IRT in SPSS: The development of a new software tool to conduct item response models

A Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

John DiTrapani, B.S.
Graduate Program in Psychology

The Ohio State University

2016

Master’s Examination Committee:

Dr. Minjeong Jeon, Advisor
Dr. Paul De Boeck
Dr. Michael Edwards
Abstract

Item Response Theory (IRT) is a modeling framework that can be applied to a large variety of different research questions in many different disciplines. Currently, there are many software packages that can be used to conduct IRT, but they can be difficult to access for a practitioner who primarily uses software like SPSS. To make IRT models more accessible for the general researcher, a tool has been created that can easily run one-parameter IRT (1PL) models on SPSS without any required downloads or add-ons. This tool takes advantage of the fact that a 1PL model is a member of the generalized linear mixed model (GLMM) family by utilizing the built-in SPSS function for GLMMs. It can run 1PL models with person and item covariates, DIF analyses, multidimensional models, IRTree models, and many other variations. It also provides the user with item and person fit statistics, item characteristic curves, item information functions, and several other options. Comparisons show that the estimates provided by this macro are very similar to those estimated from other software packages.
Vita

July 16, 1992 ........................................ Born - Virginia, USA

2014 .................................................. B.S. Psychology, Virginia Tech

2014 .................................................. B.S. Statistics, Virginia Tech

Fall 2014-Fall 2015 .............................. Distinguished University Fellow,
                                             The Ohio State University

Fall 2015-present ............................. Graduate Teaching Associate,
                                            The Ohio State University.

Publications

Research Publications


Fields of Study

Major Field: Psychology
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>Vita</td>
<td>3</td>
</tr>
<tr>
<td>List of Tables</td>
<td>6</td>
</tr>
<tr>
<td>List of Figures</td>
<td>7</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>9</td>
</tr>
<tr>
<td>2. 1PL Models in the GLMM Framework</td>
<td>12</td>
</tr>
<tr>
<td>3. The SPSSIRT Macro</td>
<td>16</td>
</tr>
<tr>
<td>3.1 Wide and long data format</td>
<td>17</td>
</tr>
<tr>
<td>3.2 SPSSIRT syntax</td>
<td>18</td>
</tr>
<tr>
<td>4. 1PL Variations and Examples Using SPSS</td>
<td>21</td>
</tr>
<tr>
<td>4.1 Example dataset</td>
<td>21</td>
</tr>
<tr>
<td>4.2 1PL model</td>
<td>22</td>
</tr>
<tr>
<td>4.3 Person covariates</td>
<td>24</td>
</tr>
<tr>
<td>4.4 Item covariates</td>
<td>27</td>
</tr>
<tr>
<td>4.5 Multidimensional models</td>
<td>29</td>
</tr>
<tr>
<td>4.6 Item and person covariates</td>
<td>31</td>
</tr>
<tr>
<td>4.6.1 DIF models</td>
<td>32</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>17</td>
</tr>
<tr>
<td>4.1</td>
<td>22</td>
</tr>
<tr>
<td>6.1</td>
<td>47</td>
</tr>
<tr>
<td>6.2</td>
<td>49</td>
</tr>
<tr>
<td>6.3</td>
<td>51</td>
</tr>
<tr>
<td>6.4</td>
<td>53</td>
</tr>
<tr>
<td>6.5</td>
<td>55</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Item intercept ($\beta_i$) estimates calculated from identical basic 1PL models estimated in SPSS, lme4, and TAM.</td>
<td>45</td>
</tr>
<tr>
<td>6.2</td>
<td>$\theta_p$ estimates calculated from identical basic 1PL models estimated in both SPSS and lme4.</td>
<td>46</td>
</tr>
<tr>
<td>6.3</td>
<td>$\theta_p$ estimates calculated from identical latent regression models estimated in SPSS, lme4 and TAM.</td>
<td>48</td>
</tr>
<tr>
<td>6.4</td>
<td>$\theta_p$ estimates calculated from identical LLTM models estimated in both SPSS and lme4.</td>
<td>50</td>
</tr>
<tr>
<td>6.5</td>
<td>$\theta_p$ estimates for the “do” dimension for SPSS, lme4, and TAM models.</td>
<td>52</td>
</tr>
<tr>
<td>6.6</td>
<td>$\theta_p$ estimates for the “want” dimension for SPSS, lme4, and TAM models.</td>
<td>54</td>
</tr>
<tr>
<td>6.7</td>
<td>$\theta_p$ estimates calculated from identical DIF models estimated in both SPSS and lme4.</td>
<td>56</td>
</tr>
<tr>
<td>B.1</td>
<td>Image taken from SPSSIRT output. This is the “Fixed Effects” table, which gives item parameter ($\beta_i$) estimates and covariate effect ($\gamma$) estimates. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.</td>
<td>81</td>
</tr>
<tr>
<td>B.2</td>
<td>Image taken from SPSSIRT output. This is the “Person Statistics” table, which gives $\theta_p$ values for every respondent, as well as person fit statistics. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.</td>
<td>82</td>
</tr>
</tbody>
</table>
B.3 Image taken from SPSSIRT output. This is the “Item Fit Statistics” table, which gives item fit statistics. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.

B.4 Image taken from SPSSIRT output. This is the “Random Effect” table, which gives the random effect variance estimate(s) ($\sigma^2$) for the model. In a multidimensional model, the covariance(s) between dimensions will also be given here. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.

B.5 Image taken from SPSSIRT output. This depicts the item characteristic curves for all 24 items in the verbal aggression dataset, created using the ICC function.

B.6 Image taken from SPSSIRT output. This depicts the item information as a function of theta. This image was created using the ICC function.
Chapter 1: Introduction

Item response theory (IRT) is a modeling framework commonly used within the domains of psychology, education, public health, and many other disciplines. The main appeal of IRT models is that they allow users to model dichotomous or polytomous scale or test data to obtain estimates of both person measures and item properties. These estimates are mapped onto the same scale, making it easy to intuitively compare a respondent’s latent ability with the difficulty of an item. Since this type of modeling is often used to acquire measures of person ability, it is traditionally very common in the area of education. However, IRT methods can and have been applied to other situations where discrete behavioral data are present, such as with scale responses in psychological research or with behavioral outcomes in public health disciplines.

Within the framework of item response theory, one of the most popular models is the Rasch model (Rasch, 1960), which is computationally simpler than many other IRT approaches and has some very useful properties. The Rasch model has been applied to settings as diverse as the measurement of computer anxiety (King, Bond, & Blandford, 2002), the variability of water repellency in fire-affected soils (Bodi et al., 2013), and the investigation of gender by item interactions on reading comprehension exams (Schwabe, McElvany, & Trendtel, 2015). The general areas of public health
and clinical psychology also commonly utilize the Rasch model. Aggen, Neale, and Kendler (2005) provide an example of how IRT models, including the Rasch model, can analyze DSM criteria on depression in more beneficial ways, while Hagquist, Bruce, and Gustavsson (2009) present an introduction on scale development using Rasch models in the field of nursing. Small et al. (2008) utilized Rasch models to examine group differences in symptomologies for depressed adolescents. In the domain of dermatology, Nijsten, Unaeze, and Stern (2006) used Rasch methods to enhance the effectiveness of a patient questionnaire measuring the impact of Psoriasis. These examples are just a few of many applications of Rasch models across many disciplines answering a large variety of research questions.

As the popularity of IRT grows, there have been a multitude of different software created that can estimate this type of model. Many options are packages within R, such as FLIRT (Jeon, Rijmen, & Rabe-Hasketh, 2014), ltm (Rizopoulos, 2006), mirt (Chalmers, 2012), mlirt (Fox, 2007), eRm (Mair & Hatzinger, 2007), TAM (Kiefer, Robitzsch, & Wu, 2016), and lme4 (Bates, Mächler, Bolker, & Walker, 2015). Other options include software specifically programmed for latent variable analyses, such as Mplus (Muthen & Muthen, 1998-2011), flexMIRT (Houts & Cai, 2013), HLM (Raudenbush, Bryk, & Congden, 2004), and Latent GOLD (Vermunt & Magidson, 2013). More general statistical software can be adapted to run item response models, such as SAS PROC NLMIXED (SAS Institute Inc., 2015), the glamm package within Stata (Rabe-Hasketh, Skrondal, & Pickles, 2004), or the aforementioned lme4 package within R (De Boeck et al., 2011). These packages all are slightly different in the possible types of analyses that can be conducted and the estimation methods used.
Despite this relatively large amount of alternatives, it can be daunting for a researcher to choose and correctly utilize an IRT modeling software for the first time. The packages often require a fair amount of background knowledge to correctly use. Also, software that was not specifically made to conduct item response models (such as Mplus, SAS, or lme4) provides output that may be difficult for the common practitioner to interpret (even one who has a background in IRT), since it may contain vocabulary inconsistent with the common IRT literature.

This paper will present a new, more accessible option for researchers familiar with the software SPSS. Using the GENLINMIXED function from the popular software SPSS (IBM Corp., Released 2013), it is entirely possible to conduct Rasch IRT models and other one-parameter item response models by taking advantage of the fact that this type of model is within the generalized linear mixed model (GLMM) family (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; De Boeck et al., 2011). A macro that can allow the practitioner to quickly and easily execute IRT models in SPSS will be introduced along with its properties and limitations. The different variations that can be done are also presented along with concrete data examples. The hope is that this macro will allow these valuable models to be more accessible to the general research community by putting them in the SPSS environment that is familiar to many researchers.
Chapter 2: 1PL Models in the GLMM Framework

Item response models are usually (but not always) applied to datasets which are composed of $P$ individuals being exposed to $I$ items. Often these observations are binary, such as an item response being “correct” or “incorrect.” Because of this format, an item response by person $p$ on item $i$ can be treated as a Bernoulli random variable with probability of a “success” $\pi_{ip}$. The generalized linear mixed modeling (GLMM) framework allows for the linear modeling of Bernoulli (and other discrete) variables through the utilization of a link function. A GLMM is composed of a link function, a linear equation of predictors, and a random component. It can be shown that a Rasch model is within the GLMM family since it contains each of the three elements necessary to a GLMM.

To illustrate how a Rasch model is within this framework, take as an example the simplest version of a Rasch model, which can be described as follows:

$$P(Y_{ip} = 1|\theta_p) = \pi_{ip} = g\text{link}(\eta_{ip}),$$

where $\pi_{ip}$ is the probability that person $p$ answers with a “1” instead of a “0” on item $i$ (or correctly answers item $i$ if the context is an exam of some sort). This probability is a function of both a linear component (denoted as $\eta_{ip}$) and a link
function, which is represented as \( g_{\text{link}}(\cdot) \). The linear component is the sum of item and person effects. In the simplest Rasch model, this sum is often:

\[
\eta_{ip} = \beta_i + \theta_p,
\]

where \( \beta_i \) is a fixed coefficient that can be interpreted as the “easiness” or “intercept” for item \( i \). As parameterized here, an item with a higher \( \beta_i \) will tend to have a higher probability of being answered as “correct” or as a “1.” This item parameter is a fixed effect, meaning we are assuming it is a fixed parameter for all respondents encountering this item. The \( \theta_p \) value indicates the latent variable value for person \( p \). This random person estimate is the random component of the GLMM. This random effect is included because it models the dependence within a cluster of observations. The model assumes that given this random effect, all observations are independent of one another. If the random effect was not included in the model, the model would then assume that all observations are independent, which we know not to be true since the data are clustered by respondents. Hence, a mixed model that includes a random effect is needed.

The distribution of the \( \theta_p \) values is usually set to be normal with a mean of zero and an unconstrained variance of \( \sigma^2 \) (which is estimated by the model). Often, this random effect is interpreted as the “ability” of person \( p \), if the data are in the context of an examination. The higher a participant’s \( \theta_p \), the higher that person’s probability of correctly answering an item. The probability of a response of “1” is therefore a function of the easiness of the item and the general ability of the person. This term \( (\eta_{ip}) \) contains both the fixed and random components of a generalized linear mixed model. In later sections, more 1PL models will be introduced that involve more
complicated $\eta_{ip}$ terms that contain different covariates, interactions between effects, and other alternatives.

The last critical element of a GLMM is the link function. It permits researchers to model data that do not range across the real number line (like binary data, which consist of only 0’s and 1’s) by mapping this data between negative infinity and infinity. A linear model of predictors can then be used to estimate this transformed parameter. A common example of a link function for binary data is the logit function. For this link function, the log-odds of the probability of a successful trial are modeled with a linear equation of predictors:

$$Y_{ip} \sim Bernoulli(\pi_{ip}),$$

$$\ln\left(\frac{\pi_{ip}}{1 - \pi_{ip}}\right) = \eta_{ip}.$$  \hspace{1cm} (2.3)

Here, $\eta_{ip}$ is defined as it was above. Within the simplest Rasch model introduced in Equation 2.1, $\eta_{ip}$ is the sum of an item’s intercept ($\beta_i$) and a person’s ability ($\theta_p$). These item and person estimates are therefore interpreted on the logit, or logistic scale. The $\ln\left(\frac{\pi_{ip}}{1 - \pi_{ip}}\right)$ term corresponds to the log-odds of the probability of a successful response (or a response of “1”). To calculate estimates of $\pi_{ip}$, Equation 2.3 can be restructured such that

$$\pi_{ip} = \frac{e^{\eta_{ip}}}{1 + e^{\eta_{ip}}}.$$  \hspace{1cm} (2.4)

Therefore, a calculated $\eta_{ip}$ of zero will result in a modeled probability of success of $\pi_{ip} = \frac{e^0}{1 + e^0} = 0.5$. According to this hypothetical model, person $p$ would have a 50% chance of obtaining a “success” on item $i$ if $-\beta_i = \theta_p$. It is therefore easy to see
that any \( \eta_{ip} \) value greater than zero corresponds to a modeled \( \pi_{ip} \) greater than 0.5, and an \( \eta_{ip} \) below zero results in a \( \pi_{ip} \) less than 0.5.

The logit link function will be the default link function in the SPSS macro being presented here, since it is used commonly in practice. Another link function that is often used is the probit link function, which is also available in the introduced SPSS macro (link function specification within the macro will be shown later). Instead of modeling the log-odds of success, this link function is based off the normal cumulative distribution function. These two methods will yield estimates that appear numerically differently, but are actually only different by a numerical constant.

Equations 2.1 and 2.2 show that standard 1PL models contain all elements of a generalized linear mixed model: a link function, a linear equation of fixed coefficients, and a random effect. Since a Rasch model falls into this model family, it is possible to take advantage of this fact by using popular statistical software that can perform generalized linear mixed models. Since SPSS has this capacity using the GENLINMIXED function, it is therefore possible to conduct 1PL analyses on this popular software.
Chapter 3: The SPSSIRT Macro

The SPSSIRT macro allows researchers to conduct 1PL analyses through the typical SPSS syntax interface without any external add-ons or expert knowledge. It is available and easily implemented on SPSS version 21 and above. Since a 1PL model is a specific case of a generalized linear mixed model, the macro utilizes the GENLINMIXED function of SPSS. One of the main advantages of the SPSSIRT tool is that the required syntax and outputted results are designed specifically in the language of IRT and are as a result simple to specify and straightforward to understand. Researchers who are moderately familiar with the connections between GLMMs and Rasch models may have trouble interpreting the typical GENLINMIXED output, since the item intercept parameters are called “fixed coefficients” and the individual random effect values are not given for each individual (only the variance of this random effect is given in the typical output). However, since the SPSSIRT macro was made specifically with IRT users in mind, the code is more concise and more IRT-inspired, and the output gives the item parameters and person latent variable estimates. This section of the paper will cover several requirements and properties of the SPSSIRT macro.
(a) An example dataset in the long format. Each item response has its own row. Since this example contains two respondents and three items, there are six rows.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Item</th>
<th>Resp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

(b) An example dataset in the wide format. Each subject has its own row. Since this example contains two respondents and three items, there are two rows and a column for each of the three items.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.1: Example item response data in long and wide formats.

3.1 Wide and long data format

For SPSS to correctly run its GENLINMIXED function, data must be presented in long form. This is true for any generalized linear mixed models in SPSS, not just the IRT models being presented here. In this data format, every item response has its own row, as opposed to wide form data, where every respondent has his or her own row. Take for an example Table 3.1, which has a short dataset in both long (Table 3.1a) and wide (Table 3.1b) form.
Long form data is frequently used when data are longitudinal in nature or if the data come from a repeated measures design. It is helpful for GLMMs since it allows both item and person variables to be entered into the data. The wide format does not allow for item covariates because there is not a row for each observation. Despite these advantages, the long format is not particularly common in IRT research, and the wide format is more frequently required for IRT software. It is relatively straightforward to use the VARSTOCASES function on SPSS to restructure a dataset from wide to long form.

3.2 SPSSIRT syntax

Before any SPSSIRT syntax can be run, the user must turn off the default “Model Viewer” setting. To do this, simply use the “Edit” dropdown and go to the “Options” button. From there, click on the “Output” tab. Here, there should be an option to switch to a “Pivot Tables” setting for viewing output. This option must be selected for the macro to provide output in an aesthetically pleasing manner. This change should only need to be made once; “Pivot Tables” should remain the selected option moving forward, even if the user ends the current session.

Once this change is made, the macro is ready to be used. The SPSSIRT macro contains two functions:

1. **Rasch**: This function is used to run every type of IRT model that can be run on SPSS.

2. **ICC**: This function allows for the creation of plotted item response curves. The user specifies the item or item intercept value that he or she wants to plot. It
also provides a plot for the item information for each individual specified item and the aggregated information value for the entire scale.

Generally speaking, the syntax for the Rasch function is structured as follows:

Rasch items = /cov = /response = /id = /dim = /difvar = /difitem = /link = logit'/theta = 'Yes'/fit = 'Yes'/save = '0'/iters = '100'/conf = '95'.

The macro code (which is included in full in Appendix A) must be copied and pasted into an SPSS syntax file, and then run within SPSS. After the macro is activated, running the above syntax will utilize the GENLINMIXED function and output IRT results. The user must specify the variable column that contains item names as a categorical variable (the “items” command), the variable containing the item responses (the “response” command), and the respondent id column as a categorical variable (the “id” command). If desired, other covariates can be included into the model with the “cov” command (models involving this command, as well as more complicated commands using the “dim,” “difitem,” and “difvar” commands, are described later). The “link” command specifies what link function to use when the IRT model is run. The default is the logit link function, but the probit link function can also be used if the user includes “/link = probit.”

Finally, the macro also allows for the specification of the number of iterations (the “iters” command, where the default is 100) and the confidence level of any hypothesis tests conducted (the “conf” command, where the default is 0.95). The “save” command controls what output values, if any, are saved into the original SPSS dataset. The default for this command is “save = 0,” which will result in no output values (parameter estimates, predicted probabilities, etc.) to be saved into the original dataset.
(all output values will be shown only in the output window). However, the user can choose what values he or she wants to save into the dataset for further use. If “save = 1” then all item and person statistics are put into the dataset, including standard errors, confidence intervals, p-values, and fit statistics. If “save = 2” then only item effects are saved, while “save = 3” will output the item parameter estimates as well as the random effect for the individual of each response.

After the model is run, output will be provided in several different ways. Most importantly, the output will contain clearly titled tables that provide estimates of item intercept parameters (the “Fixed Effects” table; a screenshot is shown in Figure B.1), person $\theta_p$ values (the “Person Statistics” table, which also contains the Dragow’s $Z_3$ person fit statistics for each respondent, which will be described later; a screenshot is shown in Figure B.2), and a table for item fit statistics such as infit and outfit, which will be discussed in full in Chapter 5 (this table can be removed by specifying “fit = NO”; a screenshot is shown in Figure B.3). The $\theta_p$ values will appear by default into the output window, but can be hidden if “theta = NO” is specified in the syntax when the model is run. The “Random Effect” table includes the variance of the person random effect being used and, in cases of multidimensional models that are described in Section 4.7, any covariances between random effects (a screenshot is shown in Figure B.4). Examples of item characteristic curves and information function graphs are given in Figures B.5 and B.6.
Chapter 4: 1PL Variations and Examples Using SPSS

4.1 Example dataset

Different 1PL models will now be investigated and run using the SPSSIRT macro. As a running example, models will be conducted using a dataset that attempts to measure a respondent’s level of verbal aggression. This dataset, which is used extensively in De Boeck and Wilson (2004) and is publicly available in the lme4 package in R, contains responses from 316 respondents to 24 different items. Each of the items presents a hypothetical, “frustrating” situation and asks whether the respondent would “want” (for one half of the items) or “do” (the other half of the items) a particular verbally aggressive behavior. These 24 items vary in several ways. First of all, there are four separate hypothetical situations (so six items per situation). Two of these four situations are scenarios in which the respondent is at fault, and the other half of the items are situations in which another person is at fault. Within each situation, the respondent is asked whether he/she would want to curse, want to shout, want to scold, actually curse, actually shout, or actually scold the other person in the scenario. This leads to 24 items (six per situation) which all contain a “mode” characteristic (do vs. want) and “situation” property (other person’s fault vs. respondent’s fault), and a behavior type description (shout, scold, or curse).
Table 4.1: The first three rows for the verbal aggression dataset in long form.

Responses to these items are, for the sake of the present analyses, dichotomized to either a “yes” or “no” response. Affirmative answers suggest that the respondent would feel that he/she would be inclined to participate in that specific item’s behavior. The respondent’s gender and aggregated score on a separate scale that attempts to measure a respondent’s level of anger are both provided as person covariates. The first three rows of this dataset are shown in Table 4.1. The rows shown correspond to the responses by respondent 1 to the “S1DoCurse”, “S1DoScold”, and “S1DoShout” items. The first three columns represent the id number, item, and response to the specific row’s item response. The “btype” column signifies the behavior type for that specific item, the “mode” variable tells whether the item asks a “do” question or a “want” question, and the “situ” column says whether or not the item’s hypothetical situation was the respondent’s fault or another person’s fault. The “Gender” and “Anger” columns are the two person covariates. In the following sections, a variety of example 1PL models will be conducted using this dataset.

4.2 1PL model

The first model that will be examined is the most straightforward version of the Rasch family. It contains no item or person covariates and only one dimension,
making it the easiest to comprehend and the simplest to run in SPSSIRT. This is also the model that was presented earlier in chapter 2. When a logit link function is used, this model is often called the one parameter logistic model, or the 1PL. Often this model is presented generally as the mixed logistic regression model (Hedeker & Gibbons, 1994). To review, the probability that person \( p \) answers correctly or aggressively (or with a “1”) on item \( i \) can be modeled as a function of the sum of the easiness/intercept of item \( i \) (\( \beta_i \)) and the ability, or aggressiveness of person \( p \) (\( \theta_p \)), as described in Equations 1-3 in chapter 2.

This model is very simply specified in the SPSSIRT syntax. In fact, the command names of the function were specifically named with the basic 1PL model in mind, so the commands are self-explanatory. If one was to conduct this model using the verbal aggression dataset, the following code would be used:

\[
\text{Rasch items} = \text{item/response} = \text{ans/id} = \text{id}.
\]

Here, the “item” variable is the column which includes the item for each item response. This variable must always be specified as a nominal variable. The “ans” column includes the actual item responses, which should be denoted as “1” for a “yes” response and as “0” for a “no” response. And finally, the last variable which needs to be specified is the “id” column, which specifies the specific respondent for a given item response.

This syntax will give all item intercept values and all person effects (as well as the variance of these person random effects, a parameter of the model). This model is commonly used to obtain a measurement error-free “score” for each respondent. For this specific dataset, the \( \theta_p \) values represent the overall tendency for a respondent
to self-report behavior that is verbally aggressive. Recall that the $\theta_p$ latent variable values will have a mean of 0 and will be normally distributed. It is worth noting that although this data do not come from a traditional test measuring some type of intellectual ability, an IRT model is still perfectly reasonable and beneficial to use. The interpretation of a $\theta_p$ value is no longer a measure of ability, but a measure of the verbal aggression of a participant.

### 4.3 Person covariates

One-parameter IRT models can be extended to include other fixed effects in addition to the aforementioned item intercept parameters. One type of fixed effect that can be included into the model is a person-level trait, which will be denoted as a person covariate. Models that contain person-level covariates can be denoted as latent regression models (Adams, Wilson, & Wang, 1997). Covariates from latent regression models can be binary, categorical, or continuous, and represent properties of a respondent (such as age, gender, measured intelligence, etc.). When these covariates are included, it is possible to investigate the association between the person property and the measured latent trait of a respondent. Therefore, this type of model can be used to explain relationships between predictors and item responses, above and beyond the difficulty of a given item and controlling for person dependence. Since models contain person covariates (and/or item covariates, which will be described later), these models are often called *explanatory* IRT models, since the goal of the model is more focused on explaining an item response with an assortment of predictors as opposed to purely measuring person traits (De Boeck & Wilson, 2004).
When person covariates are included into the model, the linear component of the IRT model, $\eta_{ip}$, can now be described as:

$$\eta_{ip} = \beta_i + \theta_p, \quad (4.1)$$

$$\theta_p = \sum_{j=1}^{J} \gamma_j X_{pj} + \theta_p^*.$$ 

In this formula, there are $J$ person covariates included into the explanatory IRT model. Each person has a value for each covariate, which represent the $X_{pj}$ values. The $\gamma_j$ values are the fixed regression coefficients for the $j$-th person covariate. These $\gamma$ estimates are similar to the $\beta$ item parameters in that they are fixed, deterministic predictors of $\theta_p$. Hypothesis tests can therefore be conducted to determine whether these covariate coefficients are different from zero. If they are statistically significantly different from zero, one can conclude that the $j$-th covariate is in fact associated with, in this example dataset, a person’s level of verbal aggression. It also should be mentioned that many covariates can be added into the model, as well as possible interactions between covariates.

Note that the random effect in this version of the model is denoted as $\theta_p^*$, not as $\theta_p$ like before. This is to emphasize that the person value in this model does not carry the same meaning as the $\theta_p$ from previously described models. In the simple Rasch model from section 4.2, $\theta_p$ represents person $p$’s latent variable value relative to all other respondents. A value of “0” would roughly be indicative of a person with an average measure of the latent variable. However, in the model with person covariates, the $\theta_p^*$ value represents the residual of person $p$ for this specific regression model. It is therefore not a direct measure of a person’s ability; it is important to note that
models with covariates are not suited to get a $\theta_p$ estimate representing the overall trait value for person $p$.

Person covariates are simple to specify within the SPSSIRT syntax. The code is the same as it was for the basic Rasch model, except now the “cov” command must be included. This command is used to specify the variable in the dataset that contains the person covariate value. To specify multiple covariates, simply list each variable, separating each with a space. Using the verbal aggression dataset, assume researchers were interested in the effect of gender and a respondent’s score on the included anger scale on the likelihood of a “successful” item response. The code to run this model would be as follows:

\[
\text{Rasch items = item/cov = Gender Anger/response = ans/id = id.}
\]

The output provided would then yield item easiness parameters for every item as well as the coefficient estimate for the effect of gender and a respondent’s anger score. For example, the above model would conclude that a respondent’s anger score is positively related to their tendency to indicate verbally aggressive responses ($\gamma_{Ang} = 0.054, SE = 0.016, p = 0.001$). However, the general tendency of males versus females was not quite statistically significant ($\gamma_{Male} = 0.303, SE = 0.184, p = 0.098$). These parameter estimates are interpreted on the logistic scale. For example, for the gender effect, this value indicates that the ratio of the odds of a “yes” answer for a typical male versus a typical female is equal to $e^{0.303} = 1.354$, meaning that without any other information, a male is about 1.35 times more likely to answer an item “aggressively.” This difference was not significant at the 0.05 level.

This model also gives the random effect values for every person, but recall that
these values are not a pure estimate of a respondent’s verbal aggression but are now an estimate of that person’s residual in the latent regression model.

4.4 Item covariates

IRT models including item property variables can also be run using SPSSIRT. For these models, a characteristic or characteristics of the items themselves are treated as predictors, allowing researchers to investigate whether certain types of items are more or less likely to yield certain item responses. A good example of an item covariate comes from the verbal aggression dataset. Recall that half of the items from this dataset are construed as asking respondents whether they would do a certain behavior, while the other half ask whether the respondent would want to perform a certain behavior. As a result, half of the items can be thought of as “do” items and the other half as “want” items. A basic IRT model can be used to examine whether or not the type of item is relevant when modeling the probability of a “yes” response.

The simplest 1PL IRT model with item covariates is the linear logistic test model, or LLTM (Fischer, 1973). Its linear component can be written as:

\[ \eta_{ip} = \sum_{k=1}^{K} \beta_k X_{ik} + \theta_p. \] (4.2)

For the LLTM, there no longer is a coefficient for every individual item. Instead, there is a fixed effect parameter for each different type of item property within the dataset (with a total of \( K \) different item covariates being modeled). For example, there will now be an intercept for all “do” items and an intercept for all “want” items if this model was conducted using the verbal aggression dataset. Therefore, all items with that given item property will have identical “easiness” values; however, it is now
possible to examine whether there are systematic differences between different types of items.

Contrary to the 1PL model with person covariates introduced in section 4.3, the $\theta_p$ values obtained from the LLTM represent a latent value estimate for person $p$. This value can be interpreted as the tendency for person $p$ to respond aggressively. However, for this specific model, it is assumed that a person’s $\theta_p$ score is the same regardless of the type of item being answered.

An LLTM using the want/do covariate will be performed as an example from the verbal aggression dataset. A separate categorical variable column must be created which contains the property of the item for that specific row. The verbal dataset already contains this column, which is denoted as “mode.” To conduct this LLTM, the following code would be run:

\begin{verbatim}
Rasch items = mode/response = ans/id = id.
\end{verbatim}

The code for this type of model can be a bit counterintuitive, since the “item” command does not specify the column containing the actual items, but instead specifies the item covariate column. Recall that the LLTM no longer has a parameter for every item, but instead has an intercept parameter for each type of item covariate. Hence, the item command is now used to specify the relevant item covariate. It is entirely possible to run a model with more than one item covariate. However, one of the quirks of the SPSSIRT macro is that only one item covariate can and must be listed in the “item” command. All further covariates (item or person) are then listed in the “cov” command.
4.5 Multidimensional models

The 1PL models outlined up to this point assume that all items are a function of one person characteristic. All items in the dataset are therefore related equally to this one variable. However, one can easily think of scales or tests which have within it more than one different variable - or dimension - being measured. In this type of scale, one item may be a function of one person variable, and another item may be related to a different person variable.

In the IRT framework, models that contain more than one person dimension are called multidimensional IRT models. In these models, $\theta_p$ is now a vector, $\theta_p$, with $D$ elements, where $D$ is the number of dimensions in the model. Therefore, each respondent will have a separate measurement for each distinct dimension. These person variables follow a multivariate normal distribution with means of zero and a covariance matrix $\Sigma$, where covariances can be found between dimensions.

The most straightforward multidimensional model with a logit link function can be described as follows:

$$ln(\frac{\pi_{ip}}{1-\pi_{ip}}) = \eta_{ip}$$

$$\eta_{ip} = \beta_i + Z_i \cdot \theta_p.$$  

In this notation, $Z_i$ is the $i$th row of an $I \times D$ matrix that denotes the dimensionality of item $i$. This $1 \times D$ vector will then be multiplied by person $p$’s ($D \times 1$) $\theta_p$ vector, resulting in the elimination of all terms except for the $\theta_p$ corresponding to the dimension of item $i$. This is essentially a complicated way of mathematically writing that each $\eta_{ip}$ value will be dependent on the dimension of item $i$, since the person
variable for person $p$ is different for each dimension.

Fortunately, the SPSSIRT macro can conduct multidimensional models without much trouble. The one limiting condition is that each item can only be “linked” to one dimension. Some IRT models have complicated dimensionality structures where certain items load onto more than one dimension, which is unfortunately not currently possible in the SPSSIRT tool. However, any structure in which an item is linked to one dimension is possible. All that needs to be added to the data is a new column which denotes the dimensionality of each item.

For an example using the verbal aggression dataset, assume that researchers are interested in measuring two different latent traits for every respondent: one for a respondent’s level of aggression in situations that are specifically his or her own fault (“self” items), and one for a respondent’s aggression in situations in which it is someone else’s fault (“other” items). Using a multidimensional IRT model, it is possible to get these two values for each person as well as the correlation between these two dimensions. The code for this model would not be too different compared to the other models that have already been described, and would look like this:

$$\text{Rasch items} = \text{item/response} = \text{ans/id} = \text{id/dim} = \text{situ}.$$  

The only change is the addition of the “dim” command, which specifies the column containing the dimension information for each item response (this column is called the “situ” variable in the verbal aggression data). In this case, all “self” items will load onto the “self” dimension, and all “other” items will be linked to the “other” dimension. The output for a multidimensional model will look identical to past models except for two differences. The first is that the “Person Estimates” table will
now contain two $\theta_p$ columns, one for the value of $\theta_{p1}$ and one for the value of $\theta_{p2}$. Each column will clearly label which dimension $\theta_p$ corresponds to. The other output change is in the “Random Effects” table, which now contains the variance of both dimensions as well as the covariance between the two. To calculate the correlation between dimensions (which ends up being 0.830 in the above example), simply use the formula:

$$\rho(\theta_1, \theta_2) = \frac{\text{cov}(\theta_1, \theta_2)}{\sqrt{\text{var}(\theta_1) * \text{var}(\theta_2)}}$$

(4.4)

There are many different multidimensional models that can be conducted, including more specific models such as the random weights linear logistic test model, or RWLLTM (Rijmen & De Boeck, 2002) or generalized item response tree models (De Boeck & Partchev, 2012). It should also be noted that identifiability can be limited when dimensions from multiple item covariates are included; see (De Boeck et al., 2011) for a detailed discussion regarding the identifiability of different 1PL multidimensional models.

### 4.6 Item and person covariates

Users can be very flexible with the different types of models that the SPSSIRT macro can run. Multiple person and/or item covariates can be included to answer many different research questions a user may have. Furthermore, multidimensional models can also be flexibly combined, allowing for different explanatory IRT models to be conducted both in a unidimensional as well as a multidimensional setting.
4.6.1 DIF models

One application of this model flexibility is the ability to examine for differential item functioning - or DIF (for a discussion of the different types of differential item functioning, see Holland and Wainer (2012)). DIF occurs when the effect of an item is different for two respondents who have equal ability but differ on some person property. This item difference can be modeled by examining how the specific item in question interacts with the person covariate responsible for the difference. This interaction term is the quantitative estimate for DIF.

An example of DIF is relatively simple to consider. For the verbal aggression dataset, perhaps a certain item will have a different effect on male respondents as opposed to females. Therefore, one simple $\beta_i$ for the item is no longer sufficient to describe its intercept, since it functions differently depending on the sex of the respondent. A new fixed coefficient must now be added to the model that takes this difference into account. The linear component of the simplest DIF model would appear as:

$$\eta_{ip} = \beta_i + \gamma X_p + \delta W_{ip} + \theta_p^e.$$ (4.5)

This model is interested in examining the DIF effect for only one item. Here, $\beta_i$ is the overall intercept for item $i$, $\gamma$ is the fixed coefficient corresponding to the effect of the one person covariate of interest ($X_p$), and $\theta_p^e$ is the random person effect as defined in Section 4.3. The only difference between this model and a typical latent regression IRT model with one person covariate is the addition of the $\delta$ term, which in this case is the effect of DIF. To correctly estimate DIF, the user must first create
a new indicator column variable, which we here call $W_{ip}$. $W_{ip} = 1$ if a given item response contains the exact combination we would like to test for DIF. For example, if a user hypothesizes that males will respond differently than females on, say, item 24, then $W_{ip} = 1$ for any response to item 24 by a male, but is equal to zero for any other type of response. Therefore, the $\delta$ coefficient only applies to the specific item/person combination of interest (which would be item 24 and males in the brief example). As a result, this $\delta$ estimate yields the differential item effect of interest.

To conduct a specific example, let us examine whether item 24 (the “S4WantShout” item) of the verbal aggression dataset contains DIF by sex. If DIF is present, it would suggest that males and females of the same overall aggression tendency (equal $\theta_p$ values) would expect to have different probabilities of answering with a “yes” to this item, due to the interaction between their gender and this specific item. Before we conduct any IRT analyses, a new column must be constructed, which will be called $w$. This column is equal to 1 in the dataset only specifically when item 24 is being answered by a male. The column is equal to zero for every other observation. After this column is added, the following code can run a model that investigates DIF:

```
Rasch items = item/cov = Gender w/response = ans/id = id.
```

Notice that $w$ is included as a fixed effect within the model. The output for this model will provide $\beta_i$ and $\theta_p$ estimates for every item and person in the dataset, respectively. It will also provide the global effect estimate of sex, which applies to responses of all items. The last fixed effect that will be presented is the effect of $\delta$, which is the DIF coefficient of interest. In this specific example, the estimate for DIF is $\delta = -0.704, SE = 0.343, p = 0.040$. This estimate signifies the difference in the
item 24 parameter between males and females. Since this coefficient is significant, it suggests that DIF is present for this item, meaning this item behaves differently depending on a respondent’s sex. Specifically, we can make the claim that for whatever reason, males respond significantly less aggressively to item 24 compared to the typical expected male response on all other items.
Chapter 5: Item and Person Fit Statistics

A beneficial property of an IRT model is that it is possible to assess the extent to which an item or a respondent “fits” the model that has been estimated. For example, if 99 out of 100 items on an exam are based on mathematics, and the last item is about a completely irrelevant topic, then this item will not fit with the rest of the model, since the response pattern will not be consistent with what the model expects. Similarly, if 99 respondents all take an exam seriously, but one respondent answers questions completely at random, then this respondent will not fit the model either. Therefore, it is possible to examine both the item and the person fit of a model, and there are a plethora of approaches to do so. The following statistics will be given by default in the SPSSIRT macro output, but will not be given if the command “fit = NO” is specified.

5.1 Person Fit

To start, the measures of person fit that SPSSIRT provides will be discussed first. Person fit statistics are used to examine how well a subject’s responses match the IRT model’s predicted responses for that given subject. There are many possible reasons as to why a respondent does not adequately fit into the model. For example, person fit statistics have been used to determine whether certain respondents have
been adequately paying attention and not guessing during an examination (Wright, 1977).

There are many different statistics that attempt to measure person fit, or a person’s “appropriateness measure.” Meijer and Sijtsma (2001) thoroughly describe many of these options, including the statistic that is included into SPSSIRT. The macro uses Drasgow’s $Z_3$, or $l_z$, person fit statistic, introduced in Drasgow, Levine, and Williams (1985). This statistic compares the overall likelihood of a person’s responses to the expected likelihood of these responses, and standardizes them onto an approximately standard normal scale. To do this, the observed likelihood of person $p$’s responses (given the model) must be estimated with the following likelihood equation:

$$L_p|\theta_p = \sum_{i=1}^{I}\{[Y_{ip}\ln(\hat{\pi}_{ip})] + [(1 - Y_{ip})\ln(1 - \hat{\pi}_{ip})]\}. \quad (5.1)$$

This likelihood value is obtained by summing over all item responses for a given respondent. $\hat{\pi}_{ip}$ is the estimated probability of a “successful” response by person $p$ on item $i$, according to the fitted 1PL model. $Y_{ip}$ is the actual observed response by person $p$ on item $i$, so it is either “1” or “0.” Once the observed likelihood is calculated, the expected likelihood, $E[L_p|\theta_p]$, must be calculated with the following formula:

$$E[L_p|\theta_p] = \sum_{i=1}^{I}\{[\hat{\pi}_{ip}\ln(\hat{\pi}_{ip})] + [(1 - \hat{\pi}_{ip})\ln(1 - \hat{\pi}_{ip})]\}. \quad (5.2)$$

One can measure the difference between the observed and expected likelihood by observing $L_p$ and $E[L_p]$. However, this difference must be standardized so that all
person fit statistics are on the same scale. This standardized fit statistic is known as $l_z$, or $Z_3$, and is calculated as follows:

$$Z_{3p} = \frac{L_p - E[L_p]}{SD[L_p]},$$

where

$$SD[L_p] = \sum_{i=1}^{I} \sqrt{\hat{\pi}_{ip}(1 - \hat{\pi}_{ip}) \times \ln(\frac{\hat{\pi}_{ip}}{1 - \hat{\pi}_{ip}})}.$$  \hspace{1cm} (5.4)

These formulas result in a $Z_3$ person fit statistic for each respondent. The $Z_3$ statistic is approximately distributed with a standard unit normal distribution (mean of zero and variance on one) (Drasgow et al., 1985). The output of the SPSSIRT macro provides the $Z_3$ person fit statistic for every respondent. This is given in the “$Z_3$ Person Fit” column within the “Person Statistics” table. A nice component of this statistic is that it can provide two separate types of model misfit. If a $Z_3$ statistic is too negative, it suggests that the response pattern actually observed from the data is inconsistent with the expected pattern. This is more consistent with the typical definition of misfit, and suggests that this respondent does not “fit” into the 1PL model being run. However, if a person has a $Z_3$ value greater than zero, it suggests that this respondent answered items more in concordance to the model than expected. Therefore, both the magnitude and the direction of the Drasgow $Z_3$ statistic provides important fit information.

Despite the multitude of person fit options to choose from, the $Z_3$ person fit statistic was chosen due to its frequent use, straightforward calculation, and interpretable
results. However, there are some drawbacks to the $Z_3$ statistic. Molenaar and Hoijtink (1990) explain some issues regarding the assumption of standard normality that is made with this statistic. It is shown that the distribution holds only if the actual, population $\theta_p$ values are used in the calculation of the statistic. Of course, researchers do not have access to the true population values. When the estimated $\theta_p$ values are used, Molenaar and Hoijtink (1990) shows that there can be serious issues when testing hypotheses with $Z_3$ values. Because of this, we recommend that users do not become reliant on “cut-offs” based on the normal distribution to determine whether a person’s misfit is “significant” or not. The statistic gives evidence of a respondent’s general misfit, but unfortunately because of these distributional issues it is difficult to test formal hypotheses of this misfit.

5.2 Item Fit

Like with respondents, it is also possible to obtain an estimate of misfit for each item of a test. These statistics can provide researchers with information on how certain items are related to others within a test. The main statistics provided in the SPSSIRT output are the outfit and infit statistics, as presented in Wright and Masters (1982) and Smith, Schumaker, and Bush (1998). Both statistics give an idea as to how well an item fits into the model. They also both are presented in two different manners: the mean-squares approach, which is distributed approximately with a chi-squared distribution with one degree of freedom, and a standardized version of the mean squares approach, which approximately follows a standard normal distribution. The SPSSIRT output provides all four of these values for every item: the mean-squared version of both outfit and infit (denoted as “Outfit (Chi)” and “Infit (Chi)”
in the output) and the standardized version of both of these statistics (denoted as “Outfit (t)” and “Infit (t)” in the output).

The outfit of an item deals with the average squared residual for every response to a given item. The mean squares (unstandardized) version of outfit for item $i$ can be calculated with:

$$outfit_i = \frac{1}{P} \sum_{p=1}^{P} \frac{(Y_{ip} - \hat{\pi}_{ip})^2}{w_{ip}}. \quad (5.5)$$

Here, the squared residual of an item response is calculated with $(Y_{ip} - \hat{\pi}_{ip})^2$ and is weighted by the variance of an item response, $w_{ip}$, where

$$w_{ip} = \hat{\pi}_{ip}(1 - \hat{\pi}_{ip}). \quad (5.6)$$

Therefore, the higher an item’s value of unstandardized outfit, the higher the typical squared residual for that item, meaning the item does not fit the estimated model well. The distribution of an outfit value is approximately chi-squared with an expected value of one, although this distribution and its critical values for hypothesis testing is reliant on factors such as the sample size, range of item parameters, and other factors. To standardize the outfit onto a roughly standard normal scale, the standard deviation of the mean squares outfit is needed:

$$SD[outfit_i] = \sqrt{\frac{\sum_{p=1}^{P} \frac{1}{w_{ip}} - 4P}{P}}, \quad (5.7)$$

where $P$ is the number of respondents in the dataset. Using this standard deviation (which will be denoted as $SD_i$) and the unstandardized outfit of item $i$ (which will be denoted as $outfit_i$), the standardized version of outfit can be calculated by:
This will provide the user with a standardized outfit value for every item. The further a standardized outfit value gets away from zero in the positive direction, the more unpredictable this item's responses were under the model, and the worse fit the item has. Although the distribution is only approximately normally distributed (and hence formal hypothesis tests of fit are not encouraged), one can start becoming concerned about the fit of an item once a standardized outfit value becomes higher than 2 or 3, although strict cut-off rule-of-thumbs should be avoided.

Outfit statistics are quite volatile to “extremely” unlikely observations. These outlier observations, such as a highly skilled respondent missing an easy item or a below-average respondent correctly answering a very difficult problem, can lead to outfit statistics that are extremely high. Therefore, if the user is interested in finding highly aberrant responses, then outfit is a decent statistic to use. However, if a researcher is more interested in how the item tends to perform across all responses and does not want a fit statistic to be skewed by outliers, then the infit of an item is probably preferred.

The infit of an item uses very similar formulas compared to outfit, and contains both a mean squared, chi-squared version as well as a standardized, approximately normally distributed version. The difference is that the overall fit for an item is now weighted by the sum of the $w_{ip}$ values, such that

$$
  infit_i = \frac{\sum_{p=1}^{P} (Y_{ip} - \hat{\pi}_{ip})^2}{\sum_{p=1}^{P} w_{ip}}.
$$

(5.9)
Like the unstandardized outfit statistic, the mean squares infit statistic (sometimes called the “weighted” total mean squares, since it weighs the squared residuals by the sum of $w_{ip}$) follows an approximate chi-squared distribution and has an expected value of one. Values greater than one indicate that the squared residuals between the expected and observed item responses are higher than they would expect to be under the model, suggesting the item may not be a good fit for the model.

To standardize the infit statistic, the standard deviation of the mean squares infit must be used:

$$SE[infit_i] = \sqrt{\frac{\sum_{p=1}^{P} w_{ip} - 4 \sum_{p=1}^{P} w_{ip}^2}{\sum_{p=1}^{P} w_{ip}}}.$$  \hspace{1cm} (5.10)

Using these equations, the standardized infit statistic for item $i$ can be calculated using equation 5.8 like with outfit. This standardized infit value will also be approximately normally distributed with a mean of zero and variance of one. The infit item statistics will not be as vulnerable to outliers, and will tend to provide a more stable estimate of an item’s performance across all responses to that item. In the SPSSIRT output, the infit values can be found in the “Item Fit Statistics” table under the “Infit (Chi)” and “Infit (t)” columns.

The infit and outfit statistics are not without their own set of issues (Karabatsos, 2000). The statistics are calculated based off the residuals of the model, meaning the statistics are dependent on the data sample. Since the parameter estimates being used to calculate these statistics are directly measured from the data, the degree to which the IRT model as a whole fits the data cannot truly be examined with outfit
and infit values. Because we are using sample values, there are issues in the distributional assumptions as well. For example, it is unwise to use stringent distributional cut-offs as critical values for infit and outfit statistics, since these critical values can in reality be influenced by factors such as the sample size (Smith et al., 1998; Karabatsos, 2000). It should also be noted that infit and outfit statistics were originally conducted within models that utilized conditional maximum likelihood estimation (Anderson, 1973), not the PQL estimation method that SPSS uses (see Section 6.1 for estimation details). Therefore, further investigation into the validity of infit and outfit under these conditions is warranted.

The $Z_3$ statistic discussed earlier is also provided as an item fit statistic in the output. Reise (1990) describes how the Drasgow $Z_3$ statistic, which is traditionally a person fit statistic, can be successfully applied as an item fit statistic as well. It was shown that it performs as well as or better than other traditional item fit methods such as Bock’s chi-squared approach (Bock, 1972) or Yen’s $Q_1$ statistic (Yen, 1981).

Item and person fit statistics can show valuable information about the measurement process of a test or scale. In the SPSSIRT macro, the $Z_3$ statistic is provided for both persons and items, while the (standardized and unstandardized) infit and outfit values are shown for all items.
Chapter 6: Estimation Methods and Comparisons with other Programs

6.1 Estimation methods

In general, SPSS uses a penalized quasi-likelihood (or PQL) approach to estimating generalized linear mixed models. This approach (Breslow & Clayton, 1993) approximates the model’s likelihood by using Taylor expansion principles to obtain a linearization of the likelihood. This “pseudo-” or “quasi-likelihood” is then maximized to obtain estimates for the fixed and random effects. Methods using quasi-likelihood techniques have shown evidence of having slightly biased parameter estimates when used in an IRT context (Breslow & Clayton, 1993; Breslow & Lin, 1995). The bias appears to result in fixed effect estimates being less extreme (closer to zero) than they should be, especially when the estimates themselves are much greater or much less than zero.

To assess the validity of the results obtained from the SPSSIRT macro, the findings from SPSS were compared to results from identical models in the R packages lme4 and TAM, which use different methods of estimation. The lme4 package utilizes the Laplace estimation technique (Doran, Bates, Bliese, & Dowling, 2007), which uses Laplace approximation to approximate the likelihood, which is then maximized. Like
the PQL approach that SPSS uses, this is an approximate estimation method, but has been shown to provide reasonable estimates for a variety of IRT models (De Boeck et al., 2011). The TAM package estimates IRT models using marginal maximum likelihood procedures (De Leeuw & Verhelst, 1986), which do not approximate the likelihood like SPSS and lme4 do. For the scoring step, when the random effect(s) for a person is obtained, SPSS and lme4 both use MAP (maximum a posteriori) techniques, while TAM uses EAP (expected a posteriori).

6.2 Model comparisons

Various models were conducted in SPSS, lme4, and TAM in order to compare estimated results. For all three packages, the basic 1PL model, the latent regression model, and a multidimensional model were conducted. On both SPSS and lme4, an LLTM and a model estimating DIF were run (these models were not readily accessible on the TAM package). For all models, the item parameters, person/item covariate effects (and their respective hypothesis tests), the specific person random effects, and the variances/covariances of random effects were compared.

6.2.1 1PL model

To start, the basic 1PL model was run on SPSS, lme4, and TAM using the verbal aggression data. Both the fixed item parameter estimates and the random person variables were compared. The estimates in Table 6.1 show how similar the item parameter estimates were. Visually, this pattern can be seen in Figure 6.1. There did
appear to be some differences as an item parameter became further away from zero; the SPSS estimates were slightly less extreme, or closer to zero, in these instances. Still, there are no discrepancies here that produce much concern. The $\theta_p$ values from this model are shown in Figure 6.2. Similar to the conclusions reached from Table 6.1, the person values from the three softwares are almost identical, except for some slight differences when the estimates get further away from zero. The one area of most potential concern between SPSS and the R packages are the estimates of the random effect variance; the estimates were 1.696 ($1.302^2$) for SPSS, 1.902 ($1.379^2$) for lme4, and 1.918 ($1.385^2$) for TAM.

Figure 6.1: Item intercept ($\beta_i$) estimates calculated from identical basic 1PL models estimated in SPSS, lme4, and TAM.
Figure 6.2: $\theta_p$ estimates calculated from identical basic 1PL models estimated in both SPSS and lme4.

6.2.2 Person covariates

Next, estimates were compared from the latent regression model described in section 4.3. This specific example will include both the anger score and gender covariates. Results suggest that the fixed and random effects from SPSS are also quite similar to the R package results for this type of model. Figure 6.3 shows the close relationship between the $\theta_p^*$ values calculated from the two software programs. The estimated variance of this random effect is 1.633 $(1.278^2)$ for SPSS, 1.817 $(1.348^2)$ for lme4, and 1.819 $(1.349^2)$ for TAM. The fixed effects also appeared to be very similar, as shown in Table 6.2. One potential concern is the fact that for the model conducted in TAM, the
<table>
<thead>
<tr>
<th>Item</th>
<th>SPSS $\beta$ (SE)</th>
<th>lme4 $\beta$ (SE)</th>
<th>TAM $\beta$ (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.167 (0.158)</td>
<td>1.229 (0.162)</td>
<td>1.221 (0.142)</td>
</tr>
<tr>
<td>2</td>
<td>0.373 (0.149)</td>
<td>0.397 (0.153)</td>
<td>0.390 (0.132)</td>
</tr>
<tr>
<td>3</td>
<td>-0.832 (0.153)</td>
<td>-0.864 (0.158)</td>
<td>-0.870 (0.137)</td>
</tr>
<tr>
<td>4</td>
<td>1.167 (0.158)</td>
<td>1.229 (0.162)</td>
<td>1.221 (0.142)</td>
</tr>
<tr>
<td>5</td>
<td>0.540 (0.150)</td>
<td>0.573 (0.154)</td>
<td>0.565 (0.133)</td>
</tr>
<tr>
<td>6</td>
<td>0.077 (0.148)</td>
<td>0.088 (0.152)</td>
<td>0.081 (0.131)</td>
</tr>
<tr>
<td>7</td>
<td>0.834 (0.153)</td>
<td>0.879 (0.157)</td>
<td>0.873 (0.136)</td>
</tr>
<tr>
<td>8</td>
<td>-0.054 (0.148)</td>
<td>-0.049 (0.152)</td>
<td>-0.056 (0.131)</td>
</tr>
<tr>
<td>9</td>
<td>-1.416 (0.164)</td>
<td>-1.473 (0.169)</td>
<td>-1.481 (0.150)</td>
</tr>
<tr>
<td>10</td>
<td>1.669 (0.169)</td>
<td>1.757 (0.175)</td>
<td>1.749 (0.156)</td>
</tr>
<tr>
<td>11</td>
<td>0.677 (0.151)</td>
<td>0.715 (0.155)</td>
<td>0.708 (0.134)</td>
</tr>
<tr>
<td>12</td>
<td>0.012 (0.148)</td>
<td>0.019 (0.152)</td>
<td>0.012 (0.131)</td>
</tr>
<tr>
<td>13</td>
<td>-0.201 (0.148)</td>
<td>-0.202 (0.152)</td>
<td>-0.210 (0.131)</td>
</tr>
<tr>
<td>14</td>
<td>-1.438 (0.165)</td>
<td>-1.495 (0.169)</td>
<td>-1.504 (0.150)</td>
</tr>
<tr>
<td>15</td>
<td>-2.840 (0.224)</td>
<td>-2.969 (0.232)</td>
<td>-2.975 (0.215)</td>
</tr>
<tr>
<td>16</td>
<td>0.507 (0.149)</td>
<td>0.537 (0.154)</td>
<td>0.530 (0.133)</td>
</tr>
<tr>
<td>17</td>
<td>-0.656 (0.151)</td>
<td>-0.678 (0.155)</td>
<td>-0.686 (0.135)</td>
</tr>
<tr>
<td>18</td>
<td>-1.459 (0.165)</td>
<td>-1.519 (0.170)</td>
<td>-1.526 (0.151)</td>
</tr>
<tr>
<td>19</td>
<td>0.677 (0.151)</td>
<td>0.715 (0.155)</td>
<td>0.708 (0.134)</td>
</tr>
<tr>
<td>20</td>
<td>-0.367 (0.149)</td>
<td>-0.377 (0.153)</td>
<td>-0.383 (0.132)</td>
</tr>
<tr>
<td>21</td>
<td>-1.911 (0.179)</td>
<td>-1.991 (0.184)</td>
<td>-1.999 (0.166)</td>
</tr>
<tr>
<td>22</td>
<td>1.034 (0.155)</td>
<td>1.090 (0.160)</td>
<td>1.082 (0.140)</td>
</tr>
<tr>
<td>23</td>
<td>-0.333 (0.149)</td>
<td>-0.341 (0.153)</td>
<td>-0.349 (0.132)</td>
</tr>
<tr>
<td>24</td>
<td>-0.998 (0.156)</td>
<td>-1.037 (0.160)</td>
<td>-1.043 (0.140)</td>
</tr>
</tbody>
</table>

Table 6.1: Item intercept ($\beta_i$) estimates (and their standard errors) for identical models conducted in SPSS, lme4, and TAM respectively. Here, the basic 1PL model was applied to the verbal aggression dataset.
Figure 6.3: $\theta_p$ estimates calculated from identical latent regression models estimated in SPSS, lme4 and TAM.

standard error estimate for the covariate regression coefficients and item intercepts were much lower than for SPSS and lme4. However, TAM appears to underestimate standard errors for fixed parameter estimates for this model. Explorations into other software resulting in standard errors for the intercept and regression coefficient standard error estimates much closer to those from SPSS and lme4, not like those from TAM.

6.2.3 Item covariates

Item response theory models containing item covariates were compared next. Specifically, a linear logistic test model (LLTM) with the situation item covariate
<table>
<thead>
<tr>
<th>Item</th>
<th>SPSS $\beta$ (SE)</th>
<th>lme4 $\beta$ (SE)</th>
<th>TAM $\beta$ (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.013 (0.360)</td>
<td>0.037 (0.376)</td>
<td>-0.002 (0.145)</td>
</tr>
<tr>
<td>2</td>
<td>-0.781 (0.357)</td>
<td>-0.792 (0.373)</td>
<td>-0.834 (0.135)</td>
</tr>
<tr>
<td>3</td>
<td>-1.988 (0.360)</td>
<td>-2.052 (0.376)</td>
<td>-2.093 (0.140)</td>
</tr>
<tr>
<td>4</td>
<td>0.013 (0.360)</td>
<td>0.040 (0.376)</td>
<td>-0.002 (0.145)</td>
</tr>
<tr>
<td>5</td>
<td>-0.614 (0.357)</td>
<td>-0.617 (0.373)</td>
<td>-0.658 (0.136)</td>
</tr>
<tr>
<td>6</td>
<td>-1.077 (0.360)</td>
<td>-1.106 (0.373)</td>
<td>-1.143 (0.134)</td>
</tr>
<tr>
<td>7</td>
<td>-0.320 (0.358)</td>
<td>-0.308 (0.374)</td>
<td>-0.351 (0.139)</td>
</tr>
<tr>
<td>8</td>
<td>-1.208 (0.357)</td>
<td>-1.238 (0.373)</td>
<td>-1.279 (0.134)</td>
</tr>
<tr>
<td>9</td>
<td>-2.572 (0.365)</td>
<td>-2.660 (0.381)</td>
<td>-2.703 (0.153)</td>
</tr>
<tr>
<td>10</td>
<td>0.517 (0.365)</td>
<td>0.563 (0.381)</td>
<td>0.525 (0.159)</td>
</tr>
<tr>
<td>11</td>
<td>-0.477 (0.358)</td>
<td>-0.472 (0.373)</td>
<td>-0.516 (0.137)</td>
</tr>
<tr>
<td>12</td>
<td>-1.143 (0.360)</td>
<td>-1.170 (0.373)</td>
<td>-1.211 (0.134)</td>
</tr>
<tr>
<td>13</td>
<td>-1.356 (0.357)</td>
<td>-1.391 (0.373)</td>
<td>-1.434 (0.134)</td>
</tr>
<tr>
<td>14</td>
<td>-2.593 (0.365)</td>
<td>-2.686 (0.382)</td>
<td>-2.726 (0.154)</td>
</tr>
<tr>
<td>15</td>
<td>-3.998 (0.398)</td>
<td>-4.156 (0.415)</td>
<td>-4.196 (0.219)</td>
</tr>
<tr>
<td>16</td>
<td>-0.648 (0.357)</td>
<td>-0.652 (0.373)</td>
<td>-0.694 (0.136)</td>
</tr>
<tr>
<td>17</td>
<td>-1.811 (0.359)</td>
<td>-1.864 (0.375)</td>
<td>-1.909 (0.138)</td>
</tr>
<tr>
<td>18</td>
<td>-2.615 (0.366)</td>
<td>-2.704 (0.382)</td>
<td>-2.748 (0.154)</td>
</tr>
<tr>
<td>19</td>
<td>-0.477 (0.358)</td>
<td>-0.474 (0.374)</td>
<td>-0.516 (0.137)</td>
</tr>
<tr>
<td>20</td>
<td>-1.522 (0.358)</td>
<td>-1.564 (0.374)</td>
<td>-1.607 (0.135)</td>
</tr>
<tr>
<td>21</td>
<td>-3.067 (0.373)</td>
<td>-3.186 (0.389)</td>
<td>-3.221 (0.169)</td>
</tr>
<tr>
<td>22</td>
<td>-0.119 (0.359)</td>
<td>-0.099 (0.375)</td>
<td>-0.141 (0.143)</td>
</tr>
<tr>
<td>23</td>
<td>-1.488 (0.358)</td>
<td>-1.534 (0.373)</td>
<td>-1.572 (0.135)</td>
</tr>
<tr>
<td>24</td>
<td>-2.153 (0.361)</td>
<td>-2.224 (0.377)</td>
<td>-2.266 (0.143)</td>
</tr>
</tbody>
</table>

(a) Item intercept ($\beta_i$) estimates (and their standard errors) for identical models conducted in SPSS, lme4, and TAM, respectively. Here, the latent regression 1PL IRT model, with a participant’s anger score and gender as covariates, was applied to the verbal aggression dataset.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\gamma$</th>
<th>SE</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger (SPSS)</td>
<td>0.054</td>
<td>0.016</td>
<td>0.001</td>
</tr>
<tr>
<td>Male (SPSS)</td>
<td>0.303</td>
<td>0.184</td>
<td>0.098</td>
</tr>
<tr>
<td>Anger (lme4)</td>
<td>0.055</td>
<td>0.017</td>
<td>0.001</td>
</tr>
<tr>
<td>Male (lme4)</td>
<td>0.322</td>
<td>0.193</td>
<td>0.095</td>
</tr>
<tr>
<td>Anger (TAM)</td>
<td>0.057</td>
<td>0.004</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Male (TAM)</td>
<td>0.321</td>
<td>0.169</td>
<td>0.058</td>
</tr>
</tbody>
</table>

(b) The person covariate effect ($\gamma$) estimates of the anger and gender covariates, estimated on SPSS, lme4, and TAM.

Table 6.2: Item intercept and covariate effects from SPSS, lme4, and TAM.
Figure 6.4: $\theta_p$ estimates calculated from identical LLTM models estimated in both SPSS and lme4.

(Other person’s fault vs. the respondent’s own fault) was run on both SPSS and lme4 (the LLTM model was not conducted on TAM, since models with item covariates are not readily accessible in this package). The same pattern emerges when estimates from these models are compared: SPSS estimates continue to be very close to the values of other software, but tend to trend towards zero, especially as the values become more extreme. The item covariate intercepts (one for “other” items and one for “self” items) can be found on Table 6.3 and the person random effect values from this model can be seen on Figure 6.4. For this LLTM, the random effect variance estimate for SPSS was 1.164 ($1.079^2$) and was 1.289 ($1.135^2$) from lme4.
### 6.2.4 Multidimensional model

Next, models involving multiple dimensions were compared across programs. For the following example, the model being conducted involves two dimensions - one for “want” items and one for “do” items, meaning the model will obtain two estimates for each respondent: one for that respondent’s tendency to want to be verbally aggressive and one for the respondent’s tendency to actually do verbally aggressive actions.

Much like the previous examples, the fixed effects and random effects were very similar across the different programs. As you can see on Table 6.4, the item intercept estimates for each of the three packages are all relatively similar, following the same consistent pattern found in earlier comparisons. Figures 6.5 and 6.6 show the random effects for the “do” and “want” dimensions, respectively. These plots indicate that even for multidimensional models, the SPSSIRT macro provides random effect values almost identical to those from other packages. As for the estimated covariance matrix of the random effects for this model, SPSS estimated the variance of the “do” dimension to be 2.252 (1.501²), while the same parameter was estimated to be 2.762 (1.662²) on lme4 and 2.804 (1.675²) from TAM. For the “want” dimension, the estimate from SPSS was 1.735 (1.317²), while the R packages produced 2.059 (1.435²)
and 2.098 (1.448^2) for lme4 and TAM, respectively. The correlation between the “do” and “want” dimensions was also calculated for all three software options. For SPSS, this correlation was estimated to be 0.789, which is very similar to the lme4 estimate (0.790) and the TAM estimate (0.775).

### 6.2.5 DIF model

The last comparison presented is a comparison of models which include an estimate of differential item functioning, or DIF, between a particular item and, in this case, gender. The example presented in section 4.6 will again be used here; it explores the possible presence of DIF in the “S4WantShout” item between genders. In SPSS,
Table 6.4: Item intercept $\beta_i$ estimates (and their standard errors) for identical multidimensional models conducted in SPSS, lme4, and TAM, respectively.

<table>
<thead>
<tr>
<th>Item</th>
<th>SPSS $\beta$ (SE)</th>
<th>lme4 $\beta$ (SE)</th>
<th>TAM $\beta$ (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.224 (0.167)</td>
<td>1.336 (0.178)</td>
<td>1.322 (0.150)</td>
</tr>
<tr>
<td>2</td>
<td>0.380 (0.158)</td>
<td>0.420 (0.166)</td>
<td>0.402 (0.138)</td>
</tr>
<tr>
<td>3</td>
<td>-0.884 (0.162)</td>
<td>-0.954 (0.171)</td>
<td>-0.968 (0.142)</td>
</tr>
<tr>
<td>4</td>
<td>1.161 (0.158)</td>
<td>1.261 (0.166)</td>
<td>1.250 (0.144)</td>
</tr>
<tr>
<td>5</td>
<td>0.538 (0.150)</td>
<td>0.592 (0.157)</td>
<td>0.581 (0.134)</td>
</tr>
<tr>
<td>6</td>
<td>0.076 (0.148)</td>
<td>0.097 (0.154)</td>
<td>0.086 (0.132)</td>
</tr>
<tr>
<td>7</td>
<td>0.870 (0.162)</td>
<td>0.948 (0.172)</td>
<td>0.935 (0.144)</td>
</tr>
<tr>
<td>8</td>
<td>-0.070 (0.157)</td>
<td>-0.074 (0.165)</td>
<td>-0.087 (0.137)</td>
</tr>
<tr>
<td>9</td>
<td>-1.489 (0.172)</td>
<td>-1.608 (0.183)</td>
<td>-1.622 (0.154)</td>
</tr>
<tr>
<td>10</td>
<td>1.661 (0.170)</td>
<td>1.797 (0.179)</td>
<td>1.786 (0.157)</td>
</tr>
<tr>
<td>11</td>
<td>0.674 (0.151)</td>
<td>0.736 (0.158)</td>
<td>0.727 (0.136)</td>
</tr>
<tr>
<td>12</td>
<td>0.011 (0.148)</td>
<td>0.026 (0.154)</td>
<td>0.016 (0.132)</td>
</tr>
<tr>
<td>13</td>
<td>-0.225 (0.159)</td>
<td>-0.242 (0.165)</td>
<td>-0.255 (0.137)</td>
</tr>
<tr>
<td>14</td>
<td>-1.511 (0.173)</td>
<td>-1.629 (0.183)</td>
<td>-1.646 (0.155)</td>
</tr>
<tr>
<td>15</td>
<td>-2.956 (0.232)</td>
<td>-3.190 (0.248)</td>
<td>-3.207 (0.222)</td>
</tr>
<tr>
<td>16</td>
<td>0.504 (0.150)</td>
<td>0.553 (0.156)</td>
<td>0.545 (0.134)</td>
</tr>
<tr>
<td>17</td>
<td>-0.655 (0.152)</td>
<td>-0.687 (0.158)</td>
<td>-0.699 (0.136)</td>
</tr>
<tr>
<td>18</td>
<td>-1.457 (0.166)</td>
<td>-1.547 (0.174)</td>
<td>-1.560 (0.153)</td>
</tr>
<tr>
<td>19</td>
<td>0.703 (0.160)</td>
<td>0.769 (0.169)</td>
<td>0.752 (0.141)</td>
</tr>
<tr>
<td>20</td>
<td>-0.399 (0.158)</td>
<td>-0.422 (0.166)</td>
<td>-0.443 (0.137)</td>
</tr>
<tr>
<td>21</td>
<td>-1.999 (0.187)</td>
<td>-2.152 (0.198)</td>
<td>-2.172 (0.171)</td>
</tr>
<tr>
<td>22</td>
<td>1.029 (0.156)</td>
<td>1.119 (0.163)</td>
<td>1.108 (0.141)</td>
</tr>
<tr>
<td>23</td>
<td>-0.334 (0.149)</td>
<td>-0.343 (0.155)</td>
<td>-0.354 (0.133)</td>
</tr>
<tr>
<td>24</td>
<td>-0.997 (0.156)</td>
<td>-1.055 (0.163)</td>
<td>-1.065 (0.142)</td>
</tr>
</tbody>
</table>
Figure 6.6: $\theta_p$ estimates for the “want” dimension for SPSS, lme4, and TAM models.

it has already been shown that the estimate for this DIF effect is -0.704 (SE = 0.343). When this identical model is conducted in lme4 (this model is not readily accessible in TAM), the resulting estimate for DIF is -0.731 (SE = 0.350). The fixed effects for these models can be found in Table 6.5 and the person random effects are shown in Figure 6.7.

6.3 Summary

To assess the validity of the SPSSIRT macro, estimates from SPSSIRT were compared to values from various IRT models conducted in the R package lme4 and, when
(a) Item intercept ($\beta_i$) estimates (and their standard errors) for identical models conducted in SPSS and lme4. Here, a 1PL IRT model investigating whether a DIF effect is present between the gender of a respondent and the “S4WantShout” item.

<table>
<thead>
<tr>
<th>Item</th>
<th>SPSS $\beta$ (SE)</th>
<th>lme4 $\beta$ (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.094 (0.163)</td>
<td>1.160 (0.168)</td>
</tr>
<tr>
<td>2</td>
<td>0.299 (0.155)</td>
<td>0.323 (0.159)</td>
</tr>
<tr>
<td>3</td>
<td>-0.907 (0.159)</td>
<td>-0.937 (0.164)</td>
</tr>
<tr>
<td>4</td>
<td>1.094 (0.163)</td>
<td>1.155 (0.168)</td>
</tr>
<tr>
<td>5</td>
<td>0.467 (0.156)</td>
<td>0.498 (0.160)</td>
</tr>
<tr>
<td>6</td>
<td>0.003 (0.154)</td>
<td>0.015 (0.158)</td>
</tr>
<tr>
<td>7</td>
<td>0.761 (0.158)</td>
<td>0.805 (0.163)</td>
</tr>
<tr>
<td>8</td>
<td>-0.128 (0.154)</td>
<td>-0.123 (0.158)</td>
</tr>
<tr>
<td>9</td>
<td>-1.492 (0.170)</td>
<td>-1.548 (0.175)</td>
</tr>
<tr>
<td>10</td>
<td>1.597 (0.175)</td>
<td>1.682 (0.180)</td>
</tr>
<tr>
<td>11</td>
<td>0.603 (0.159)</td>
<td>0.642 (0.161)</td>
</tr>
<tr>
<td>12</td>
<td>-0.062 (0.154)</td>
<td>-0.055 (0.158)</td>
</tr>
<tr>
<td>13</td>
<td>-0.275 (0.154)</td>
<td>-0.277 (0.159)</td>
</tr>
<tr>
<td>14</td>
<td>-1.513 (0.170)</td>
<td>-1.570 (0.175)</td>
</tr>
<tr>
<td>15</td>
<td>-2.917 (0.223)</td>
<td>-3.041 (0.237)</td>
</tr>
<tr>
<td>16</td>
<td>0.433 (0.155)</td>
<td>0.463 (0.160)</td>
</tr>
<tr>
<td>17</td>
<td>-0.730 (0.157)</td>
<td>-0.755 (0.162)</td>
</tr>
<tr>
<td>18</td>
<td>-1.535 (0.171)</td>
<td>-1.593 (0.176)</td>
</tr>
<tr>
<td>19</td>
<td>0.603 (0.157)</td>
<td>0.643 (0.161)</td>
</tr>
<tr>
<td>20</td>
<td>-0.441 (0.155)</td>
<td>-0.451 (0.159)</td>
</tr>
<tr>
<td>21</td>
<td>-1.986 (0.184)</td>
<td>-2.062 (0.190)</td>
</tr>
<tr>
<td>22</td>
<td>0.961 (0.161)</td>
<td>1.019 (0.166)</td>
</tr>
<tr>
<td>23</td>
<td>-0.408 (0.155)</td>
<td>-0.416 (0.159)</td>
</tr>
<tr>
<td>24</td>
<td>-0.909 (0.176)</td>
<td>-0.940 (0.181)</td>
</tr>
</tbody>
</table>

(b) The covariate effect ($\gamma$) estimate of gender and the DIF effect ($\delta$) estimate for the “S4WantShout” item (item 24), estimated on SPSS and lme4.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\gamma/\delta$</th>
<th>SE</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male (SPSS)</td>
<td>0.320</td>
<td>0.187</td>
<td>0.087</td>
</tr>
<tr>
<td>DIF (SPSS)</td>
<td>-0.704</td>
<td>0.343</td>
<td>0.040</td>
</tr>
<tr>
<td>Male (lme4)</td>
<td>0.328</td>
<td>0.196</td>
<td>0.095</td>
</tr>
<tr>
<td>DIF (lme4)</td>
<td>-0.713</td>
<td>0.350</td>
<td>0.036</td>
</tr>
</tbody>
</table>

Table 6.5: Item intercept and covariate effects from SPSS, lme4, and TAM.
possible, the R package TAM. The results from the SPSSIRT macro were quite close to estimates from the identical models conducted in the other software. Although there were some minor discrepancies in the estimates when the values got further from zero, there were no significant issues. The effectiveness of these R packages to correctly conduct these analyses is established and well-documented (De Boeck et al., 2011); it stands to reason that since SPSSIRT yielded similar output to these packages, then SPSSIRT does a reasonable job with these models, at least for this particular dataset.

Of course, further analysis must be done using different datasets and conditions
to test the effectiveness of SPSSIRT. Just because SPSSIRT was able to perform well with the verbal aggression dataset does not necessarily mean that it will provide valid results in all situations. Still, it is a positive first step.
Chapter 7: Discussion

7.1 Limitations and future directions

The SPSSIRT macro presented here provides a tool that can be usefully applied to many research settings. However, the program does have its share of limitations. Unfortunately, the macro does not currently have an option to run more complex models such as the two parameter (2PL) or three parameter (3PL) logistic IRT models. These models take into account the differing discrimination and guessing parameters for each item, something that is currently not possible in SPSSIRT.

Another shortcoming is the fact that the SPSSIRT macro presently can only run models with dichotomous outcomes, excluding models with polytomous data such as the graded response model (Samejima, 1997) or the partial credit model (Masters, 1982). However, an option to run a one-parameter version of the graded response model, which uses the multinomial logit link function, is currently being developed.

One of the most inconvenient features of the SPSSIRT output is the lack of a reliable estimate of overall model fit. Typically, the GENLINMIXED function in SPSS does provide values for AIC and BIC, however when compared to other packages, we have found that these values do not match the conclusions from other software, and are therefore not recommended for use. A future direction for the macro would
be a deeper investigation into the given AIC and BIC values, as well an exploration into other possible criteria for IRT model fit, such as the \( M_2 \) statistic, RMSEA for IRT, and other similar limited-information goodness of fit criteria (Maydeu-Olivares, 2013).

Although the results from the SPSSIRT macro were shown to be similar to output from other packages, it is important to keep in mind that this comparison was only made using one dataset. Future research should examine how generalizable the macro is with other data. Also, a simulation study that investigates the SPSSIRT estimation effectiveness in different situations would be a necessary next step. Specifically, an examination of the results when sample size, number of dimensions in the data, and other conditions are varied would be particularly insightful. However, until this type of study is done, it cannot be said with certainty how generalizable the PQL estimation results from SPSS are to different situations.

Currently, the macro is configured so that the user only needs to specify a simple line of syntax for a particular model to be run. However, SPSS has a point-and-click interface, where models can be specified by clicking and dragging variable names into specific categories. An IRT “point-and-click” option is currently being developed, which will allow users to run all the models described here without writing any code. This development should make the SPSSIRT tool even more convenient for researchers.

### 7.2 Conclusion

An SPSS macro has been introduced that has the capacity to run a multitude of item response models in an accessible, user-friendly manner. The macro can run
basic 1PL models, models with item and/or person covariates, multidimensional IRT models, DIF models, and any combination of these. The macro’s output provide item parameters, item and person covariate fixed effects, person random effect values, random effect covariance matrices, item and person fit statistics, and other options like item characteristic curves and item information plots.

When these models are conducted on SPSS, estimates appear quite similar to results from other software packages. Since SPSS uses penalized quasi-likelihood estimation techniques, the parameter estimates tend to have a non-extreme pattern, although this discrepancy is hardly noticeable except for more non-zero estimates (for the verbal aggression data at least). This macro can therefore serve as a reliable tool to conduct IRT in a convenient way. The hope is that researchers who are new to IRT or who are worried about complicated software packages can learn IRT through the SPSSIRT macro. Also, the SPSSIRT macro can be a great pedagogical outlet for instructors introducing IRT to students who are familiar with SPSS.
References


Appendix A: The SPSSIRT Macro Syntax

DEFINE Rasch (items=!charend('/'))/cov=!charend('/')
/response=!charend('/') /id=!charend('/')
/dim=!charend('/') !default(XXXXXXXX)
/difitem=!charend('/') !default(YYYYYYYY)
/difvar=!charend('/') !default(YYYYYYYY)
/link=!charend('/') !default(logit) /save=!charend('/') !default(0)
/dataset=!charend('/') !default(DataSet1)
/outdata=!charend('/') !default(coefdata)
/plot=!charend('/') !default(NO) /theta=!charend('/') !default(YES)
/center=!charend('/') !default(Person)
/iters=!charend('/') !default(100)/conf= !charend('/') !default(95)).

MATRIX.
print/title = "************ SPSSIRT ***********".
*******************************************************************
END MATRIX.

DATASET ACTIVATE !dataset.

ALTER TYPE !items(A20).

*VARIABLE LEVEL !difvar(scale).

OUTPUT MODIFY
  /REPORT PRINTREPORT=NO
  /SELECT TABLES
  /IF COMMANDS=\"Alter Type(LAST)\"
    LABELS=[EXACT\"Altered Types\"] INSTANCES=[1]
  /DELETEOBJECT DELETE=YES
  /SELECT HEADINGS
  /IF COMMANDS=\"Alter Type(LAST)\"
    LABELS=[EXACT\"Title\"] INSTANCES=[1]
/DELETEOBJECT DELETE=YES.

!IF (!difitem ^= YYYYYYY) !THEN
!IF (!difvar ^= YYYYYYY) !THEN
RECODE !items (MISSING=SYSMIS) (!difitem=1) (ELSE=0) INTO item_ind.
VARIABLE LABELS item_ind 'Item Indicator'.
EXECUTE.
COMPUTE DIF = item_ind * !difvar.
EXECUTE.

RECODE DIF (0=100) (1=101).
VARIABLE LEVEL DIF(scale).

!IFEND
!IFEND

OMS

/SELECT TABLES
/IF COMMANDS=[’Generalized Linear Mixed Models’]
SUBTYPES=[’Case Processing Summary’ , ’Classification’
’ Contrasts’ , ’Covariance Parameters Summary’ , ’Data Structure’
’ Estimates’ , ’Fixed Coefficients’ , ’Fixed Effects’ , ’Grand Means’
’ Model Summary’ , ’Notes’ , ’Overall Test’ , ’Results’
’ Random Effects’ , ’Residual Effect’]
/DESTINATION VIEWER=NO.
DATASET DECLARE !outdata.

OMS

/SELECT TABLES
/IF COMMANDS=[’Generalized Linear Mixed Models’]
SUBTYPES=[’Fixed Coefficients’]
/DESTINATION FORMAT=SAV NUMBERED=TableNumber_
OUTFILE=!outdata VIEWER=NO.

!IF (!difitem = YYYYYYY) !THEN
!IF (!difvar = YYYYYYY) !THEN
!IF (!dim = XYYYYYY) !THEN
GENLINMIXED

/FIELDS TARGET=!response TRIALS=NONE OFFSET=NONE
/TARGET_OPTIONS DISTRIBUTION=BINOMIAL LINK=!link
/FIXED EFFECTS=!items !cov USE_INTERCEPT=FALSE
DATASET ACTIVATE !outdata.
compute Var1 = substr(Var1, (char.index(Var1,'=')+1), 20).
RENAME VARIABLES (Var1 = !items).
RENAME VARIABLES (Coefficient = Beta).
EXECUTE.

OUTPUT MODIFY
 /REPORT PRINTREPORT=NO
 /SELECT TEXTS
 /IF COMMANDS="CTables(LAST)"
 LABELS=[EXACT("Active Dataset")]] INSTANCES=[1]
/DELETEOBJECT DELETE=YES.

!IF (!center = item) !THEN
COMPUTE placeholder = 1.
AGGREGATE
/BREAK=placeholder
/ave_beta = MEAN(Beta).
COMPUTE newbeta = Beta-ave_beta.
EXECUTE.
DELETE VARIABLES Beta placeholder ave_beta Lower Upper.
RENAME VARIABLES (newbeta = Beta).
!IFEND

VARIABLE LABELS Beta 'Beta'.
EXECUTE.

!IF (!center =~ item) !THEN
CTABLES
/MRSETS COUNTDUPLICATES=YES
/SMISSING VARIABLE 
/VLABELS VARIABLES=!items Beta Std.Error Lower Upper DISPLAY=LABEL
/TABLE !items [C] BY Beta [S][MEDIAN ''] + Std.Error [S][MEDIAN ''] 
+ t [S][MEDIAN ''] + Sig [S][MEDIAN ''] + Lower [S][MEDIAN ''] 
+ Upper [S][MEDIAN ''] 
/SLABELS VISIBLE=NO 
/CATEGORIES VARIABLES=!items ORDER=A KEY=VALUE EMPTY=EXCLUDE 
/TITLES 
TITLE='Fixed Effects'.
OUTPUT MODIFY
/REPORT PRINTREPORT=NO 
/SELECT HEADINGS 
/IF COMMANDS=["CTables(LAST)"] 
LABELS=[EXACT("Title") ] INSTANCES=[1] 
/DELETEOBJECT DELETE=YES.
!IFEND

!IF (!center = item) !THEN
CTABLES
/MRSETS COUNTDUPLICATES=YES 
/SMISSING VARIABLE
/VLABELS VARIABLES=!items Beta Std.Error DISPLAY=LABEL
/TABLE !items [C] BY Beta [S][MEDIAN ''] + Std.Error [S][MEDIAN ''] 
+ t [S][MEDIAN ''] + Sig [S][MEDIAN ''] 
/SLABELS VISIBLE=NO 
/CATEGORIES VARIABLES=!items ORDER=A KEY=VALUE EMPTY=EXCLUDE 
/TITLES
TITLE='Item Statistics'.
OUTPUT MODIFY
  /REPORT PRINTREPORT=NO
  /SELECT HEADINGS
  /IF COMMANDS=['CTables(LAST)']
    LABELS=[EXACT("Title")]
    INSTANCES=[1]
  
/DELETEOBJECT DELETE=YES.
!IFEND

DELETE VARIABLES TableNumber_ TO Label_ t TO Sig.
EXECUTE.

!IF (!link = logit) !THEN
DELETE VARIABLES ExpCoefficient Lower_A Upper_A.
!IFEND

SORT CASES BY !items.
VARIABLE LABELS Beta 'Beta'.
EXECUTE.

OMS
  /SELECT TABLES
  /IF COMMANDS=['Alter Type']
  /DESTINATION VIEWER=NO.
  ALTER TYPE !items(A20).
EXECUTE.

OUTPUT MODIFY
  /REPORT PRINTREPORT=NO
  /SELECT HEADINGS
  /IF COMMANDS=['Alter Type(LAST)']
    LABELS=[EXACT("Title")]
    INSTANCES=[1]
  
/DELETEOBJECT DELETE=YES.

DATASET ACTIVATE !dataset.

OUTPUT MODIFY
  /REPORT PRINTREPORT=NO
  /SELECT TEXTS
  /IF COMMANDS=['CTables(LAST)']
    LABELS=[EXACT("Active Dataset")]
    INSTANCES=[1]
  
/DELETEOBJECT DELETE=YES.
SORT CASES BY !items.
ALTER TYPE !items(A20).
MATCH FILES file = * / table = !outdata
/BY !items.
EXECUTE.

OUTPUT MODIFY
/RREPORT PRINTREPORT=NO
/SELECT HEADINGS
/IF COMMANDS="["Alter Type(LAST)""]
LABELS=[EXACT("Title")]
INSTANCES=[1]
/DELETEOBJECT DELETE=YES.
OMSEND.

DATASET CLOSE !outdata.
DATASET ACTIVATE !dataset.

!IF (!link = logit) !THEN
COMPUTE Theta = ln(Pred_Prob_01/(1 - Pred_Prob_01)) - Beta.
!IFEND

!IF (!link = probit) !THEN
COMPUTE Theta = Idf.Normal(Pred_Prob_01,0,1) - Beta.
!IFEND

COMPUTE LogLike = (!response*ln(Pred_Prob_01))
+ (1-!response)*(ln(1-Pred_Prob_01)).
COMPUTE Expected = (Pred_Prob_01*ln(Pred_Prob_01))
+ (1-Pred_Prob_01)*ln(1-Pred_Prob_01).
COMPUTE var = Pred_Prob_01*(1-Pred_Prob_01)*((ln(Pred_Prob_01)
- ln(1-Pred_Prob_01))**2).
COMPUTE resid =
(((response - Pred_prob_01)/sqrt(Pred_Prob_01*(1-Pred_Prob_01)))**2.
COMPUTE sq = (response - Pred_Prob_01)**2.
COMPUTE w = (Pred_Prob_01)*(1 - Pred_Prob_01).
COMPUTE w2 = w**2.
COMPUTE winv = 1/w.

AGGREGATE
/BREAK=!id
/perslike = SUM(LogLike)
/persexp = SUM(Expected)
/persvar = SUM(var)
/Sum_Score = SUM(!response)
/information = SUM(w).
COMPUTE persstd = sqrt(persvar).
COMPUTE Z3_person = (perslike - persexp)/persstd.

!IF (!dim = XXXXXXXX) !THEN
COMPUTE Theta_SE = 1/sqrt(information).
EXECUTE.
VARIABLE LABELS Theta_SE 'Person Standard Error'.
!IFEND

VARIABLE LABELS z3_person 'Z3 Person Fit'.
VARIABLE LABELS Sum_Score 'Raw Scale Score'.

AGGREGATE
/BREAK=!items
/cnt = N
/itlike = SUM(LogLike)
/itexp = SUM(Expected)
/itvar = SUM(var)
/sumresid = sum(resid)
/sumsq = sum(sq)
/sumw = sum(w)
/sumw2 = sum(w2)
/sumwinv = sum(winv).
COMPUTE itstd = sqrt(itvar).
COMPUTE Z3_item = (itlike - itexp)/itstd.
COMPUTE outfitchi = sumresid/cnt.
COMPUTE soutfit = sqrt(sumwinv - 4*cnt)/cnt.
COMPUTE out_t = (((outfitchi**(1/3)) - 1)*(3/soutfit)) + (soutfit/3).
COMPUTE infitchi = sumsq/sumw.
COMPUTE infitchi = sqrt(sumw - 4*sumw2)/sumw.
COMPUTE in_t = (((infitchi**(1/3)) - 1)*(3/sinfit)) + (sinfit/3).
EXECUTE.

VARIABLE LABELS z3_item 'Z3 Item Fit'.
VARIABLE LABELS infitchi 'Infit (Chi)'.
VARIABLE LABELS outfitchi 'Outfit (Chi)'.
VARIABLE LABELS in_t 'Infit (t)'.
VARIABLE LABELS out_t 'Outfit (t)'.
EXECUTE.

DELETE VARIABLES LogLike Expected var perslike persexp persvar persstd itlike itexp itvar itstd resid sq w w2 cnt sumresid sumsq sumw sumw2 sumwinv sinfit soutfit information.
EXECUTE.

CTABLES
   /MRSETS COUNTDUPLECTURES=YES
   /SMISSING VARIABLE
   /VLABELS VARIABLES=!items Beta Z3_item infitchi in_t outfitchi out_t DISPLAY=LABEL
   /TABLE !items [C] BY Beta [S][MEDIAN ''] + Z3_item [MEDIAN ''] + infitchi [MEDIAN ''] + in_t [MEDIAN ''] + outfitchi [MEDIAN ''] + out_t [MEDIAN '']
   /SLABELS VISIBLE=NO
   /CATEGORIES VARIABLES=!items ORDER=A KEY=VALUE EMPTY=EXCLUDE
   /TITLES
       TITLE='Item Fit Statistics'.

OUTPUT MODIFY
   /REPORT PRINTREPORT=NO
   /SELECT HEADINGS
   /IF COMMANDS=['CTables(LAST)'] LABELS=[EXACT("Title")]
   /DELETEOBJECT DELETE=YES.

!IF (!theta = YES) !THEN
!IF (!dim = XXXXXXXX) !THEN
CTABLES
   /MRSETS COUNTDUPLECTURES=YES
   /SMISSING VARIABLE
   /VLABELS VARIABLES=!id DISPLAY=NAME
   /VLABELS VARIABLES=Theta Theta_SE Z3_person DISPLAY=LABEL
   /TABLE !id [C] BY Theta [S][MEDIAN ''] + Theta_SE [S][MEDIAN ''] + Z3_person [S][MEDIAN '']
   /SLABELS POSITION=ROW VISIBLE=NO
   /CATEGORIES VARIABLES=!id ORDER=A KEY=VALUE EMPTY=EXCLUDE
   /TITLES
       TITLE='Person Statistics'.

OUTPUT MODIFY
   /REPORT PRINTREPORT=NO

73
/SELECT HEADINGS
/IF COMMANDS="CTables(LAST)"
   LABELS=[EXACT("Title")]
   INSTANCES=1
/DELETEOBJECT DELETE=YES.
!IFEND

!IF (!dim ~ = XXXXXXXX) !THEN
CTABLES
   /FORMAT EMPTY=BLANK MISSING=’.’
   /MISSING VARIABLE
   /VARIABLES=!id Theta !dim Z3_person DISPLAY=LABEL
   /TABLE id [C] BY Theta [S][MEDIAN ’’] > !dim [C]
   + Z3_person [S][MEDIAN ’’]
   /LABELS VISIBLE=NO
   /CATEGORIES VARIABLES=!id !dim ORDER=A KEY=VALUE EMPTY=EXCLUDE
   /TITLES
      TITLE=’Person Statistics’.
OUTPUT MODIFY
   /REPORT PRINTREPORT=NO
   /SELECT HEADINGS
   /IF COMMANDS="CTables(LAST)"
      LABELS=[EXACT("Title")]
      INSTANCES=1
   /DELETEOBJECT DELETE=YES.
!IFEND
!IFEND

VARIABLE LABELS Theta ’Person Ability’.

!IF (!dim = XXXXXXXX) !THEN
!IF (!plot = YES) !THEN
GRAPH
   /SCATTERPLOT(BIVAR)=Sum_Score WITH Theta
   /MISSING=LISTWISE.
OUTPUT MODIFY
   /REPORT PRINTREPORT=NO
   /SELECT HEADINGS
   /IF COMMANDS="Graph(LAST)"
      LABELS=[EXACT("Title")]
      INSTANCES=1
   /DELETEOBJECT DELETE=YES.
DATASET ACTIVATE !dataset.
DATASET DECLARE tempdata.
AGGREGATE
   /OUTFILE='tempdata'
74
/BREAK=!items
/Beta=FIRST(Beta).
DATASET ACTIVATE tempdata.
compute x004=-4.5.
compute x003=-4.25.
compute x002=-4.
compute x001=-3.75.
compute x01=-3.5.
compute x02=-3.25.
compute x03=-3.
compute x04=-2.75.
compute x05=-2.5.
compute x06=-2.25.
compute x07=-2.
compute x08=-1.75.
compute x09=-1.5.
compute x10=-1.25.
compute x11=-1.
compute x12=-.75.
compute x13=-.5.
compute x14=-.25.
compute x15=0.
compute x16=4.5.
compute x16=4.25.
compute x17=4.
compute x18=3.75.
compute x01=3.5.
compute x02=3.25.
compute x03=3.
compute x04=2.75.
compute x05=2.5.
compute x06=2.25.
compute x07=2.
compute x08=1.75.
compute x09=1.5.
compute x10=1.25.
compute x11=1.
compute x12=.75.
compute x13=.5.
compute x14=.25.
EXECUTE.
VARSTOCASES
/make Theta from x004 TO x014.
COMPUTE Probability = exp(Theta + Beta)/(1 + exp(Theta + Beta)).
compute Information = (Probability*(1-Probability)).
EXECUTE.
OUTPUT MODIFY
/REPORT PRINTREPORT=NO
/SELECT TABLES
/IF COMMANDS=\"Variables to Cases(LAST)\"
  LABELS=[EXACT("Processing Statistics")]
  INSTANCES=[1]
/DELETEOBJECT DELETE=YES
/SELECT TABLES
/IF COMMANDS=\"Variables to Cases(LAST)\"
  LABELS=[EXACT("Generated Variables")]
  INSTANCES=[1]
/DELETEOBJECT DELETE=YES
/SELECT TEXTS
/IF COMMANDS=\"Variables to Cases(LAST)\"
  LABELS=[EXACT("Active Dataset")]
  INSTANCES=[1]
/DELETEOBJECT DELETE=YES.
DATASET ACTIVATE tempdata.
GRAPH
/LINE(MULTIPLE)=MEAN(Probability) BY Theta BY !items
/TITLE='Item Characteristic Curve'.
OUTPUT MODIFY
/REPORT PRINTREPORT=NO
/SELECT HEADINGS
/IF COMMANDS=\"Graph(LAST)\"
  LABELS=[EXACT("Title")]
  INSTANCES=[1]
/DELETEOBJECT DELETE=YES.
GRAPH
/LINE(MULTIPLE)=MEAN(Information) BY Theta BY !items
/TITLE='Item Information Curve'.
OUTPUT MODIFY
/REPORT PRINTREPORT=NO
/SELECT HEADINGS
/IF COMMANDS=\"Graph(LAST)\"
  LABELS=[EXACT("Title")]
  INSTANCES=[1]
/DELETEOBJECT DELETE=YES.
DATASET DECLARE tempdata2.
AGGREGATE
OUTFILE='tempdata2'
/BREAK=Theta
/test_info=SUM(Information).
DATASET CLOSE tempdata.
DATASET ACTIVATE tempdata2.
COMPUTE SEofM=1/sqrt(test_info).
EXECUTE.
VARIABLE LABELS SEofM 'Standard Error of Measurement'.
RENAME VARIABLES (test_info=Information).
VARSTOCASES
   /MAKE Value FROM Information SEofM
   /INDEX=Index(Value)
   /KEEP=Theta
   /NULL=KEEP.
OUTPUT MODIFY
   /REPORT PRINTREPORT=NO
   /SELECT TABLES
      /IF COMMANDS="Variables to Cases(LAST)"
         LABELS=[EXACT("Processing Statistics")]
         INSTANCES=[1]
         /DELETEOBJECT DELETE=YES
   /SELECT TABLES
      /IF COMMANDS="Variables to Cases(LAST)"
         LABELS=[EXACT("Generated Variables")]
         INSTANCES=[1]
         /DELETEOBJECT DELETE=YES
   /SELECT TEXTS
      /IF COMMANDS="Variables to Cases(LAST)"
         LABELS=[EXACT("Active Dataset")]
         INSTANCES=[1]
         /DELETEOBJECT DELETE=YES
   /SELECT HEADINGS
      /IF COMMANDS="Variables to Cases(LAST)"
         LABELS=[EXACT("Title")]
         INSTANCES=[1]
         /DELETEOBJECT DELETE=YES.
GRAPH
   /LINE(MULTIPLE)=MEAN(Value) BY Theta BY Index
   /TITLE='Grouped Item Information Curve'.
OUTPUT MODIFY
   /REPORT PRINTREPORT=NO
   /SELECT HEADINGS
      /IF COMMANDS="Graph(LAST)"
         LABELS=[EXACT("Title")]
         INSTANCES=[1]
         /DELETEOBJECT DELETE=YES.
DATASET CLOSE tempdata2.
DATASET ACTIVATE !dataset.
DELETE VARIABLES Sum_Score.

!IF (!save = 0) !THEN
DELETE VARIABLES Pred_Prob_01 Beta Std.Error Theta Z3_person Z3_item outfitchi out_t infitchi in_t.
!IF (!center =~ item) !THEN
DELETE VARIABLES Upper Lower.
!IFEND
!IF (!dim = XXXXXXXX) !THEN
DELETE VARIABLES Theta_SE.
!IFEND
!IFEND

!IF (!save = 2) !THEN
DELETE VARIABLES Pred_Prob_01 Std.Error Theta Z3_person Z3_item outfitchi out_t infitchi in_t.
!IF (!center =~ item) !THEN
DELETE VARIABLES Upper Lower.
!IFEND
!IF (!dim = XXXXXXXX) !THEN
DELETE VARIABLES Theta_SE.
!IFEND
!IFEND

!IF (!save = 3) !THEN
DELETE VARIABLES Pred_Prob_01 Std.Error Z3_person Z3_item outfitchi out_t infitchi in_t.
!IF (!center =~ item) !THEN
DELETE VARIABLES Upper Lower.
!IFEND
!IF (!dim = XXXXXXXX) !THEN
DELETE VARIABLES Theta_SE.
!IFEND
!IFEND

!IF (!difitem =~ YYYYYYY) !THEN
!IF (!difvar =~ YYYYYYY) !THEN
DELETE VARIABLES item_ind DIF.
!IFEND
!IFEND

!ENDDEFINE.
Appendix B: SPSSIRT Output Images
Figure B.1: Image taken from SPSSIRT output. This is the “Fixed Effects” table, which gives item parameter ($\beta_i$) estimates and covariate effect ($\gamma$) estimates. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.

<table>
<thead>
<tr>
<th>Item</th>
<th>Data</th>
<th>Std. Error</th>
<th>t</th>
<th>Sig.</th>
<th>95% Confidence Interval Lower</th>
<th>95% Confidence Interval Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1DoCurse</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S1DoScold</td>
<td>-3.73</td>
<td>.345</td>
<td>2.512</td>
<td>.012</td>
<td>-0.092</td>
<td>1.164</td>
</tr>
<tr>
<td>S1DoShout</td>
<td>-7.328</td>
<td>.633</td>
<td>3.131</td>
<td>.000</td>
<td>-1.433</td>
<td>-5.332</td>
</tr>
<tr>
<td>S1WantCurs</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S1WantBcol</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S1WantBshout</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S2DoCurse</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S2DoScold</td>
<td>-3.73</td>
<td>.345</td>
<td>2.512</td>
<td>.012</td>
<td>-0.092</td>
<td>1.164</td>
</tr>
<tr>
<td>S2DoShout</td>
<td>-7.328</td>
<td>.633</td>
<td>3.131</td>
<td>.000</td>
<td>-1.433</td>
<td>-5.332</td>
</tr>
<tr>
<td>S2WantCurs</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S2WantBcol</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
<tr>
<td>S2WantBshout</td>
<td>1.157</td>
<td>.1675</td>
<td>7.494</td>
<td>.000</td>
<td>-1.675</td>
<td>2.275</td>
</tr>
</tbody>
</table>

81
Figure B.2: Image taken from SPSSIRT output. This is the “Person Statistics” table, which gives $\theta_p$ values for every respondent, as well as person fit statistics. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.

<table>
<thead>
<tr>
<th>Id</th>
<th>Theta</th>
<th>Person Standard Error</th>
<th>Z3 Person Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.46</td>
<td>0.46</td>
<td>-2.39</td>
</tr>
<tr>
<td>2</td>
<td>-2.42</td>
<td>0.70</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>-0.27</td>
<td>0.46</td>
<td>-0.57</td>
</tr>
<tr>
<td>4</td>
<td>0.48</td>
<td>0.46</td>
<td>0.44</td>
</tr>
<tr>
<td>5</td>
<td>-0.27</td>
<td>0.46</td>
<td>1.73</td>
</tr>
<tr>
<td>6</td>
<td>-0.08</td>
<td>0.46</td>
<td>-0.05</td>
</tr>
<tr>
<td>7</td>
<td>0.48</td>
<td>0.46</td>
<td>-1.16</td>
</tr>
<tr>
<td>8</td>
<td>-1.52</td>
<td>0.54</td>
<td>0.45</td>
</tr>
<tr>
<td>9</td>
<td>-0.27</td>
<td>0.46</td>
<td>-1.46</td>
</tr>
<tr>
<td>10</td>
<td>1.78</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>11</td>
<td>0.67</td>
<td>0.47</td>
<td>-0.30</td>
</tr>
<tr>
<td>12</td>
<td>1.08</td>
<td>0.49</td>
<td>0.34</td>
</tr>
<tr>
<td>13</td>
<td>-0.27</td>
<td>0.46</td>
<td>1.81</td>
</tr>
<tr>
<td>14</td>
<td>1.53</td>
<td>0.53</td>
<td>0.37</td>
</tr>
<tr>
<td>15</td>
<td>0.29</td>
<td>0.46</td>
<td>0.19</td>
</tr>
<tr>
<td>16</td>
<td>0.10</td>
<td>0.46</td>
<td>0.67</td>
</tr>
<tr>
<td>17</td>
<td>0.67</td>
<td>0.47</td>
<td>-1.81</td>
</tr>
<tr>
<td>18</td>
<td>-0.46</td>
<td>0.46</td>
<td>-0.29</td>
</tr>
<tr>
<td>19</td>
<td>-2.85</td>
<td>0.82</td>
<td>1.28</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>0.46</td>
<td>1.44</td>
</tr>
<tr>
<td>21</td>
<td>-0.65</td>
<td>0.47</td>
<td>1.79</td>
</tr>
<tr>
<td>22</td>
<td>2.05</td>
<td>0.59</td>
<td>1.44</td>
</tr>
<tr>
<td>23</td>
<td>2.36</td>
<td>0.64</td>
<td>1.44</td>
</tr>
</tbody>
</table>
Figure B.3: Image taken from SPSSIRT output. This is the “Item Fit Statistics” table, which gives item fit statistics. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.

<table>
<thead>
<tr>
<th>Item</th>
<th>Beta</th>
<th>23 Item Fit</th>
<th>Infit (Ch)</th>
<th>Outfit (Ch)</th>
<th>Outfit (I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1DoCurse</td>
<td>.517</td>
<td>1.63</td>
<td>.60</td>
<td>-1.27</td>
<td>.06</td>
</tr>
<tr>
<td>S1DoShout</td>
<td>.540</td>
<td>1.63</td>
<td>.60</td>
<td>-1.27</td>
<td>.06</td>
</tr>
<tr>
<td>S2DoCurse</td>
<td>.591</td>
<td>1.63</td>
<td>.60</td>
<td>-1.27</td>
<td>.06</td>
</tr>
<tr>
<td>S2DoShout</td>
<td>.591</td>
<td>1.63</td>
<td>.60</td>
<td>-1.27</td>
<td>.06</td>
</tr>
</tbody>
</table>

Figure B.4: Image taken from SPSSIRT output. This is the “Random Effect” table, which gives the random effect variance estimate(s) ($\sigma^2$) for the model. In a multidimensional model, the covariance(s) between dimensions will also be given here. This specific table comes from the verbal aggression dataset, when a basic 1PL IRT model is conducted.

<table>
<thead>
<tr>
<th>Random Effect Covariance</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Z</th>
<th>Sig</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.056</td>
<td>.170</td>
<td>10.001</td>
<td>.000</td>
<td>1.394 - 2.983</td>
</tr>
</tbody>
</table>
Figure B.5: Image taken from SPSSIRT output. This depicts the item characteristic curves for all 24 items in the verbal aggression dataset, created using the ICC function.
Figure B.6: Image taken from SPSSIRT output. This depicts the item information as a function of theta. This image was created using the ICC function.