well under using hit rate alone.

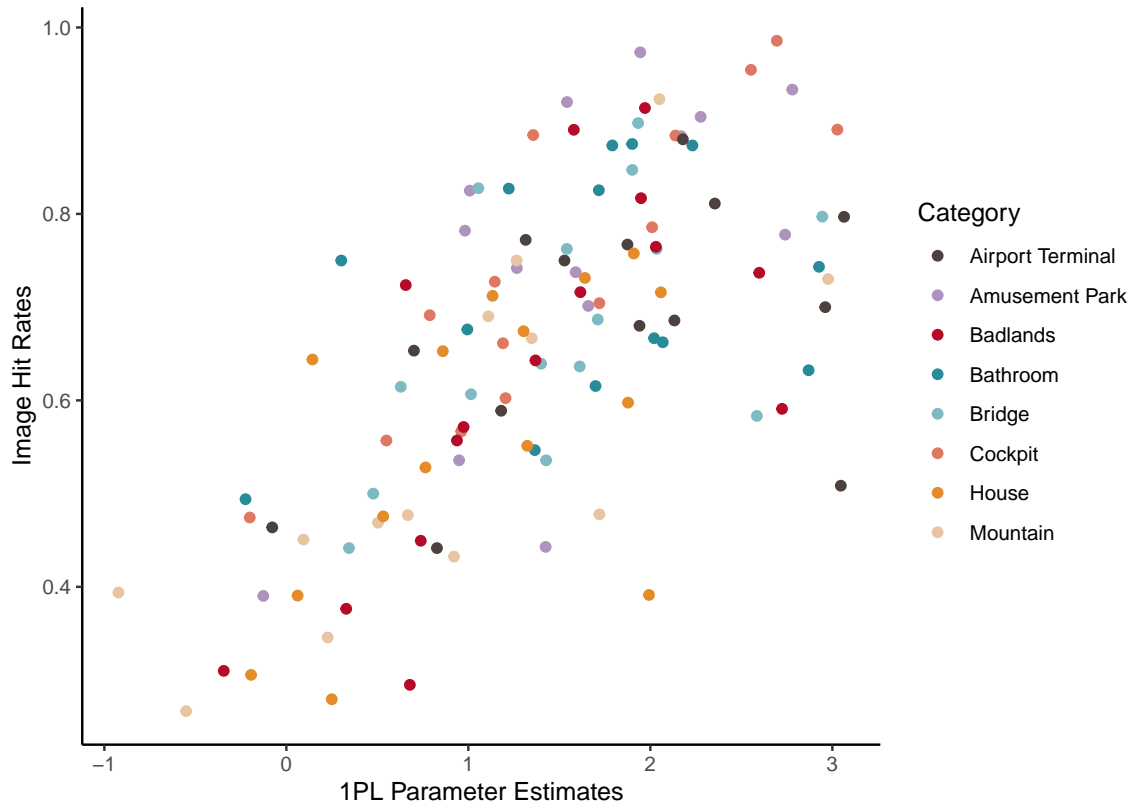
The parameter estimates for the 115 items in this model are plotted against their

Figure 4. Item information plots by image category. The curves shows the range of theta values that each item provides the most information for. Many of these curves overlap quite a bit showing that some items are redundant in the range of theta the best predict.



hit rates in Figure 5. This model keeps a similar overall trend as the others, but with fewer items. However, because they weren't specifically removed, there are still a few items that are inconsistent in their hit rates and parameter estimates. Item information for the model is found in Figure 6. The items cover a similar range as the previous models, but are not as redundant in the information they provide. A few of the categories, like bathroom and bridge, do have items that are closely clustered together and could possibly use with some more reduction if wanted.

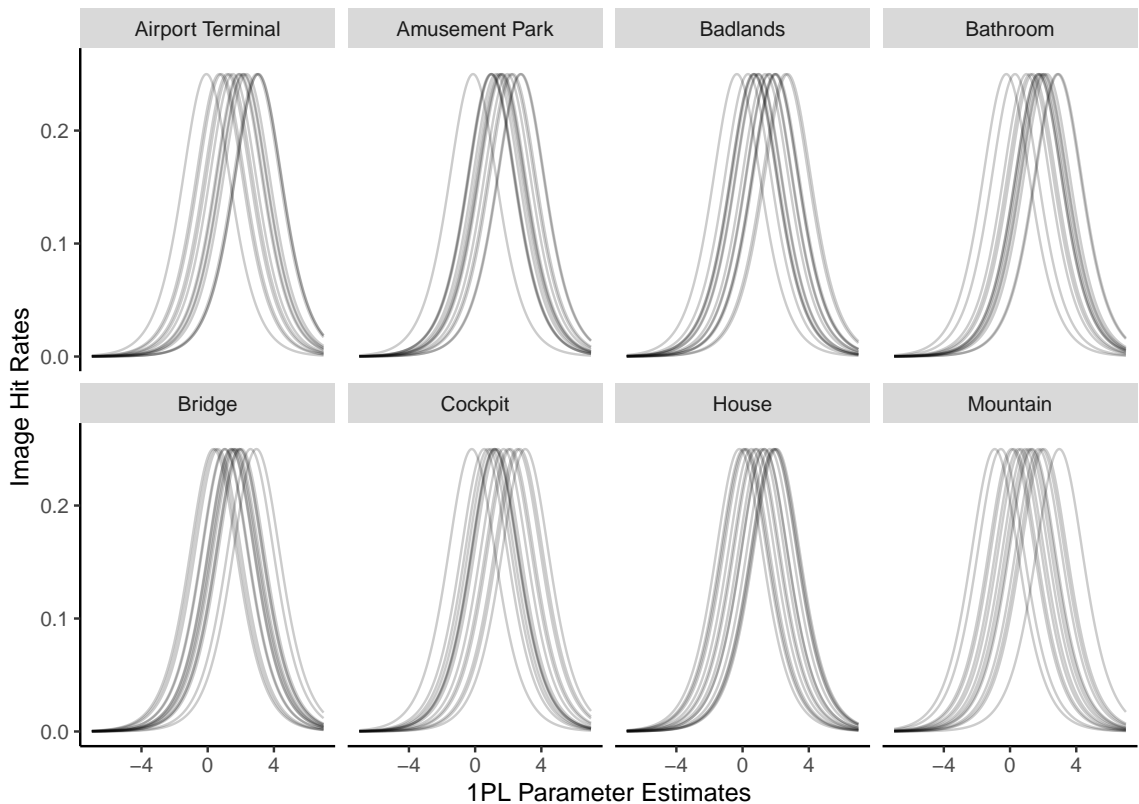
Figure 5. The scatter plot presents item parameter estimates from the 1PL model with about half the number of items.



Adding Item Covariates

After removing the items with large parameter values and thinning, additional item or person level covariates may be included in the model. One additional variable of interest within this data set is fixation count. Typically, when a person is familiar with something they do not require as many fixations to process the image as they would for something that is unfamiliar. So, an increase in the number of fixations could be indicative of an incorrect response. To help with estimation, fixation count is first centered and then added in to the model below. The images were also presented multiple times during the study, so data from the second recall phase is added into the analysis along with a variable for the presentation time.

Figure 6. Item information for the model with half the number of items. The remaining items still cover a similar range of theta values as the larger model, but is able to do so with fewer items. There are still a few dense areas of some categories where more items could potentially be removed.

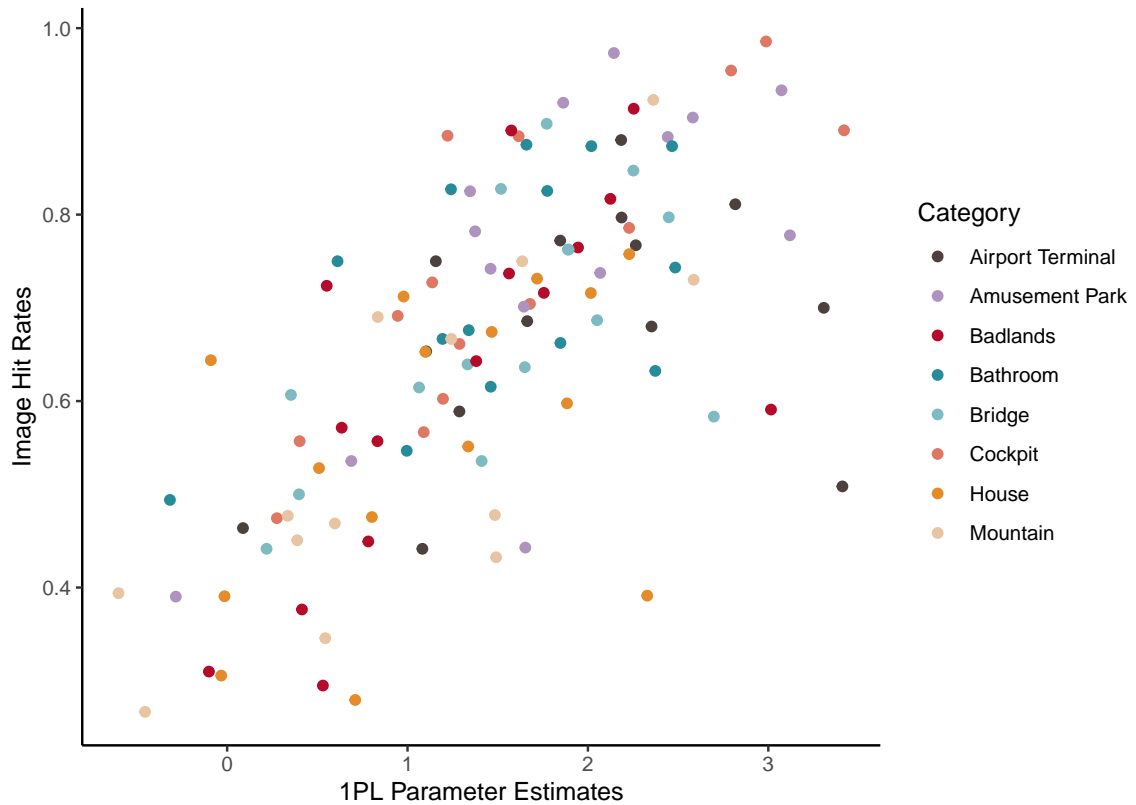


```
modex2 <- glmer(BinRes ~ -1 + Image + FixCent + Pres + (1 | ParID),
  data = matching,
  family = 'binomial',
  glmerControl(optimizer = 'bobyqa'))
```

Centered fixation count has a parameter estimate of -0.111 ($SE = 0.06$), which is small but does imply that the probability of correctly recalling an image decreases as the number of fixations increases. The images are also more readily remembered in the second recall phase than the first ($b = 1.587$, $SE = 0.12$). There are slight

changes in item parameter estimates with the addition of fixation count but only by hundredths of a point or less. The inclusion of the additional data in the model decreases the Brier score to 0.104.

Figure 7. Comparison of parameter estimates from 1PL model with halved items and additional covariates and the images corresponding hit rates.



Alternative Response Variables and Link Functions

GLMMS provide added flexibility to the choice of outcome variable and link function used in model estimation. Using alternative link functions to the logit means that items can be specifically tailored to the researcher's preferred outcome of interest. For example, the researcher might want a range of items that increase variability in the number of fixations participants make or for the amount of time participants are

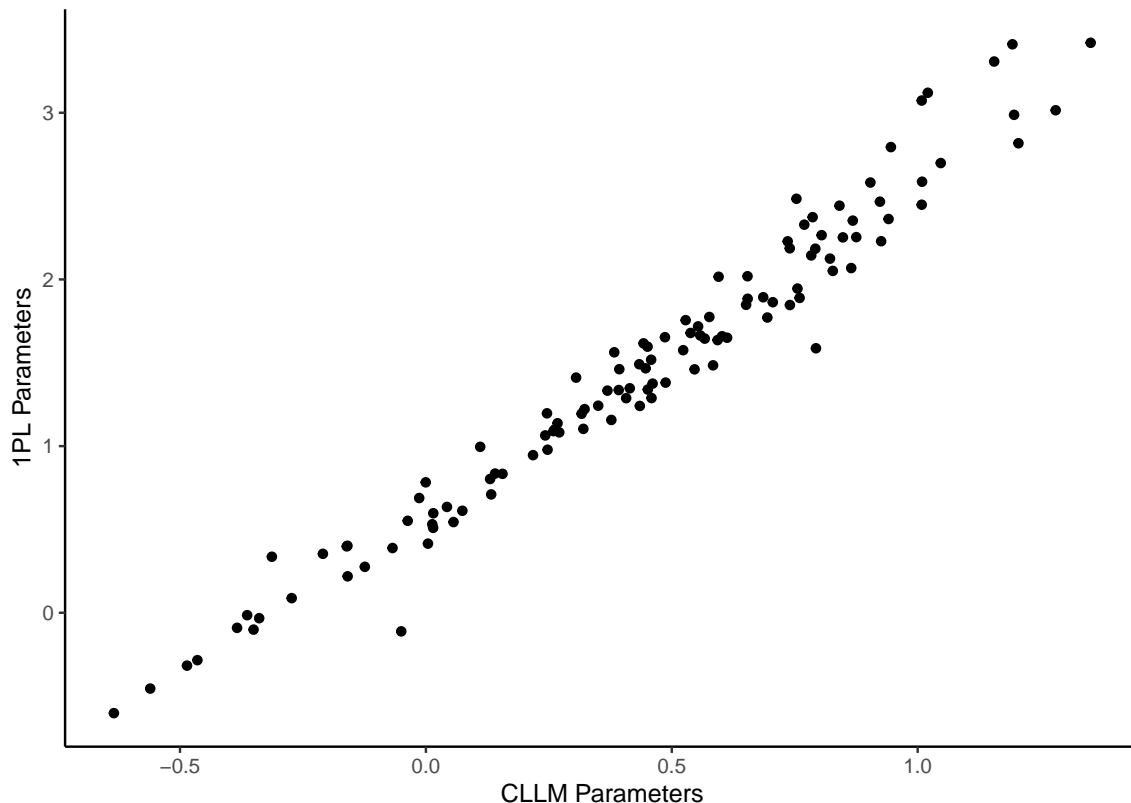
spending at each fixation. Alternatively, changing the link function might allow for added model functionality as is the case for the complementary log-log link. In this section I explore the use of CLLMs and Poisson models as alternatives to the 1PL model.

Complementary Log-Log Model. The complementary log-log link function can be used in place of the logit link function to approximate the information provided by the 3PL model without increasing the number of parameters that need to be estimated. This link function allows for asymmetric response curves instead of the symmetric ones assumed in a traditional 1PL model. It is useful if the researcher wants to account for participants potentially guessing the correct response for the task instead of knowing it for certain. Since this model is less complex than the 3PL, it works better with smaller sample sizes that are typically found in eye tracking studies. The CLL link can be easily specified using **lme4**, via the same code syntax as the previous models. The only change is specifying that `family = binomial(link = 'cloglog')`, as can be seen in the code below.

```
mod.CLLM <- glmer(BinRes ~ -1 + Image + FixCent + Pres + (1 | ParID),
                 data = matching4,
                 family = binomial(link = 'cloglog'))
```

Fitting this model to the halved item set and the other two predictors of interest results in parameter estimates that are similar in relation to those of the 1PL model as seen in Figure 8. They also have a similar relationship to hit rates as the previous models did (Figure 9). Compared to the 1PL model with the added covariates, this model predicts almost as well and provides the flexibility of being a less restrictive model. The Brier score for the CLLM is 0.105. Items are also less likely to be remembered as fixation count increase but they are more readily remembered at the second presentation.

Figure 8. Comparison of parameter estimates from 1PL model and CLLM using the halved item set and other covariates.

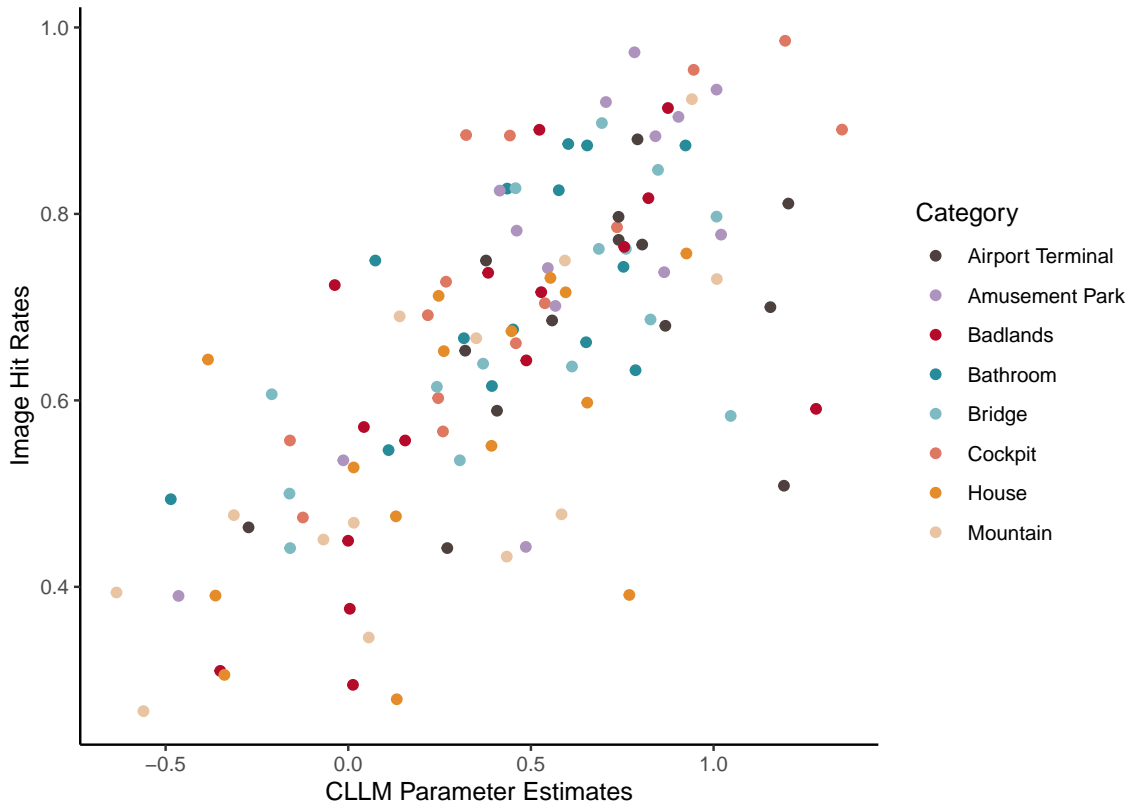


Item information for the CLLM can be obtained using the formula in Equation 17. The information for these items have a slightly different distribution than those from the 1PL model due to the asymmetrical nature of the CLLM. The information range of the images in the CLLM is more bounded than in the 1PL model, mostly in the upper range of theta. Figure 10 presents the plots of the information functions. Items are most informative from -2.5 to 2.5 and provide very little information in the tails.

$$I_i(\theta) = \left[\frac{1 - P_i(\theta)}{P_i(\theta)} \right] [\log(1 - P_i(\theta))]^2 \quad (17)$$

Poisson Model. Eye tracking response data can be of any type, from binary to continuous data. When the outcome of interest changes, the link function for the

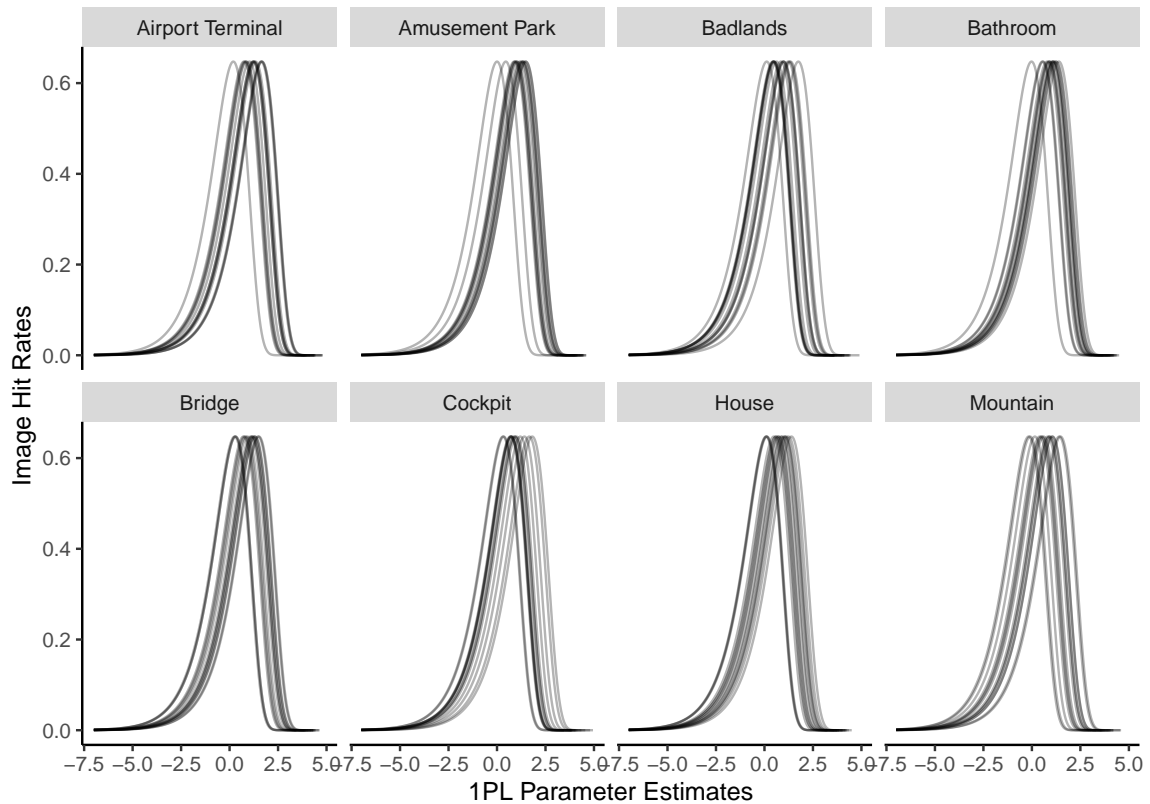
Figure 9. Comparison of parameter estimates from CLLM model with halved items and additional covariates and the hit rates for corresponding images.



measurement models can be changed to appropriately account for this. Items can then be assessed based on this alternative outcome instead of an often-forced dichotomous or continuous outcome. This allows for item selection that is specifically tailored to the outcome of interest, and use of an outcome that may be more directly related to the actual visual or attention driven process.

For example, in the case of image memorability, fewer fixations on an image often mean that the image is more memorable. The image is able to be processed quickly and holistically, leading to a correct recall. Counter to this, a novel or less memorable image might take more fixations and processing time. Applying this to the previous models, the researcher could select a range of images that induce various

Figure 10. Item information for images in CLLM for each category.



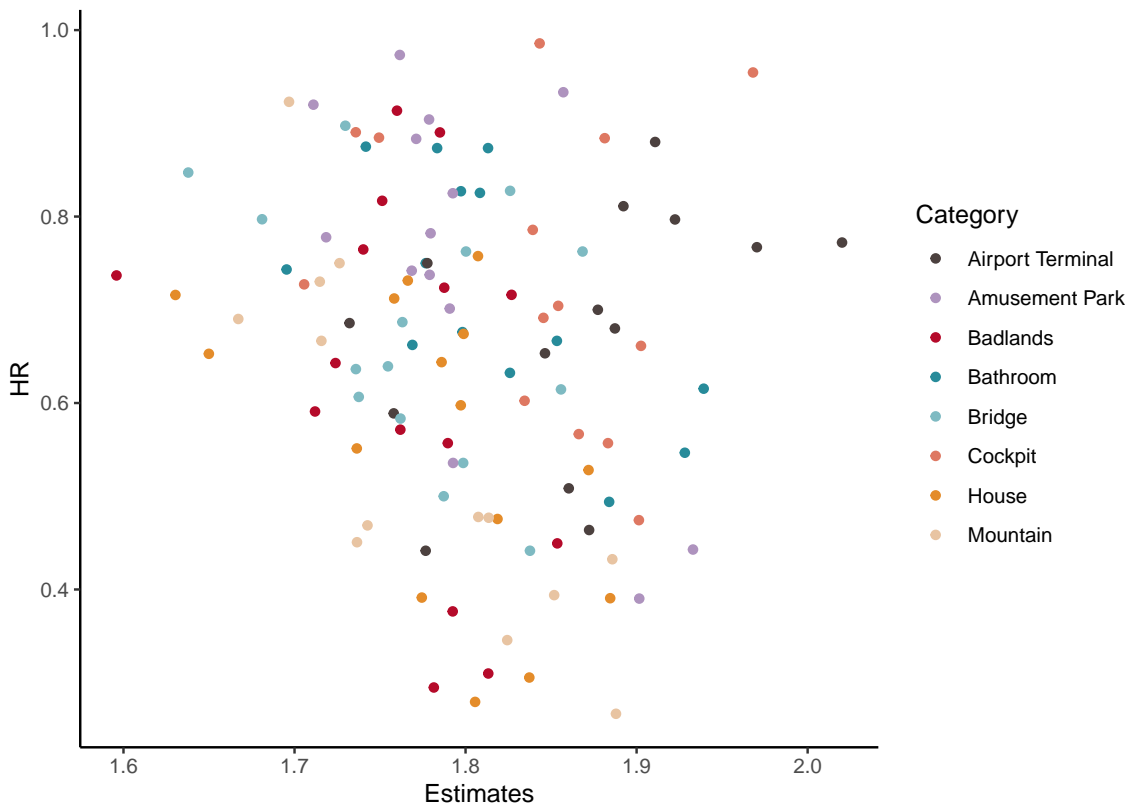
amounts of scanning behavior. This can be used as an indication of image complexity instead of relying on the participant's response to a memory recall task.

An example of this model is presented in the code below. It follows a similar format as the previous one, but changes the outcome variable to be total fixation count for the images and the distributional family to be Poisson. Similarly, other item covariates of interest could be added to the model. Here, the variable for recall presentation time has been added.

```
poimod <- glmer(Fixation ~ Item + Pres + (1 | Person),
               data = data,
               family = 'poisson')
```

Figure 11 plots the item parameters from the Poisson model against the hit rates. There is a slight negative relationship between the estimates and the hit rates, $r = -0.2$. However, inspection of the plot reveals that the airport terminal category might not follow the same trend as the others. In fact, it is quite the opposite with a positive correlation of $r = 0.47$. Removing the airport terminal items from the correlation increase it to $r = -0.31$. So, in general, the ability to recall items is related to the average number of fixations for the item. Further, the more fixations there are, the less likely the image is to be remembered.

Figure 11. Scatter plot of item parameter estimates from Poisson regression predicting fixation count plotted again item hit rates. There is a slight negative correlation between parameter estimates and hit rate for the items. However, there are a few images in the airport category that break this trend.



In this chapter, various item response models from a GLMM framework are applied to real world eye tracking data to illustrate their ability to aid in item assessment and reduction. First, a large subset of 236 items from [Bylinskii et al. \(2015\)](#) were fit to a 1PL equivalent model and were reduced down to 115 by first removing items that provide no benefit in discriminating between participants and then by using item information plots to remove items that overlap in information they provide. The reduced item set is then used to fit models with additional item covariates, such as image category and fixation count, and to models with alternate link functions like complementary log-log and Poisson. This illustrates the flexibility of these models to fit the variety of data types that comes with eye tracking data. In the next chapter, a simulation is conducted to assess how model performance holds across a decreasing number of items and with various sample sizes.

Chapter IV

Simulation

The primary goal of this project is to assess whether fewer images can be used in eye tracking studies while obtaining similar results. In this chapter, I conduct a series of simulations to investigate the stability of model performance when image numbers are decreased. In the first simulation, stability is assessed with decreasing the number of images, varying sample size, and image thinning method. The second and third simulations are designed to further investigate results from the first simulation by testing models that predict a single value, changing the data simulation process, and testing more image thinning methods. The final simulation then assess how the models perform under misspecification. In the sections below, I will first discuss the simulation designs and then the results of the simulations.

Simulation 1

The first simulation investigates the performance of the item response models discussed in the application section with a decreasing number of images. If the models perform similarly, then using a less complex model would be more desirable both theoretically and practically. It is also important to consider how the images are selected for each successive model and how sample size effects results.

Design

The data are simulated using the fitted half-item with covariates 1PL model and Poisson model from the application section. These models consist of 115 images. Both models use recall phase as a covariate and the 1PL model contains the centered fixation variable as a covariate. First, a design matrix is created with N number of people fully crossed with all items for both recall phases. Then, using the `simulate()`

function, fixation counts are generated from the Poisson model. The simulated fixation counts are then added to the design matrix as raw fixation counts and as a scaled variable. The binary response variable is simulated using the same data matrix that the fixation counts were but with the inclusion of the centered fixation count variable. This is further referred to as the “all” item data set. Finally, the larger data set is broken down into two “half” size and two “quarter” size data sets.

The items in these subsets are thinned out from the “all” item data set in two different ways. The first thinning method orders the items by their original parameter estimates within each of their image categories and then retaining every other item for the halved data set. This is repeated for the quarter item data set. The other data sets are created by randomly selecting an equivalent number of items from the full item pool. For this simulation, data sets were generated with 125 and 250 people. The all item data set consists of 115 items, the halved data sets have 59 items, and the quartered data sets have 32 items.

Before models are fit to the data, each of the data sets have 20% of their total observations (people x items) randomly sampled to create a testing data set and the other 80% of observations are used as the training data. This is done to calculate test Brier Scores and MSE values to allow for a comparison between the models. Traditional comparison methods such as AIC, BIC, or Likelihood Ratio tests are unable to be appropriately used because changing the number of items in the model changes the number of overall observations in the data set when it is in long format. These metrics are not designed to make comparisons between models that are fit to “different” data sets. Randomly selecting these observations also creates a training data set that is partially crossed, so not every person necessarily saw every item at each presentation point.

After the data are separated into test and training sets, they are used to fit 15 models (3 link functions by 5 data sets). First, using the all item data set, a 1PL and

CLLM are fit with the binary correct/incorrect response as the outcome and a Poisson model is fit using the raw fixation count as the outcome variable. Using these models, new predicted outcome values are generated using the test data set. Brier scores are then calculated for the 1PL and CLL models and MSE is calculated for the Poisson model. This process is then repeated for the thinned half item, random half item, thinned quarter item, and random quarter item data sets. The full simulation consists of 100 replications for each of the two sample sizes. Brier scores and MSE are recorded and used to compare the precision of the models as the number of items decrease. Warnings for each model are also documented in case of model convergence or other issues.

Results

Predictive ability for each of the models is compared using the means and standard deviations of the selected loss functions, which are calculated across the 100 replications for every possible condition. Ideally, there will be a minimal loss of prediction as the number of images decrease across the models. Means and standard deviations for the simulation with a total of 125 people are in Table 3 and the ones for the simulation with a total of 250 people are in Table 4. The tables are broken down into two sections. The top section presents the results based on all of the replications while the bottom section presents the results when models that had convergence issues are removed. For the models that used 125 people, there were no issues with the Poisson models, 3 or fewer replications with issues for the 1PL models, and 5 or less replications with issues for the 115 image and 32 image CLL models. Both the thinned and randomly chosen half item CLL models struggled with estimation and around half have convergence issues. This was also the case for the 59 image CLL models using 250 people. The other models at this sample size did not have any issues.

Replications	Image Subsets	1PL	CLLM	Poisson
All Reps				
	115	0.106(0.0028)	0.1065(0.0027)	5.907(0.1084)
	Categorized 59	0.1082(0.004)	0.1087(0.0039)	5.9053(0.1617)
	Random 59	0.107(0.0059)	0.1075(0.0059)	5.9321(0.1786)
	Categorized 32	0.1159(0.0046)	0.1167(0.0043)	6.0687(0.2276)
	Random 32	0.1073(0.0086)	0.1078(0.0085)	5.9651(0.2585)
Converged Reps				
	115	0.106(0.0028)	0.1064(0.0026)	5.907(0.1084)
	Categorized 59	0.1082(0.004)	0.1093(0.0036)	5.9053(0.1617)
	Random 59	0.1069(0.0059)	0.1068(0.0058)	5.9321(0.1786)
	Categorized 32	0.116(0.0046)	0.1168(0.0043)	6.0687(0.2276)
	Random 32	0.1073(0.0086)	0.1078(0.0085)	5.9651(0.2585)

Table 3

Means and standard deviations of the 1PL and CLLM Brier scores and the Poisson model MSE for the simulation condition with 125 observations. The Brier scores and MSE are calculated from a hold out sample of 20% of the total number of observations.

Replications	Image Subsets	1PL	CLLM	Poisson
All Reps				
	115	0.1054(0.0018)	0.1058(0.0018)	5.8709(0.0816)
	Categorized 59	0.1074(0.0028)	0.1079(0.0027)	5.9011(0.1088)
	Random 59	0.1047(0.0055)	0.1052(0.0055)	5.8889(0.1164)
	Categorized 32	0.1156(0.0038)	0.1162(0.0037)	6.0041(0.1492)
	Random 32	0.1062(0.0084)	0.1066(0.0085)	5.9271(0.1656)
Converged Reps				
	115	0.1054(0.0018)	0.1058(0.0018)	5.8709(0.0816)
	Categorized 59	0.1074(0.0028)	0.1077(0.0025)	5.9011(0.1088)
	Random 59	0.1047(0.0055)	0.1046(0.0056)	5.8889(0.1164)
	Categorized 32	0.1156(0.0038)	0.1162(0.0037)	6.0041(0.1492)
	Random 32	0.1062(0.0084)	0.1066(0.0085)	5.9271(0.1656)

Table 4

Means and standard deviations of the 1PL and CLLM Brier scores and the Poisson model MSE for the simulation condition with 250 observations. The Brier scores and MSE are calculated from a hold out sample of 20% of the total number of observations.

In general, across both sample sizes, models with more images were better at predicting new observations than those with fewer images. However the difference between the metrics are in fractions of points. Also, the random image selection performed slightly better than the thinning method, but was more variable in score values. This difference is small and could be due to randomly selecting more easy images than difficult ones because eye tracking stimuli tend to be on the easier side more often than not. The models performed similarly across each of the sample sizes.

Starting with the 1PL model, Brier scores range between about .105 and .116 discrepancy between the predicted probability correct and actual response for both of

the sample sizes. The model containing all of the images performed the best in the 125 person condition, but the randomly selected, the 59 image 1PL model performed the best in the 250 person condition. The quarter item 1PL model was the worst performing in each of the sample size conditions.

The CLL models performed similarly to the 1PL models and have Brier scores ranging from about .106 to .117. While the models with 59 images had issues with convergence, the average Brier score was not too different when removing the models with issues. Interestingly, the difference between the average scores of the thinned and randomized half image models did increase when removing the non-converged models. In the smaller sample size condition, the thinned model lost some predictive ability while the random one improved. In the larger sample size condition, both of the models improved, but the random item selection improved slightly more.

The Poisson models have the least amount of issues and MSE values range from about 5.9 to 6.1. In the smaller sample size, the model using half the number of images has the lowest MSE and the model with a quarter of the number of images has the highest. For the larger sample size, the model with all of the images has the lowest MSE and the 32 image model had the highest. Once again, the random item selection out performed the thinned items except in the case of the half image model in the 125 observation condition.

One potential concern across all of the models is how small the standard deviations are. These suggest that there is very little variability in Brier scores and MSE across the replications and could be due to an issue with how the data is generated, how the test data is selected, or just due to eye tracking items generally being easier difficulty. These potential issues are assessed in the next section of simulations.

In sum, decreasing the number of items used in model estimation is minimally detrimental to a model's predictive ability, at least in the case of this simulation. Model performance is also similar for the 1PL and CLL models, although the

half-item CLL model has major issues with model convergence. While the cause of these issues are unknown and worth investigation, they could be a product of how the data is generated from the source models or with how the items are pared down.

Simulation 2

Several concerns arose from the first simulation after seeing the results. This section and the following section attempt to address some of these concerns. The models in the previous simulation were all able to predict responses surprisingly well. Eye tracking tasks and associated images can be notoriously “easy”, because people are typically very quick at processing a visual stimulus. This could be the reason the models are so good at predicting responses. This simulation attempts to address this by generating new data sets and then assessing how a test subset performs when all participants are predicted to have the same response or probability of response. Here, two response options are tested. First, a model predicting that every participant gets every item correct is assessed. Then, every participant is predicted to have an 84% chance of getting an item correct. This number is the overall average hit rate of the original data set from [Bylinskii et al. \(2015\)](#).

Design

The data generation portion of this simulation follows the same steps as in simulation 1, but the test data are assessed as if the models predict a constant value. For each of the 100 replications, full data sets consisting of 115 items are generated with sample sizes of 125 and 250 people. Then, 20% of the crossed item by person observations are held out as a test data set. Instead of fitting models for this simulation, the response variable of the test data is just compared to a constant value. In this case, the constants chosen are 1 and .84. The value 1 represents a model that assumes a person will respond to an item correctly 100% of the time. The value of .84

is the overall hit rate from the original data set. It assumes a person will be correct 84% of the time.

Results

Total Sample Size	Image Set	Predicted Value = 1	Predicted Value = .84
125	115 Images	0.1604(0.0107)	0.1347(0.0073)
	Categorized 59 Images	0.1662(0.0114)	0.1386(0.0077)
	Random 59 Images	0.161(0.0145)	0.1351(0.0099)
	Categorized 32 Images	0.1857(0.0141)	0.1518(0.0096)
	Random 32 Images	0.1593(0.0182)	0.1339(0.0124)
250	115 Images	0.1604(0.0068)	0.1347(0.0046)
	Categorized 59 Images	0.1661(0.0073)	0.1386(0.005)
	Random 59 Images	0.1619(0.0121)	0.1357(0.0082)
	Categorized 32 Images	0.1851(0.0079)	0.1515(0.0054)
	Random 32 Images	0.1615(0.0157)	0.1354(0.0107)

Table 5

Means and standard deviations of the Brier scores for decreasing image test sets being compared to a model that predicts a constant value. The first column is the initial number of people in the data set, the second column is the set of images the constants are compared to, and the other columns are the average Brier scores for a model that always predicts a response will be correct and one that always predicts a 84% chance of being correct.

The average Brier scores and associated standard deviations for simulation 2 are presented in Table 5. Brier scores are similar across the sample sizes for each of the constants. This is also the case when decreasing the number of images for all data

sets except for the thinned quarter image sets. For this set, the average Brier score is about .02 larger than its counter parts. A similar pattern is found in simulation 1 but is not as large of a difference. Thinning items in this way always keeps the most difficult items, and if the model is predicting an image will always be remembered correctly, it is expected that the Brier scores will be larger.

Brier scores are overall larger for simulation 2 than they are for simulation 1. Fitting the data to a model, be it a 1PL or CLLM, has overall better predictive ability than just using the average hit rate or assuming that an image will always be correct. So, while the images are still on the easy side, the models are better to use and probably worth the added complexity.

Simulation 3

The third set of simulations are designed to address potential concerns from the first simulation. The primary concern this simulation is designed to address, is the size of the standard deviations of the loss functions. They are very small considering the number of replications, suggesting that the Brier scores and MSE for all the reps are very similar to each other. This could just be caused by chance or it could be caused by how the data are generated. This section of the chapter addresses these concerns by tweaking the data generation method, changing test data sampling method, and testing a new item thinning method.

Design

The overall simulation design and procedures for this set of simulations is similar to the one above with a few changes to how the data are generated, the number of items used in the models, and adds additional models to test. The first change is with how the **lme4** `simulate()` function generates the outcomes for the data sets. The first simulation used the argument `use.u = TRUE`, which conditions on the random effects and is similar to saying the test data come from the same people as the

training data, making prediction easier than it otherwise would be. This was removed for the current simulation and replaced with `re.form = NA` to better approximate the random effect from the original data set and get more variability in prediction.

In order to help decrease the amount of time it takes to conduct a full simulation, the starting number of items in this simulation are halved. Instead of using the full item list, like in the simulation above, the data generating process starts with using the thinned, halved-item data set. This item set is then halved into three new data sets. This includes thinning by grouping items into image categories by parameter size and selecting every other one, randomly selecting items, and a new thinning method of ordering all the items by their parameter values, regardless of their image category, and then removing every other item. Thinning this way yields one or two fewer items than when using categories.

Finally, the sample size and test sample selection are modified for this simulation. There are still two sample size conditions but they consist of a total of 125 and 625 observations. The number of observations for the larger sample size was increased in an attempt to reduce model convergence issues. Test sets are now made from the last 20% of the “participants” (i.e. in the 125 person condition, persons 101-125 are considered the test sample), making the training data fully crossed now and the test data consist of totally novel observations.

Results

Table 6 presents the results for simulation three. Once again, average Brier scores are presented for the 1PL and CLL models while MSE is used for the Poisson models. Compared to the results in simulation 1, models do not perform quite as well for directly comparable item subsets but show a similar trend of not predicting as well as the number of items decreases. However, this difference is very small. The CLLMs with 59 items also had issues with convergence with only 17 of the 100 replications

converging in the 125 observations condition and 6 in the 625 observations condition. Other than that, only three other replications had convergence issues in other models in the smaller sample size and one other replication had an issue in a model in the larger sample size. Changing the data generating method did increase variability in the data sets, but only very slightly.

Sample Size	Image Subset	1PL	CLLM	Poisson
125	59	0.1175(0.0092)	0.1175(0.0091)	6.8017(0.469)
	Categorized 32	0.122(0.0091)	0.122(0.009)	6.7586(0.4837)
	Not Categorized 30	0.1181(0.0093)	0.1181(0.0092)	6.7469(0.4933)
	Random 32	0.1186(0.0114)	0.1186(0.0113)	6.8043(0.5023)
	Categorized 16	0.1497(0.01)	0.1497(0.0097)	7.0867(0.5875)
	Not Categorized 15	0.1212(0.0108)	0.1212(0.0107)	6.6898(0.5119)
	Random 16	0.1193(0.0166)	0.1192(0.0165)	6.8012(0.5136)
625	59	0.1177(0.0051)	0.1177(0.005)	6.8521(0.2061)
	Categorized 32	0.1223(0.0047)	0.1224(0.0046)	6.8161(0.2138)
	Not Categorized 30	0.1177(0.0054)	0.1177(0.0053)	6.796(0.2299)
	Random 32	0.1167(0.0076)	0.1167(0.0076)	6.8455(0.242)
	Categorized 16	0.1502(0.0053)	0.1502(0.0051)	7.118(0.2496)
	Not Categorized 15	0.1215(0.0055)	0.1216(0.0054)	6.7474(0.2642)
	Random 16	0.1185(0.0139)	0.1185(0.0139)	6.8387(0.285)

Table 6

Means and standard deviations from the revised simulation 3. The table is broken down by overall generated sample size, image thinning method, and model. The columns for the 1PL model and CLLM is the average Brier score across the 100 replications and the Poisson column is the average MSE across all reps.

Within the simulation itself, the average loss function values are similar between the sample sizes but the standard deviations are slightly larger for the smaller sample size than for the larger one. Brier scores are also nearly identical for the 1PL models and the CLLMs across all image subsets and usually start to differ at around the fifth decimal space. Prediction is generally better with more items for the 1PL models and CLLMs. The Poisson models a bit more variable and some of the models with fewer items are better at predicting on average than the model with all 59 items. The Poisson models using the thinned, not categorized images performed the best out of the other image categories. Finally, within the image subsets, the thinned, categorized subsets performed worse than the thinned, not categorized subsets and the randomly selected image subsets. Once again, this is probably because more difficult items are kept using this method than they are with the others. Randomly selecting items was also just the smallest bit better than the no category thinning in most of the 1PL and CLLM situations, but not for the Poisson models.

Simulation 4

The final simulation is conducted to assess model misspecification. Here, data are generated from a 2PL and then fit to the GLMMs from the previous simulations and chapter 3 which approximate a Rasch model. Similar to the previous simulations, item numbers are decreased for successive models and compared against each other. These reduced data sets are also fit to a 2PL model and compared to their corresponding GLMM.

Design

Once again, this simulation follows the same general procedures as the previous ones, specifically simulation 3, with the major differences being how the data are generated, removal of item specific covariates, and the addition of 2PL model estimation. In order to get an idea of what 2PL parameter estimates should look like,

a 2PL model is fit to a subset of the 59 images used in the previous simulation from original data set using **mirt** (Chalmers, 2012). Responses were only taken from the first recall phase and no covariates, such as fixation count or recall phase, were added to the model. Many of the items had large and unrealistic parameter estimates. These items were removed from consideration of an appropriate parameter range if $d > \pm 4$, which indicates an item is always correct or incorrect. Parameter values were also removed from consideration if a was negative or greater than four, indicating that the items were more likely to be correct for lower levels of the latent trait or were just exceedingly large. Most of the time the removed item met both of these conditions for parameters a and d . Using the remaining 24 item parameters, a distribution of potential values for 2PL parameters are generated for creating the simulated data sets. Parameter values for a are generated from a normal distribution truncated at zero with $m = .844$ and $sd = .68$, while d is drawn from a normal distribution with $m = 1.23$ and $sd = .83$.

To simulate the data, item parameters, a and d , are first generated for each of the 59 items. These vectors and the number of participants are passed into the `simdata()` function from **mirt**. For this simulation, total samples sizes of 125 and 625 are once again used. The returned set of response patterns are converted to long format and combined with the parameter values, participant IDs, and item number. This data set with all the items then goes through the various thinning processes that are mentioned in simulation 3 to create half and quarter item data sets. Item parameters are generated randomly and category isn't taken into account, so items for the first reduced sets are thinned by ordering all of them by parameter value and taking every other. For the other thinned sets, items are just randomly selected. The data sets are further broken down into training and test data sets consisting of the last 20% of the participant ID numbers (i.e. participants 1-100 are the training data and 101-125 are the test data).

This process is repeated for each of the sample sizes for 125 replications. The number of replications are over-sampled in case any data sets are generated where items have a perfect correct or incorrect response rate. These data sets are removed from the pool because **mirt** cannot estimate models with items that every participant gets correct or incorrect. If any data sets are leftover beyond 100 after this, the last sets of replications are removed so that there are 100 replications for each of the sample sizes.

After the data sets are generated, each of the data sets are fit to a 1PL, CLLM, and 2PL model to assess model performance. Brier scores are once again calculated to compare the accuracy of the 1PL and CLLMs as the number of items decrease. In addition to Brier scores, AIC values are also calculated to compare the 1PL and CLLMs to the 2PL models within each of the item set categories. AIC is used in favor of BIC, because the penalty term in BIC requires a sample size which is not clearly defined for MLMs due to the clustered nature of the data (McCoach & Black, 2011; Hamaker, van Hattum, Kuiper, & Hoijtink, 2011; Lorah & Womack, 2019; Cho, Wu, & Naveiras, 2023).

Results

Overall, the 1PL and CLLM did not perform as well when the data are generated from a 2PL model. Even the constant value models from simulation 2 were better at predictions. Table 7 presents the average Brier scores and standard deviations for the 1PL and CLL models across the decreasing number of images. Within the simulation itself, Brier scores across all conditions are very similar to each other. They all hover between .18 to .19 and the only thing the increase in sample size does, is to make the standard deviations slightly smaller. The CLLMs with all 59 items and about one fourth of the items did have some convergence issues, but in this case there were less than ten replications with issues in each of the conditions.

Sample Size	Image Subset	1PL	CLLM
125	59	0.186(0.0153)	0.1859(0.0149)
	Not Categorized 30	0.1852(0.0163)	0.1851(0.016)
	Random 30	0.1863(0.0174)	0.1863(0.0172)
	Not Categorized 15	0.1876(0.0179)	0.1877(0.0177)
	Random 15	0.1838(0.0211)	0.1837(0.0209)
625	59	0.1878(0.0099)	0.1875(0.0098)
	Not Categorized 30	0.1869(0.0098)	0.1867(0.0097)
	Random 30	0.1893(0.0123)	0.189(0.0121)
	Not Categorized 15	0.1899(0.0109)	0.1897(0.0108)
	Random 15	0.1867(0.0169)	0.1866(0.0167)

Table 7

Means and standard deviations of Brier scores from simulation 4. The table is broken down by overall generated sample size, image thinning method, and model. Brier scores are much higher than in the previous 1PL and CLLM simulations.

Sample Size	Image Set	1PL AIC	CLLM AIC	2PL AIC
125	59	6080.44(221.3)	6114.47(219.09)	5970(220.86)
	Not Categorized 30	3124.01(122.84)	3140.05(120.55)	3079.47(124.62)
	Random 30	3146.21(163.77)	3161.66(162.09)	3097.84(160.82)
	Not Categorized 15	1610.95(66.6)	1617.36(66)	1595.91(66.6)
	Random 15	1589.15(95.57)	1596.09(94.83)	1574.49(94.49)
625	59	30077.73(889.01)	30251.21(877.95)	29307.76(851.67)
	Not Categorized 30	15465.81(459.75)	15546(453.5)	15143.71(442.91)
	Random 30	15610.77(634.17)	15688.51(625)	15269.19(614.66)
	Not Categorized 15	7985.9(238.64)	8019.2(233.66)	7859.54(238.36)
	Random 15	7837.79(439.04)	7875.48(434.56)	7699.23(420.05)

Table 8

Means and standard deviations for AIC values from simulation 4.

Sample Size	Item Subset	2PL < 1PL	2PL < CLLM	1PL < CLLM
125	59	1.00	1.00	1.00
	Not Categorized 30	0.98	1.00	0.95
	Random 30	1.00	1.00	0.94
	Not Categorized 15	0.87	0.95	0.88
	Random 15	0.84	0.90	0.92
625	59	1.00	1.00	0.99
	Not Categorized 30	1.00	1.00	0.96
	Random 30	1.00	1.00	0.97
	Not Categorized 15	1.00	1.00	0.96
	Random 15	1.00	1.00	0.94

Table 9

Proportion of replications that follow the trend of average AIC values from simulation

4.

Sample Size	Image Set	1PL - 2PL	CLLM - 2PL	CLLM - 1PL
125	59	110.44	144.47	34.03
	Not Categorized 30	44.54	60.57	16.03
	Random 30	48.37	63.82	15.45
	Not Categorized 15	15.04	21.45	6.41
	Random 15	14.65	21.60	6.95
625	59	769.97	943.45	173.48
	Not Categorized 30	322.10	402.28	80.18
	Random 30	341.58	419.32	77.74
	Not Categorized 15	126.36	159.66	33.30
	Random 15	138.56	176.25	37.69

Table 10

Average difference in AIC values from simulation 4.

Table 8 presents the average AIC value for each of the item conditions and all of the models, including the 2PL, for comparison. Within each of the item sets, 2PL has the best fit followed by the 1PL models, and finally the CLLMs. This pattern makes sense because the data are generated from a 2PL model and the asymmetry provided by the CLLM is an unnecessary trait for the data.

Raw AIC values are difficult to compare across situations because the value is heavily affected by sample size. So, Table 10 presents the average change in AIC values across the model types. Here the 2PL model is compared to each of the other models and the 1PL is compared to the CLLM. [Burnham and Anderson \(2004\)](#) suggests that models with a difference less than or equal to 2 both have substantial support for similarity, models with a difference between 4 and 7 have less support for being similar, and values greater than 10 have no support for being a similar model.

The difference values in Table 10 suggest that the 1PL and CLLM models are not as good as the 2PL model. So, in this data situation, the 1PL and CLLM are not able to compete with the 2PL at fitting the data.

Simulation Discussion

The primary motivator of the simulations is to determine model effectiveness as the number of images used decreases and how stable the models are across multiple replications. Simulation 1 compares the performance of three different models (1PL, CLLM, and Poisson) by decreasing the number of images through thinning the total number within image category by ordered parameter estimates and by randomly selecting half of the items. Overall, Brier scores and MSE for the models are better when there are more images, but only by fractions of a point. The randomly selected image sets tend to outperform the thinned image sets. This is likely due to the quantity of easy images being greater than hard images. So, the easy images have a higher chance of being selected when chosen randomly while the hardest images are always kept when using the thinning method.

Results from the first simulation suggest potential concerns with the models and data generation process. The Brier scores and MSE are surprisingly low for all of the conditions and there is very little variability within the replications. Simulation 2 and 3 attempt to address these concerns by fitting much simpler models and modifying the data generating process. In simulation 2, constants are used as predictions for the generated data. The first study predicted that images would be correctly remembered all of the time, and the second study used the overall hit rate from the original data set. Both of these situations did not perform as well as the models, but still had fairly small Brier scores. This in conjunction with the results from simulation 1 suggest that just adding some images improves model prediction, but large numbers of images might not be necessary.

Simulation 3 adjusts the data generating process and adds a new thinning method to try and increase the variability of loss functions within the simulation replications. Overall, results were similar to the first simulation, but the average loss functions were slightly larger and had a bit more variability across replications. The new thinning method orders images by parameter estimate without consideration of category. This method typically fell between the original thinning method and the random selection method in terms of predictive ability. Choosing images this way provides a more balanced option and would be best when the type of image used is not of concern.

Finally, simulation 4 assess how the models perform when the data are generated from a more complex model. Data are generated from a 2PL model and then fit to the 1PL and CLL models. Prediction for the 1PL and CLLM is the worst of the other simulations, but kept a similar pattern when decreasing the number of images. The standard deviations of the Brier scores were also very small in this situation. The 2PL models also fit much better than the others, according to AIC.

In the final chapter, I will summarize the dissertation, consider the larger implications of results, and discuss future directions.

Chapter V

General Discussion

Eye tracking studies are often very time intensive for both the researcher and the participant. The researcher must spend time finding the tens or hundreds of stimuli to be used in the study and then program them into the experimental presentation. The participant then spends upwards of several hours looking at these stimuli and responding to the task set before them by the researcher. The goal of this dissertation is to assess if the number of images used in a study can be reduced using methods from item response theory and what impact, if any, this reduction of items might have on model quality. Reducing the number of images used in a study not only saves researcher and participant time in the current study, but also provides the researcher and potentially other researchers with a pool of established, well performing images for use in future studies. Fewer high quality items also reduce error in model estimation more than using a large number of variable quality items ([Lapedriza et al., 2013](#)).

Image quality was assessed using item response theory because it is an established method for measurement development and has easy to understand and interpret results. IRT provides a way to select a range of items to differentiate between participants across ability level and to remove items that provide little information or are redundant. For this dissertation, eye tracking images were fit to a 1PL model that is parameterized as a GLMM. This method is used because it allows for item or person covariates to be added to the model. Eye tracking data often comes with additional covariates such as response latency, fixation rate, image categories, etc. These models also allow for images to assess an outcome other than a dichotomous response by changing the link function. For eye tracking data, this could be a

response such as fixation counts.

Dissertation Summary

In chapter three, a series of GLMMs are fit to real world data to determine how practical the models are for use with a typical eye tracking experiment. The raw data consists of 630 different target images from 21 different image categories presented to a total of 67 participants. The data are first tested with 1PL models to see how many images could be used without causing estimation issues. From this point, image number was then decreased to see if model performance decreased as the number of images decreased. While the models could not converge using the total number of items, the largest 1PL model I found that would still converge contained 236 images from eight categories. Images with large parameter estimates, suggesting an image is always/never recalled correctly, and images with similar parameter estimates were then removed as they do not provide any additional information. In general, as the number of images were decreased, model performance did slightly decrease, but not enough to justify the use of more complex models. Results were similar for models with additional covariates and using different link functions (e.g. complimentary log-log and Poisson).

While the application had promising results, more evaluation needed to be conducted to make sure the results generalized across multiple replications and to compare other methods of image reduction. So, in Chapter 4, a series of simulations were conducted to determine if results from the application chapter hold up across multiple replications. Various thinning methods, such as random selection and non-categorical thinning, were also compared.

Overall, decreasing the number of images used in a model has little impact on the performance of the model in this data situation. However, the small size of the Brier scores and limited variability between replications is a bit of a concern. Eye tracking

images tend to be on the “easy” side of the difficulty scale and could be contributing to the low Brier scores and variance. The models did perform better than just assuming all images would be recalled correctly or when using an average hit rate.

Model misspecification was also explored in the final part of the simulation section. It is important to correctly specify models for the data being used. It was found that simpler models could not fit data generated from the 2PL, even though the 2PL parameters were chosen based on the original data. There appeared to be problems with fitting the 2PL to the original data set, though, which calls into question the generality of this result. When test fitting 2PL models to the original data in order to get a realistic range of potential parameter values, several of the images have extremely large values for difficulty and discrimination parameters. These models also have difficulty with converging until the images with high hit rates are removed, and even then, new images often took their place with large parameter estimates. It may be that that added complexity of the 2PL model is unnecessary in this situation.

Limitations

While the models are flexible and illustrate how a large number of images can be paired down without losing predictive ability, there are still a few limitations with them and from the data. Of primary concern, is the ratio of number of images to number of participants in the study. Eye tracking experiments often take time because participants are asked to view potentially hundreds of images over multiple experimental sessions. Therefore, sample sizes are typically on the smaller size. While the long format of the data and multiple image presentations help alleviate the small sample size issue, some initial consideration for image selection must still be done. Bayesian estimation methods should also perform better under these conditions.

One way to initially thin out images, is to consider the image content and equality of that content. In the [Bylinskii et al. \(2015\)](#) image set, some images within certain

categories contained features that might be considered out of category. For example, some of the images in the “Badlands” category contained people facing the camera and because people are very good at recognizing faces, these images were almost never missed. Participants almost solely fixate on faces instead of other places in the image. Another example is distinctive signs and ads in the “Airport Terminal” category. Images with very distinct and well known signs were almost never recalled incorrectly. In both of these cases the image category content of the image is overshadowed by more distinctive features of the image and when assessing the memorability of a certain category of images, images with these distinctive features should probably not be used and could be argued that they are even not indicative of the image category. A potential solution worth investigating for this issue is the use of AI generated images. The generated images could still be representative of the category, but contain features that are unfamiliar to people, such as nonsense businesses and ads.

A somewhat related concern, is being able to acquire a selection of items of various difficulties. In the applied data set, the number of easy items far outweigh the number of difficult items. Finding items of sufficient difficulty for eye tracking experiments can be a challenge in and of itself too. While fitting the initial models, when extremely easy images are removed from the data set, others would just take their place in the next model, leading to a cycle of cutting out images with near perfect recall. Having too many easy items also made model estimation difficult and caused convergence issues. There are few items that are considered difficult, so it is worth going back to add more difficult images if there isn't a good spread of items overall. The most difficult ones tend to stick around during the thinning process.

Another potential limitation is the scope of outcomes the images from the applied data are used to measure. The primary study focused on image memorability and while the alternative link functions performed well with the data, the images were not chosen to assess another outcome such as fixation count. The CLL parameters ranked

items similarly to the 1PL model and the Poisson model worked well with the fixation counts. There was only a slight negative relationship between the item parameters in the Poisson model and image hit rate though. While this illustrates that the models work well with alternative outcomes, it would be good to test out the Poisson models with a data set consisting of images designed to influence the number of fixations. The images used here are more natural, so introducing abstract or manipulated images could expand the range of fixation counts.

One interesting issue that came up during the simulation studies is how much difficulty the CLLM model with 59 items had with convergence. While some other models and reps did not converge, it was few and far between when compared to this specific image subset and model combination. This was a much greater issue in the simulation based on the application models and the vast majority of replications had issues with convergence. Interestingly, this was less of a problem when the data were simulated from the 2PL model. This issue occurs with a specific subset of images, so researchers should be aware that images might need to be screened another way (i.e. removing too “easy” images) in order to get models to converge smoothly.

Finally, the models in this dissertation were primarily compared with themselves and not across the different link functions. Since the estimates from the 1PL and CLLM were very similar to each other, it would be worth investigating how the models compare to each other using traditional IRT fit statistics and item response functions. It is possible that the CLLM is overly complex for eye tracking data and doesn't provide any additional information above and beyond that from the 1PL.

Future Directions

This dissertation provides initial groundwork for standardizing images used for eye tracking studies, but there are several different design or statistical directions future projects can take to build and refine the process. The first direction would be trying

out a different type of image to see if it is possible to get more variability in parameter range. For example, maybe abstract images can be refined in a way that a more balanced range of easy to difficult to remember images. Alternatively, exploring more options for outcome measures, such as response latency or fixation lengths, and comparing results from those with each other would be useful and could provide better measure than simple binary responses.

The models themselves can also be extended to use alternative estimation methods and make use of less aggregated data. One of the perks of using the GLMM formulated item response models, is that they are able to handle the partially-crossed nature of the example data set. This lets the researcher assess more items with fewer participants. However, too few participants makes estimating the models difficult. Using Bayesian methods could make model estimation easier for smaller sample sizes and lead to fewer convergence issues through the use of appropriate priors. The models could potentially handle more items with a “normal” eye tracking study sample size.

The models could also be extended to allow for the use of raw eye tracking data where each row of the data set is a single fixation and contains the precise location coordinates for the fixation. [Robinson \(2017\)](#) proposes an item response model that makes use of location data to determine if items behave differently in different geographic locations. This could be extended to the eye tracking context to assess if people that focus on certain areas of an image are more likely to remember the image later than people who focus on other locations. It also wouldn't be unreasonable to expect fixations to be clustered next to each other and tend to be correlated, so an individual fixation would not provide independent information about an image. Person level covariates, such as demographic data, could also be included in the model in order to investigate differences in aspects of the individual instead of just investigating the items.

Instead of using raw fixations, scan path similarities could also be incorporated to see how participants gaze behavior differs between each other and if certain gaze patterns have an association with a “correct” response for the image. The difference between scan paths can be assessed by using string edit difference or cosine similarities. Using this in conjunction with repeated presentation of images can be used to select for images that either induce stable fixation patterns or variable fixation patterns.

Alternative item reduction methods could also be considered and could take a variety of forms. For example, using the fixation locations in conjunction with the item parameters could help a researcher select for images that have a range of difficulties for the desired task and induce more variable scanning behaviors. Alternatively, if another metric like hit rate for an image is also known, a distance measure, such as Mahalanobis distance, can be used to eliminate images with parameter estimates and hit rates that disagree with each other. This assumes there is a relationship between the parameters and alternative measure though.

Experiments can be further sped up by considering how much time is actually need during stimulus presentation to observe the effect of interest. For example, if differences in latency till first fixation is indicative of the experimental effect of interest, any fixation after that point is potentially non-informative. Therefore, stimulus presentation time can be cut down and could save time over many image presentations.

Finally, the methods can also be used to create a library of standardized sets of images researchers can access for use in experiments. Images can be classified by content, difficulty, and for specific paradigm uses. Ideally, researchers would be able to provide the specifics of their experimental design to a program, and the program would give them a sample of images specifically validated for their type of study.

Conclusion

In this paper, I first discussed various types of eye tracking data and traditional models used to analyze the data. I then proposed a method using item response models formulated as a GLMM to reduce the number of images needed to be used in a study in order to help conserve participant and researcher time, retain information in the data, and maintaining or even improve results from data analysis. This method is then illustrated using data from an image memorability study and reinforced with follow up simulations. In conclusion, researchers conducting eye tracking studies should consider the images they are using in their study. Images selection should be refined and retain high quality images that provide as much information as possible. This dissertation provides a starting point for standardizing eye tracking stimuli to reduce the number of images needed in a study and retain maximal information, saving participants and researchers valuable resources.

References

- Aracena, C., Basterrech, S., Snáel, V., & Velásquez, J. (2015). Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 2632–2637).
- Arai, M., Van Gompel, R. P., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*(3), 218–250.
- Barr, D. J. (2008a). Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.
- Barr, D. J. (2008b). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*(1), 18–40.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01
- Bates, D., Mullen, K. M., Nash, J. C., & Varadhan, R. (2014). minqa: Derivative-free optimization algorithms by quadratic approximation [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=minqa> (R package version 1.2.4)
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Brown-Schmidt, S. (2009). The role of executive function in perspective taking during online language comprehension. *Psychonomic Bulletin & Review*, *16*(5), 893–900.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.
- Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and

- extrinsic effects on image memorability. *Vision Research*, *116*, 165–178.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06
- Cho, S.-J., Brown-Schmidt, S., De Boeck, P., & Shen, J. (2020). Modeling intensive polytomous time-series eye-tracking data: A dynamic tree-based item response model. *Psychometrika*, 1–31.
- Cho, S.-J., Brown-Schmidt, S., & Lee, W.-y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: an application to intensive binary time series eye-tracking data. *Psychometrika*, *83*(3), 751–771.
- Cho, S.-J., Wu, H., & Naveiras, M. (2023). The effective sample size in Bayesian information criterion for level-specific fixed and random-effect selection in a two-level nested model. *British Journal of Mathematical and Statistical Psychology*, *77*(2), 289–315. doi: <https://doi.org/10.1111/bmsp.12327>
- Dalrymple, K. A., Jiang, M., Zhao, Q., & Elison, J. T. (2019). Machine learning accurately classifies age of toddlers based on eye tracking. *Scientific Reports*, *9*(1), 1–10.
- De Ayala, R. J. (2013). *The Theory and Practice of Item Response Theory*. New York: Guilford Publications.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, *39*, 1–28.
- De Boeck, P., Partchev, I., et al. (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software*, *48*(1), 1–28.
- Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 1–17. doi: 10.1167/8.2.6

- Frazier, T. W., Strauss, M., Klingemier, E. W., Zetzer, E. E., Hardan, A. Y., Eng, C., & Youngstrom, E. A. (2017). A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *Journal of the American Academy of Child & Adolescent Psychiatry, 56*(7), 546–555.
- Haji-Abolhassani, A., & Clark, J. J. (2013). A computational model for task inference in visual search. *Journal of vision, 13*(3), 1–24. doi: <https://doi.org/10.1167/13.3.29>
- Hamaker, E. L., van Hattum, P., Kuiper, R. M., & Hoijtink, H. (2011). Model selection based on information criteria in multilevel modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis*. New York: Routledge/Taylor & Francis Group.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York, NY: Springer Science & Business Media.
- Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition, 108*(3), 831–836.
- Hessels, R. S., Kemner, C., van den Boomen, C., & Hooge, I. T. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavior Research Methods, 48*(4), 1694–1712.
- Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (i2mc). *Behavior Research Methods, 49*(5), 1802–1823.
- Holmqvist, K., Örbom, S. L., Hooge, I. T., Niehorster, D. C., Alexander, R. G., Andersson, R., . . . others (2023). Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods, 55*(1), 364–416.
- Indrarathne, B., Ratajczak, M., & Kormos, J. (2018). Modelling changes in the cognitive processing of grammar in implicit and explicit learning conditions: Insights from an eye-tracking study. *Language Learning, 68*(3), 669–708.

- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489–1506.
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203.
- Jovancevic, J., Sullivan, B., & Hayhoe, M. (2006). Control of attention and gaze in complex environments. *Journal of Vision*, *6*(12), 1431–1450. doi:
<https://doi.org/10.1167/6.12.9>
- Kaakinen, J. K. (2021). What can eye movements tell us about visual perception processes in classroom contexts? commentary on a special issue. *Educational Psychology Review*, *33*(1), 169–179.
- Kärrsgård, I., & Lindholm, A. (2003). *Eye movement tracking using hidden Markov models* (Doctoral dissertation, Chalmers University of Technology). Retrieved from <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ef861907c9754ceda8ea9c69a7353fe5b4bc4003>
- Kim, J., Singh, S., Thiessen, E. D., & Fisher, A. V. (2020). A hidden Markov model for analyzing eye-tracking of moving objects: Case study in a sustained attention paradigm. *Behavior Research Methods*, *52*, 1225–1243.
- Knoeferle, P., Crocker, M. W., Scheepers, C., & Pickering, M. J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition*, *95*(1), 95–127.
- Koutsogiorgi, C. C., & Michaelides, M. P. (2022). Response tendencies due to item wording using eye-tracking methodology accounting for individual differences and item characteristics. *Behavior Research Methods*, *54*(5), 2252–2270.
- Krol, M., & Krol, M. (2017). A novel approach to studying strategic decisions with eye-tracking and machine learning. *Judgment & Decision Making*, *12*(6), 596–609.
- Lapedriza, A., Pirsiavash, H., Bylinskii, Z., & Torralba, A. (2013). Are all training

examples equally valuable? *arXiv preprint arXiv:1311.6510*.

- Lorah, J., & Womack, A. (2019). Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behavior Research Methods*, *51*, 440–450. doi: <https://doi.org/10.3758/s13428-018-1188-3>
- Mack, D. J., Belfanti, S., & Schwarz, U. (2017). The effect of sampling rate and lowpass filters on saccades—a modeling approach. *Behavior Research Methods*, *49*(6), 2146–2162.
- Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *31*(1), 133–156.
- Mastergeorge, A. M., Kahathuduwa, C., & Blume, J. (2021). Eye-tracking in infants and young children at risk for autism spectrum disorder: A systematic review of visual stimuli in experimental paradigms. *Journal of Autism and Developmental Disorders*, *51*, 2578–2599.
- McCoach, D. B., & Black, A. C. (2011). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing, Inc.
- McMurray, B., Klein-Packard, J., & Tomblin, J. B. (2019). A real-time mechanism underlying lexical deficits in developmental language disorder: Between-word inhibition. *Cognition*, *191*, 104000.
- Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494.
- Mould, M. S., Foster, D. H., Amano, K., & Oakley, J. P. (2012). A simple nonparametric method for classifying eye fixations. *Vision Research*, *57*, 18–25.
- Mozuraitis, M., Chambers, C. G., & Daneman, M. (2015). Privileged versus shared knowledge about object identity in real-time referential processing. *Cognition*,

142, 148–165.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention.

Vision Research, 45(2), 205–231.

Nuthmann, A., Einhäuser, W., & Schütz, I. (2017). How well can saliency models predict fixation selection in scenes beyond central bias? A new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience*, 11, 1–21. doi: <https://doi.org/10.3389/fnhum.2017.00491>

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, 26(6), 2708–2725.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1), 107–123.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>

Rizopoulos, D. (2007). ltm: An r package for latent variable modeling and item response analysis. *Journal of Statistical Software*, 17, 1–25.

Robinson, S. (2017). *Local item response theory for detection of spatially varying differential item functioning* (Doctoral dissertation, University of Arkansas). Retrieved from <http://scholarworks.uark.edu/etd/2520>

Schoemann, M., Schulte-Mecklenbeck, M., Renkewitz, F., & Scherbaum, S. (2019). Forward inference in risky choice: Mapping gaze and decision processes. *Journal of Behavioral Decision Making*, 32(5), 521–535.

Shi, S. W., Wedel, M., & Pieters, F. (2013). Information acquisition during online decision making: A model-based exploration using eye-tracking data. *Management Science*, 59(5), 1009–1026.

- Shim, H., Bonifay, W., & Wiedermann, W. (2022). Parsimonious asymmetric item response theory modeling with the complementary log-log link. *Behavior Research Methods*, 1–20.
- Torralba, A., Oliva, A., Castelhana, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113(4), 766. doi: <https://psycnet.apa.org/doi/10.1037/0033-295X.113.4.766>
- Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research*, 43(3), 333–346.
- van Renswoude, D. R., Raijmakers, M. E., Koornneef, A., Johnson, S. P., Hunnius, S., & Visser, I. (2018). Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior Research Methods*, 50(2), 834–852.
- Wang, K., Su, H., & Ji, Q. (2019). Neuro-inspired eye tracking with eye movement dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9831–9840).
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5(2), 245–262.
- Yen, W. M. (1987). A comparison of the efficiency and accuracy of bilog and logist. *Psychometrika*, 52(2), 275–291.
- Zhan, L. (2018). Using eye movements recorded in the visual world paradigm to explore the online processing of spoken language. *JoVE (Journal of Visualized Experiments)*(140), e58086.
- Zhang, Y., & Hornof, A. J. (2011). Mode-of-disparities error correction of eye-tracking data. *Behavior Research Methods*, 43(3), 834–842.

Vita

Benjamin Graves studied psychology at Henderson State University where he received a Bachelor of Science degree in December 2013. While there, he focused his studies on animal and human learning and behavior, trained chickens under a renowned animal behaviorist, and presented award winning research on alternative measures of achievement. He then attended Missouri State University where he obtained a Master of Science degree in experimental psychology in July 2016. His thesis entitled “Methods of Measuring Visual Scanning of Upright and Inverted Ecological Images” explores differences in how people visually scan ecological images that have been flipped upside down when compared to their upright counterparts. Ben then attended the University of Missouri where he completed his PhD in quantitative psychology in May 2024. Here he developed his quantitative skills and combined it with his experience with eye tracking paradigms. His dissertation entitled “A Modern Test Theory Approach to Selecting Eye Tracking Stimuli” proposes a method of streamlining eye tracking through the use of item response models. Ben currently works for the Missouri Prevention Science Institute as the director of data strategy.