



Explanatory IRT Analysis Using the SPIRIT Macro in SPSS

Jack DiTrapani ^{a,✉}, Nicholas Rockwood ^a & Minjeong Jeon ^b

^aDepartment of Psychology, The Ohio State University

^bUniversity of California, Los Angeles


Abstract ■ Item Response Theory (IRT) is a modeling framework that can be applied to a large variety of research questions spanning several disciplines. To make IRT models more accessible for the general researcher, a free tool has been created that can easily conduct one-parameter logistic IRT (1PL) analyses using the convenient point-and-click interface in SPSS without any required downloads or add-ons. This tool, the SPIRIT macro, can fit 1PL models with person and item covariates, DIF analyses, multidimensional models, multigroup models, rating scale models, and several other variations. Example explanatory models are presented with an applied dataset containing responses to an ADHD rating scale. Illustrations of how to fit basic 1PL models as well as two more complicated analyses using SPIRIT are given.

Keywords ■ Item response theory, 1PL model, generalized linear mixed models, explanatory IRT models, IRTrees. **Tools** ■ SPSS.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer.

✉ ditrapani.4@osu.edu

 *JDT*: 0000-0002-6044-7272; *NR*: 0000-0001-5931-183X; *MJ*: 0000-0002-5880-4146

 [10.20982/tqmp.14.2.p081](https://doi.org/10.20982/tqmp.14.2.p081)

Introduction

Item response theory (IRT) is a modeling framework used to model responses to scale or test items that are categorical. Since one of the goals of item response modeling is often to acquire measures of person ability, it has historically been used extensively for education research. However, IRT methods 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. IRT analyses have 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). Public health and clinical psychology researchers also commonly utilize item response models. Aggen, Neale, and Kendler (2005) provide an example of how IRT models can be used to analyze Diagnostic and

Statistical Manual of Mental Disorders (DSM) criteria for depression in more beneficial ways, while Hagquist, Bruce, and Gustavsson (2009) present an introduction on scale development using IRT in the field of nursing. Small et al. (2008) utilized IRT models to examine group differences in symptomologies for depressed adolescents. In the domain of dermatology, Nijsten, Unaeze, and Stern (2006) used IRT methods to enhance the effectiveness of a patient questionnaire measuring the impact of Psoriasis. These examples are just a few of the many diverse applications of item response models across various disciplines.

Item response models are often labeled by the number of parameters that characterize each item. This paper focuses on one-parameter logistic (1PL) item response models¹ which, as the name suggests, include only one parameter (the easiness) for each item. Although two- and three-parameter logistic (2PL and 3PL) item response models can be considered as more general versions of the 1PL model, the 1PL model has some practical and theoretical advantages. Because 1PL models contain fewer parame-

¹All of the research using IRT described earlier utilized variations of the 1PL model.



ters, less data are needed to obtain accurate parameter estimates. For example, 2PL or 3PL models need hundreds, if not thousands, of respondents before the parameter estimates can be considered trustworthy (Hulin, Lissak, & Drasgow, 1982), while the 1PL estimates can be more reliable with potentially as few as around 100 respondents (Linacre, 1994; Edelen & Reeve, 2007). This added parsimony comes with obvious computation benefits as well; 1PL models are typically quicker to estimate than 2PL (and certainly 3PL) equivalents.

The 1PL model is often referred to as the Rasch model (Rasch, 1960), and we will use these terms interchangeably in this paper. However, some strong proponents of the Rasch model view it as distinct from the general IRT framework due to some fundamental measurement properties that are not shared by other more complex item response models. For example, the Rasch model ensures that all respondents who answer an equal amount of items correctly will finish with identical “scores.” Within the 2PL framework, counterintuitive phenomena can sometimes occur. One such possibility is that a given respondent can have a higher modeled probability of success relative to another respondent on an “easy” item but a lower modeled probability of success relative to the other respondent for a “difficult” item (Wright, 1977). When using the Rasch model, the relative ordering of respondents is modeled to be equivalent regardless of the item being answered.

Motivation

Despite the attractive benefits and flexibility of the 1PL IRT model, it is still not as commonly used in psychological research. Much of the current psychological literature still bases measurement on classical test theory principles, which can be overly simplistic relative to IRT approaches (Embretson & Reise, 2013). Part of this disconnect could be due to researcher apprehension of trying novel analyses within foreign software environments.

Foster, Min, and Zickar (2017) surveyed 343 practitioners in organizational psychology and inquired about IRT usage. Of the 343 respondents, only 30.90% responded saying they use IRT. Of those who did not use IRT, 35.21% admitted to either never having learned the method or finding it too complex. An additional 15.65% mentioned that software being too expensive was the main reason why they did not attempt to utilize item response models.

Currently, there are dozens of software packages that can estimate 1PL IRT models. Many options are packages within R, such as ltm (Rizopoulos, 2006), mlirt (Fox, 2007), eRm (Mair & Hatzinger, 2007), mirt (Chalmers, 2012), FLIRT (Jeon, Rijmen, & Rabe-Hasketh, 2014), and TAM (Kiefer, Robitzsch, & Wu, 2016). Other options include software specifically programmed for IRT, such as BILOG

(Zimowski, Muraki, Mislevy, & Bock, 1996), IRTPRO (Cai, Thissen, & du Toit, 2011), flexMIRT (Houts & Cai, 2013), ConQuest (Adams, Wu, & Wilson, 2015), and Winsteps (Linacre, 2016). These options are excellent for IRT, but they can often be expensive or intimidating for those using IRT for the first time.

Other options include more general statistical software that can be adapted to run item response models, such as SAS PROC NL MIXED (SAS Institute Inc., 2015), the gllamm package within Stata (Rabe-Hasketh, Skrondal, & Pickles, 2004), the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015; De Boeck et al., 2011), and Mplus (Muthen & Muthen, 1998-2011). These options are all powerful and flexible in the models they can fit. However, they were not made specifically with IRT in mind, meaning they provide output that may be difficult for the common practitioner to interpret in the context of IRT. The packages are very broad in scope, resulting in output that may contain vocabulary inconsistent with the common IRT literature.

This paper presents a new, more accessible option to fit IRT models for researchers familiar with the popular SPSS software (IBM Corp., Released 2013). The newly-developed SPIRIT macro allows practitioners to use the convenient point-and-click interface available in SPSS to quickly and easily fit a wide variety of one-parameter item response models. The hope is that this macro will allow these valuable models to be more accessible to the general research and teaching community by putting them in the SPSS environment that is comfortable and familiar to many researchers.

We begin by detailing 1PL item response models before providing an overview of the SPIRIT macro. Following, we present three example analyses that demonstrate how to formulate item response models to address substantive research questions, as well as how to fit these models using the point-and-click SPSS interface for SPIRIT. The corresponding syntax is provided in the appendix.

1PL Item Response Models

IRT models are typically fit to *dichotomous* or *polytomous* item responses. Items in which the response options consist of only two categories are referred to as *dichotomous* items, which are commonly used in educational and cognitive testing (e.g., correct vs. incorrect) and behavioral checklists (e.g., displays symptom vs. does not display symptom). Items in which the response options consist of more than two categories are referred to as *polytomous* items. Within social and behavioral science research, a common type of polytomous item is a Likert-scale item, such as when respondents express their agreement with a given statement on a 5-point scale. Here we describe the 1PL IRT model for dichotomous items. Later, when



we present some example analyses, we will demonstrate a method for extending the model to include polytomous items.

If Y_{ip} is defined as the (dichotomous) response by the p -th ($p = 1, \dots, P$) respondent on the i -th ($i = 1, \dots, I$) item, we can model Y_{ip} with a 1PL model as

$$Y_{ip} \sim \text{Bernoulli}(\pi_{ip}), \tag{1}$$

$$\log\left(\frac{\pi_{ip}}{1 - \pi_{ip}}\right) = \eta_{ip},$$

where “ $\sim \text{Bernoulli}(\pi_{ip})$ ” is read “is distributed as a Bernoulli random variable with probability π_{ip} .” A Bernoulli random variable can only take the values of 1 or 0, and the probability of the variable taking the value of 1 is equal to its single parameter, which is π_{ip} here. Following, π_{ip} is the probability of respondent p providing a response of “1” (a “correct” response) to item i . Symbolically, we can write this as $Pr(Y_{ip} = 1) = \pi_{ip}$. The ip subscripts suggest that the probabilities may change depending on both the respondent and the item, and it is this heterogeneity in item response probabilities that we wish to model.

Rather than modeling the probabilities directly, we model a transformation of the probabilities. The term $\pi_{ip}/(1 - \pi_{ip})$ corresponds to the odds that person p responds correctly to item i and so taking the (natural) log of this term results in the log-odds, or *logit*. For 1PL models, and consequently all models discussed here, the logit of responding with a “1” is modeled by some linear function, η_{ip} , of person and item characteristics. Thus, η_{ip} is on the log-odds scale. We can convert η_{ip} to the odds scale by exponentiating (i.e., $e^{\eta_{ip}}$), or to the probability scale using the formula

$$\pi_{ip} = \frac{e^{\eta_{ip}}}{1 + e^{\eta_{ip}}},$$

which is known as the logistic function. Each of these transformations is monotonic, meaning that higher values of η_{ip} correspond to both higher odds and higher probability of a “1” response. When we fit our example models in a later section, we will return to these transformations to obtain a better understanding of the substantive meaning of our estimated parameters that form η_{ip} .

Although all models discussed here will take the form of Equation 1, there is flexibility in the specification of the linear function η_{ip} . To fit the *basic* 1PL, or Rasch model, η_{ip} is defined as follows:

$$\eta_{ip} = \beta_i + \theta_p. \tag{2}$$

Here, β_i corresponds to the fixed item intercept for the i -th item. Since higher values of β_i correspond with a higher probability of a “successful” response, the intercept parameter can be labeled as a measure of the “easiness” of that

specific item. In the IRT literature, the item easiness parameter is often replaced by an item difficulty parameter, which is simply the negative of the easiness parameter, $-\beta_i$.

The term θ_p refers to respondent p 's latent, or unobserved, trait. Because early developments in IRT stem from educational testing, θ_p is often referred to as respondent p 's “ability”, as a higher value of θ_p corresponds to a higher probability of a “successful” response. We use the term “trait” instead, since the model may be applied to other types of item responses beyond testing. The latent θ_p values are assumed to be sampled from a normal distribution with a mean of zero and variance σ^2 . Because θ_p is modeled as a random variable, it is referred to as a random effect. From a modeling standpoint, random effects are synonymous with latent variables and we will use these terms interchangeably. The individual θ_p values are not actually model parameters, but we can obtain predicted θ_p values using a scoring technique, such as EAP or MAP, after estimating the item intercepts and σ^2 .

Person and/or item covariates, as well as multiple latent variables, can be included as additional variables in the η_{ip} term, allowing for a highly flexible model that can be used to address interesting substantive research questions. We provide examples of these extensions in a later section. First, we describe the SPIRIT macro and its capabilities.

The SPIRIT Macro

The SPIRIT macro allows researchers to conduct 1PL analyses through the typical SPSS point-and-click and/or syntax interfaces without any external software add-ons. It is freely available and easily implemented in SPSS versions 21-24. The macro takes advantage of the fact that 1PL item response models can be formulated within the generalized linear mixed model (GLMM) family (e.g., Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; De Boeck et al., 2011). As De Boeck et al. (2011) describe, these models fall within the GLMM framework because they contain a link function (e.g., the logistic link) and a linear component (η_{ip}) that is a function of both fixed effects (e.g., β_i) and random effects (e.g., θ_p). Therefore, the macro utilizes the GENLINMIXED function of SPSS.

One of the main advantages of SPIRIT is that the model specification and outputted results are specifically designed in the context of IRT, as even researchers who are moderately familiar with the connections between GLMMs and 1PL models may have trouble interpreting the typical GENLINMIXED output. In addition, the individual person random effect values are not given for each individual (only the variance of this random effect is given in the typical output). However, since SPIRIT was designed specifi-



cally for IRT analyses, the output gives the item parameters and predicted person latent variable estimates.

SPIRIT can be used to fit a large variety of item response models, including the basic 1PL (or Rasch) model, latent regression (Adams, Wilson, & Wang, 1997), linear logistic test model (LLTM; Fischer, 1973), multidimensional models, multigroup models (where the distribution mean and variance of the random effect differs by group membership), and models exploring differential item functioning (DIF) for dichotomous responses. It can also execute rating scale models for polytomous response data (Andrich, 1978). SPIRIT is flexible in that the user can mix and match from the above model options, leading to potentially complex models that can address a variety of substantively interesting research questions.

Like most item response software, SPIRIT is flexible in its handling of missing data. Missing item responses, which may result from the use of computerized adaptive testing or other scenarios where respondents are only exposed to a subset of items, pose no serious challenges, as long as the missingness is assumed to occur at random or completely at random. More sophisticated approaches would need to be applied if the missingness was not due to randomness, such as a respondent purposefully skipping questions (De Boeck & Partchev, 2012). For models that include person covariates, respondents that have missing values on the covariates are excluded from the analysis, as the response probabilities are modeled as conditional on these unknown values.

An optional feature of the macro is the calculation of item (infit, outfit, lz; Wright & Masters, 1982; Drasgow, Levine, & Williams, 1985), and person (lz) fit statistics.² SPIRIT also has the ability to plot the item characteristic curves and information functions for all items when a basic 1PL model is specified. The SPIRIT User Guide, which comes with the macro, provides information on data requirements, installation instructions, and detailed descriptions of the specific options of the macro. An overview of several pertinent requirements and properties of SPIRIT is provided in Appendix A.

Illustration

This section contains several examples of item response analyses conducted using SPIRIT. The models of interest include the basic 1PL model, an explanatory model with person and item covariates, and a model with a polytomous item response variable. For each, the model specification within SPIRIT and the interpretation of the results is provided.

Data

Throughout this section, models will be fit to a dataset of simulated responses to a scale that attempts to measure a child's level of attention-deficit/hyperactivity disorder (ADHD). This simulated dataset is based on a real data that was accessed through the National Database for Autism Research (NDAR). The size and format of the generated dataset is identical to that of the actual dataset, as are the basic conclusions from data analyses. We decided to use a simulated dataset to be able to publicly share the data, allowing readers to directly follow along with the following illustrations on their own computers.

The simulated dataset contains responses from 234 respondents to 18 different items. Each of the items come from the ADHD Rating Scale-IV: Home Version (DuPaul et al., 1998), which consists of items that correspond to specific symptoms of ADHD as noted in the DSM-IV. This scale was given to a parent or guardian of each child, and each of these respondents were asked to report on the severity of their child's symptoms. Some examples of items from this scale are to report how often the child "fidgets with hands or feet or squirms in seat" or "does not seem to listen when spoken to directly." The answer to each item ranged from "Never or Rarely" (0) to "Very Often" (3). Each item response was also dichotomized, such that a response of 2 or 3 was coded as a "1," which suggests that the child frequently displays that item's symptom. An original response of 0 or 1 was coded as a "0," suggesting that the specific symptom is not prevalent for that child.

Information on the child and respondent is also present in the dataset. This includes the age (in months) and gender of the child as well as the gender of the respondent. Age is mean-centered for interpretability purposes. Dimensionality is also expected to be present in the data (DuPaul et al., 1998). Specifically, there are nine items that intend to measure the level of "Inattention" symptoms present in a child (such as the second item presented earlier), and nine items attempting to measure the "Hyperactivity-Impulsivity" level of a child (such as the first item given earlier).

An example of how this dataset looks is shown in Table 1. The data must be presented in long form in order for the SPIRIT macro to be used. Therefore, there are multiple rows for each respondent. In Table 1, the three rows being presented correspond to the responses on items 1, 2, and 3 by respondent 1.

In the dataset, the "SubGen" variable denotes the gender of the child ("M" = Male, "F" = Female) and the "RespGen" variable corresponds to the gender of the respondent. The "Item" variable is an indicator variable denot-

²Fit statistics are only available for the basic 1PL model with no covariates.

**Table 1** ■ First three rows for the generated ADHD dataset in long form.

ID	Age	SubGen	RespGen	Item	Dim	Resp	RespDi
1	16.54	F	F	p_c_adhdrs_1	I	1	0
1	16.54	F	F	p_c_adhdrs_2	H	0	0
1	16.54	F	F	p_c_adhdrs_3	I	1	0

ing which item is being answered, and “Dim” is the dimension corresponding to that specific item (“I” for Inattention and “H” for Hyperactivity). Finally, “Resp” is the actual response given (ranging from 0 to 3), and “RespDi” is the dichotomized response (0 or 1).

Example Models

Several types of IRT models can be fit using SPIRIT, many of which are applied to dichotomous outcome variables (such as the “RespDi” variable described in the ADHD dataset). To fit the basic 1PL model described earlier using SPIRIT, the user must specify the response variable in the “Response Variable” box (“RespDi” in this example), the item variable in the “Items” box (“Item” in this example), and the ID variable in the “ID” box (“ID” in this example). See Figure 1 for a screenshot of how this model is specified. It is also possible to obtain the individual predicted θ_p values in the output by clicking the “Theta Values” option. The “Fit Statistics” and “Plots” options are available for this basic 1PL model only. All of the output is supplied in the typical SPSS output window. For this example, each individual θ_p score can be interpreted as the predicted severity of ADHD behaviors for that particular reporter/child pair relative to the other reporters/children in the data.

Relative to other possible models that can be estimated using SPIRIT, this basic 1PL model is quite simple. Two more complicated models will now be explored to illustrate SPIRIT’s flexibility. Descriptions of other possible models are illustrated in DiTrapani (2016) and the SPIRIT User Guide, which can be found by clicking the “Help” button in the SPIRIT interface.

Explanatory Model Example

In addition to models designed specifically for measurement, the SPIRIT macro can be used to estimate item response models designed to answer substantive research questions. One approach is to use SPIRIT’s ability to include person and item covariates to investigate associations between the covariates and the scale’s item responses. This type of item response modeling is often termed *explanatory IRT* (De Boeck & Wilson, 2004). The explanatory model is identical in structure to the basic 1PL described in Equation 2. Now, however, the η_{ip} linear term can be extended to include coefficients corresponding to particular item or person covariates (as well as more com-

plicated terms such as interactions between covariates).

As an illustration using the simulated ADHD dataset, perhaps we are interested in whether or not Hyperactivity, or “H” behaviors, and Inattention, or “I” behaviors, are reported at different rates. This question could be of interest to psychologists who are curious about the respondents being sampled or to psychometricians who want to learn more about how this particular scale tends to be answered. To address this question, we can include the item type (“H” or “I”) as an item covariate in the model.

Suppose we are also interested in whether male or female reporters (fathers and mothers of the child, in most cases) tend to be more or less likely to report ADHD-related behaviors. Previous research has suggested that there may be discrepancies between how parents rate their children. Seiffge-Krenke and Kollmar (1998) shows that although mother and father ratings are correlated, fathers tended to be less likely to rate their children as having problem behaviors. A meta-analysis of 60 studies assessing behavioral problems in children found this same effect (although it was relatively small here); mothers were more likely to report behavioral problems than fathers (Duhig, Renk, Epstein, & Phares, 2000). In the context of ADHD rating scales, Sollie, Larsson, and Mørch (2012) also concluded that fathers were less likely to report ADHD behaviors in their children.

Knowledge of differences in the way mothers and fathers report symptoms could impact how treatment is developed for a given child, or how ADHD measurements are interpreted. The sex of the reporter can therefore be included as a person covariate into the model to investigate this research question. The effect of this covariate quantifies the overall average difference between the tendency for mothers to report ADHD symptoms relative to fathers, holding everything else constant.

Interactions between covariates can also be added to answer more specific questions. For example, we can include an interaction between (mean-centered) age of the child (in months) and the type of the item. This interaction investigates whether the effect of child age on a “successful” reported response changes depending on the type of item. In other words, does the age of a child relate to “H” items differently than “I” items? This may be of interest to researchers since this term would examine what explanatory role child age plays for both dimensions sep-

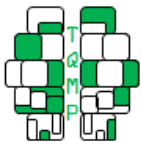
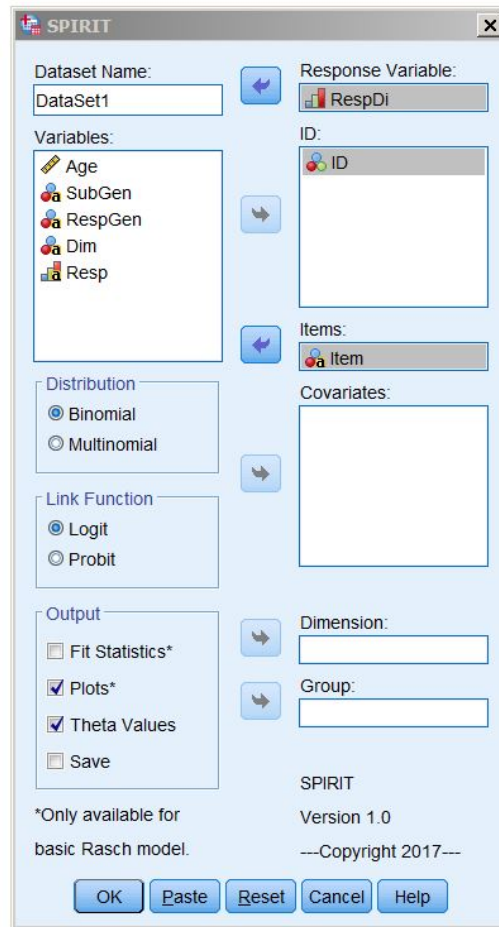


Figure 1 ■ How to specify the basic 1PL model in SPIRIT.



arately. This phenomenon would be similar in nature to DIF, since two subsets of items behave differently depending on a characteristic of the child (age, in this case).

The proposed model is a form of a linear logistic test model (LLTM; Fischer, 1973), which is designed to explain individual item intercepts using item covariates. Therefore, there is not a separate intercept/easiness for every item but instead an equivalent intercept/easiness for entire subsets of items (i.e., for this example all “H” items have the same intercept/easiness). Since the basic version of this model treats all items within a subset as equivalent, which may be an unrealistic assumption, an extension can be made that includes a random error term for each specific item, which acts as an item-level residual.

The only difference between the basic 1PL model and the explanatory model is the structure of the linear term, η_{ip} . In the basic model, η_{ip} is only a function of the person and item effects (θ_p and β_i). The η_{ip} term is more compli-

cated for the explanatory model, as it contains the person and item covariates. Specifically, it is now

$$\eta_{ip} = \beta_0 + \beta_1(H_i) + \gamma_1(Male_p) + \gamma_2(Age_p) + \delta(H_i * Age_p) + \theta_p + \epsilon_i.$$

Here, H_i equals 1 for all “H” items and 0 for all “I” items. As a result of this parameterization, β_0 corresponds to the “easiness” of an “I” item being endorsed/reported (when all other predictors equal zero), and β_1 corresponds to the difference in easiness values for all “H” items relative to “I” items for children of average age, holding all other variables constant. Higher “easiness” estimates here suggest that items are reported relatively more often. γ_1 can be interpreted as the difference between male and female reporters in how likely a response of “1” is reported, and is positive if male reporters tend to respond with a “1” more often than females (this gender difference is constrained



by the model to be equivalent across both “I” and “H” items). γ_2 is the regression coefficient for the effect of the age of the child on the probability of a “1” response to “I” items. Finally, δ is the interaction effect between a child’s age and the type of item being responded to; specifically, this parameter would be positive if the effect of age on reported symptoms is higher for “H” items than for “I” items. With this parameterization, γ_2 can be interpreted as the regression coefficient of age specifically for “I” items, and $\gamma_2 + \delta$ can be interpreted as the coefficient of age for “H” items, since the δ term is only present when $H_i = 1$.

The θ_p and ϵ_i values are the random effect terms for the explanatory IRT model. The person random effect, θ_p , is the residual term for person p after controlling for the person covariate effects of the model. These values are assumed to be sampled from a normal distribution with a mean of zero and a variance of σ_p^2 . ϵ_i is the error term for the specific item being responded to. It can be interpreted as item-specific residuals after item covariates have been controlled for. These item random effects are also assumed to come from a normal distribution with a mean of zero and a variance of σ_i^2 . Treating item effects as random is a somewhat unconventional approach in IRT modeling, but it is particularly useful when used in a model containing item covariates. This is an example of a random-item LLTM in which the random effects are cross-classified rather than nested (De Boeck, 2008; Janssen, Schepers, & Peres, 2004).

Specifying this model is straightforward using SPIRIT’s point-and-click interface. Figure 2 shows how to set up the model. Since the “Item” variable is being treated as a random effect in this model, it is actually necessary to leave the “Items” box blank and to place the “Item” variable in the “ID” box.³

The “Covariates” box is where the fixed effects of interest are included. For this model, this includes an intercept column. SPIRIT does not automatically include an intercept, so if a global intercept/easiness (β_0) is desired, it must manually be included as a covariate in the model. The intercept is important to include here because if it is not added, SPIRIT will calculate a separate easiness estimate for both “I” and “H” items. This would not allow us to directly test the effect of “H” items relative to “I” items. When this β_0 approach is used, one of the item types becomes a reference group, and the difference between the two item groups can be assessed with the β_1 coefficient. The interpretation of this β_1 parameter is identical to that of a regression coefficient for a categorical predictor in a typical regression model.

To manually create the global intercept, a new numeric variable must be created that simply equals “1” for all rows

of the dataset. In this example, the created variable is named “Int” (the name of the variable can be named arbitrarily, as long as it is equal to “1” across all rows). This should be the first variable listed in the “Covariates” box. By manually creating a variable that is a vector of all “1’s”, we are able to trick SPSS into making that predictor act like a global intercept.

The covariates box for this model includes another manually created variable, “DimH.” This column equals 1 for all “H” items and 0 for all “I” items. Creating the variable in this manner allows us to test the difference between the probabilities of endorsing “I” items and “H” items.

The last variable that must be created manually is a variable that is used to test the interaction between child age and item type. We will call this variable “D”; it will be equal to “DimH” \times “Age.” Therefore, this variable will equal 0 for all responses to I items, since DimH equals 0 for these items. For H items, the covariate will equal the child’s age. Including this variable in the covariates box will allow us to examine whether the effect of H vs. I items changes as a function of age. SPSS syntax demonstrating how to construct these new variables using the original simulated ADHD dataset is provided in Appendix B.

In addition to these variables, the sex of the reporter and the mean-centered age of the child are also included in the covariates box. As Figure 2 shows, the final covariates box should include the Int, Age, D, RespGen, and DimH variables. For this model, we are not interested in any of the optional commands listed in the bottom left of the interface.

The SPIRIT output gives the variance estimates for both the person and item random effects, as well as estimates for all fixed effects. The estimated variance for the person residuals is 3.415 (SE = 0.405), while the variance estimate for the item error term is 0.311 (SE = 0.123). The “Fixed Effects” table shows the estimates for the explanatory intercepts and coefficients of interest. These estimates are displayed in Figure 3.

The global intercept is estimated to be -0.555 (SE = 0.239), which is on the log-odds scale since a logistic link function was used. This easiness estimate suggests that for a child with an average age, average reported ADHD behaviors ($\theta_p = 0$), being reported on a typical Inattention item ($\epsilon_i = 0$) by a female reporter, $\eta_{ip} = -0.555$. As described previously, a value on the logit scale can be transformed to the probability scale. Using this conversion, the estimated probability of having a response of “1” to a typical Inattention item for this individual is

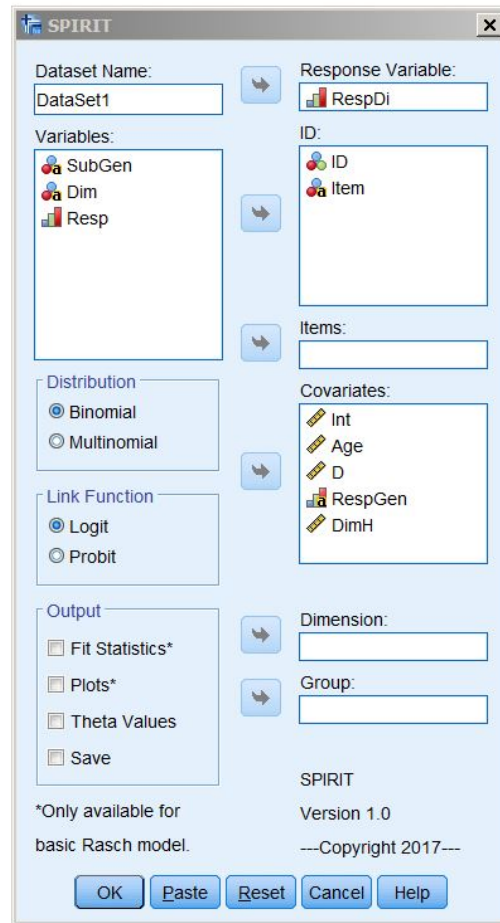
$$\pi_{ip} = \frac{e^{-0.555}}{1 + e^{-0.555}} = 0.365.$$

The difference between the easiness estimates of Hy-

³All random effects need to be placed in the “ID” box.



Figure 2 ■ How to specify the explanatory model described here.



peractivity (H) items relative to Inattention (I) items is estimated to be -0.776 (SE = 0.277). Following, the predicted probability of a “1” on an average “H” item with a female reporter of an average-aged child with average ADHD reported symptoms is 0.210, which corresponds to an η_{ip} value of $-0.555 - 0.776 = -1.326$. It can also be concluded that, for children of an average age, the probability of endorsing the item statement is significantly higher for “I” items relative to “H” items, since the effect of the DimH estimate is negative and significant at the 5% level ($p = 0.005$).

The effect for the sex of the reporter is also included in the “Fixed Effects” output. Here, the sex difference is estimated to be -1.006 (SE = 0.332), suggesting that males tend to be less likely to report ADHD behaviors in their children. This echoes the findings from previous mother-father reporter comparisons, within the context of ADHD rating

scales as well as other behavioral scales. Because the effect is in logits, we can exponentiate the term, $e^{-1.006} = 0.366$, to transform the coefficient to the odds scale. This suggests that males are estimated to be only 0.366 times as likely to report ADHD behaviors in their children relative to their female counterpart. The γ_2 effect for child age is not statistically significant here, as the estimated coefficient is -0.002 ($p = 0.648$). Note here that this estimate is specifically the effect of age for “I” items, as described earlier.

The interaction term between child age and item type was found to be statistically significant in the negative direction ($\hat{\delta} = -0.012, p < 0.001$). This finding implies that child age interacts negatively with Hyperactivity (H) items relative to Inattention (I) items. More specifically, the estimate of the effect of age for “H” items is $\gamma_2 + \delta = -0.002 - 0.012 = -0.014$. Although there is no specific hypothesis test here that explicitly examines whether this -0.014 esti-



Figure 3 ■ Fixed effect estimates given in the output from the explanatory model described here.

		Fixed Effects					
		Beta	Std. Error	t	Sig.	95% Confidence Interval Lower	95% Confidence Interval Upper
Effect	Age	-.002	.0052	-.457	.648	-.013	.008
	D	-.012	.0032	-3.849	.000	-.019	-.006
	DimH	-.776	.2766	-2.807	.005	-1.319	-.234
	F	.000
	Int	-.555	.2394	-2.319	.020	-1.024	-.086
	M	-1.006	.3322	-3.030	.002	-1.658	-.355

mate is significantly different from zero, we do have statistical evidence that the effect of age is more negative for “H” items than “I” items. This would suggest that there are very little age differences in responses to “I” items, but that there is a much more negative effect of age for “H” items: the older a child’s age, the less likely he or she tends to be reported to have Hyperactivity symptoms. From this analysis, we can conclude from this simulated data that male reporters/guardians tend to be less likely to report child ADHD symptoms, that “H” behaviors tend to be reported less frequently than “I” items, and that the effect of age is more negative for “H” behaviors relative to “I” behavior items. This implies that for “H” items, older children are less likely than younger children to have reported ADHD symptoms (i.e., the effect of age is negative). However, no such age association exists for “I” items, resulting in the interaction effect. Although the conclusions from the simulated dataset matches those found using the real data, we intend for these findings to be used for illustrative purposes only.

IRTtree Model Example

The SPIRIT macro can also fit IRT models for polytomous (ordinal) response variables. One traditional model that SPIRIT can run is the rating scale model (Andrich, 1978), which can be viewed as a constrained polytomous item response model in which the distances between an item’s threshold parameters are constrained to be equal for all items in the dataset. This analysis can be performed using SPIRIT by specifying “Multinomial” in the “Distribution” box in the point-and-click interface.

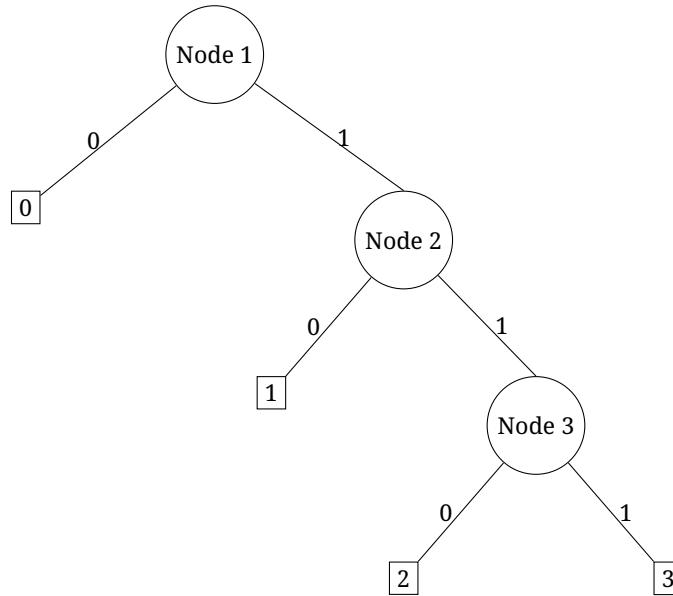
SPIRIT can also fit a variety of item response trees, or “IRTrees” (e.g., De Boeck & Partchev, 2012; Jeon & De Boeck,

2015), which can flexibly model polytomous data. One specific application of IRTrees that investigates polytomous responses in a unique way will be highlighted here. This model treats an ordinal outcome variable as a sequence of dichotomous outcomes; for example, using the ADHD dataset where the possible responses were 0, 1, 2, or 3, a response of “1” corresponds to the respondent first answering greater than 0 (i.e., a response of 1, 2, or 3 instead of exactly 0). Then, given that he/she did not answer with a zero, it is assumed that the respondent then answers with a 1 (i.e., exactly 1 instead of greater than 1). This conceptualization implies that the item response process is a function of sequential choices that the respondent makes. This conceptualization is similar to that of the sequential IRT model (Tutz, 1990). A tree diagram that visually depicts this process is shown in Figure 4. Thinking about ordinal responses in this manner allows for the estimation of item parameters that cannot typically be assessed using traditional IRT approaches. If the sequential IRTree model is fit to the ADHD data, every item will have three item parameters that reflect this sequential process; one corresponds to the probability that a respondent chooses a response greater than “0” versus exactly “0”, the next involves the probability that a respondent answers exactly “1” versus greater than “1”, given that the response is already greater than “0”, and so on. It therefore allows researchers to model polytomous data with software that can only handle dichotomous responses.

Since any ordinal response can be thought of as a series of dichotomous, sequential outcomes, the response can be modeled with a logistic model, making it executable within SPIRIT. This structure can be interesting from a researcher’s perspective because it inherently hypothesizes



Figure 4 ■ A visual depiction of a sequential IRTree model. Each square represents an observed polytomous response, which can be explained as a function of each dichotomous “node.” For example, an observed response of “2” can be recoded as a function of three dichotomous outcomes: a “1” on Nodes 1 and 2, and a “0” on Node 3.



that each response is actually a function of the specific sequential process being modeled. By modeling this process using IRTrees, the fitted model can be used to assess the plausibility of the particular hypothesized response process. Other examples of hypothesized processes that IRTrees can examine are the potential differentiation of fast and slow intelligence (Partchev & De Boeck, 2012; Di-Trapani, Jeon, De Boeck, & Partchev, 2016) and the possible presence of individual extreme response styles on Likert scales (e.g. Böckenholt & Meiser, 2017).

To fit the sequential IRTree model in SPIRIT, the data must manually be altered such that each ordinal response can be described as a series of dichotomous outcomes. Each of these dichotomous scenarios are denoted as “sub-trees”, which consist of “nodes” and “branches.” In Figure 4, each circle in the tree represents a node, and the two lines coming out of the nodes represent the two branches, or possible outcomes, for that specific node. For example, from the earlier illustration, an ordinal response of “1” must be extended in the dataset such that the respondent “successfully” answered with a “1” at node 1 (meaning the observed ordinal response was greater than 0; this is synonymous with saying that the “right branch” was selected at node 1), and then “unsuccessfully” at node 2 (meaning that the observed ordinal response was exactly 1 and

not greater than 1; the “left branch” was selected at node 2). Since the example dataset contains an ordinal variable with 4 possible outcomes, there should be 3 nodes (“0” vs. greater than “0”, “1” vs. greater than “1”, and “2” vs. “3”). Not every response interacts with every node. For example, a response of “1” is only a function of nodes 1 and 2. Table 2 details how each of the four possible responses would be recoded in terms of the three nodes and Table 3 shows how a response of “1” would appear in the dataset. There is now a row for every item and node combination, so three rows per item response for the ADHD dataset.⁴ Once this format is obtained, the SPIRIT macro can be used to fit the model.

Within the new dataset, the response variable now corresponds to the response of respondent p to node d of item i , Y_{idp} . The sequential model follows a similar form as the models introduced earlier, except the linear term, η , now includes a d subscript to indicate the node. For this model, we model the linear function as

$$\eta_{idp} = \beta_{id} + \gamma_1(Male_p) + \theta_p,$$

where β_{id} corresponds to the intercept/easiness for each item-node combination. That is, rather than estimating one intercept for each item, as in the first example, we now have $I \times D$ intercepts, where I is the number of items and

⁴SPSS syntax for appropriately transforming the ADHD data is provided in Appendix B.



Table 2 ■ How responses must be recoded for a sequential model to be implemented. For example, a response of “2” must be recoded as a “1” on nodes 1 and 2, and a “0” on node 3.

Resp	Node 1	Node 2	Node 3
0	0	NA	NA
1	1	0	NA
2	1	1	0
3	1	1	1

Table 3 ■ Three rows corresponding to a response of “1” in the form needed for the sequential model.

ID	Age	SubGen	RespGen	Item	Dim	Node	Item_Node	Resp
1	16.54	F	F	1	I	n1	1_n1	1
1	16.54	F	F	1	I	n2	1_n2	0
1	16.54	F	F	1	I	n3	1_n3	NA

D is the number of nodes in the tree. Here, there will be $D = 3$ item parameters estimated for each of $I = 18$ items. The respondent’s sex is included as a covariate with corresponding effect γ_1 and so θ_p is a person residual term that captures between-person variation not explained by sex.

Besides the addition of the reporter sex covariate, the important difference between this model and the basic 1PL model is that the response variable and intercepts now correspond to particular item-node combinations. Therefore, to fit this model in SPIRIT, the new response variable is provided for the “response” option and the “item” option must be filled with the *Item_Node* variable to allow for a separate intercept for every item/node combination. The id, covariate, and other options are all treated as before. Here, a sequential model is fit using the generated ADHD dataset, with the reporter sex included as a person covariate. A screenshot of how this model is specified within SPIRIT is shown in Figure 5.

The output contains the estimate for the effect of the reporter being male, as well as 54 item intercepts (one for each item/node combination). As in the earlier example model, the male reporter effect is significantly negative, with an estimate of -1.060 (SE = 0.313, $p = 0.001$). The estimate for the random effect variance is 3.416. SPIRIT can also provide the predicted residual θ_p values for each individual (notice the “Theta Values” box was checked in Figure 5). Figure 6 displays the person residual values for several individuals. We can conclude from this model that the reporter sex effect found in the earlier example explanatory model is also evident when treating the responses as ordinal.

The sequential model allows for users to model ordinal responses in an intuitive manner using a logistic link function. It also provides item easiness parameters at every

item/node combination, which give subtly different information about an item than item parameters from a rating scale model. For example, the item response tree model described here would estimate three separate intercepts for item 1: 1.530, -0.545, and -1.675. The 1.530 estimate corresponds to the likelihood that a typical female respondent responds to item 1 with something greater than “0” relative to exactly “0”. In other words, responding with a “0” is somewhat unlikely for this item. The second intercept/easiness, -0.545, gives the tendency for this item to be responded as a “2” or “3” (instead of a “1”), given that the response is not a zero. And finally, the -1.675 estimate represents the chances that a “3” is reported instead of a “2,” conditional on the response not being “0” or “1.” We can see from these estimates that a zero is relatively unlikely, but the higher values of “3” and “2” are somewhat unlikely as well.

This model provides flexibility that the rating scale model (SPIRIT’s other polytomous item response model) cannot. The rating scale model constrains every item’s intercepts to have equal distance from one another, while the sequential IRTree does not. Notice that for the aforementioned item 1, the “distance” between the first and second estimates (1.530 and -0.545) is 2.085. In the rating scale model, this distance between parameters is constrained to be constant across all items, such that the first two intercepts for item 2 (and 3, and 4, and so on) would also have to be exactly 2.085 apart. The sequential IRTree alleviates this constraint and flexibly estimates different intercepts for every item.⁵ Note that other IRT models such as the graded response model or partial credit model do not have this constraint either, but they are currently not available on SPIRIT. However, these existing polytomous models can not perform the specific multidimensionality options de-

⁵The intercepts/thresholds from the rating scale model carry a slightly different meaning as well. See the User Guide for an explanation on SPIRIT’s rating scale model.

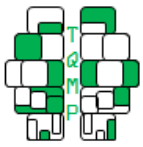
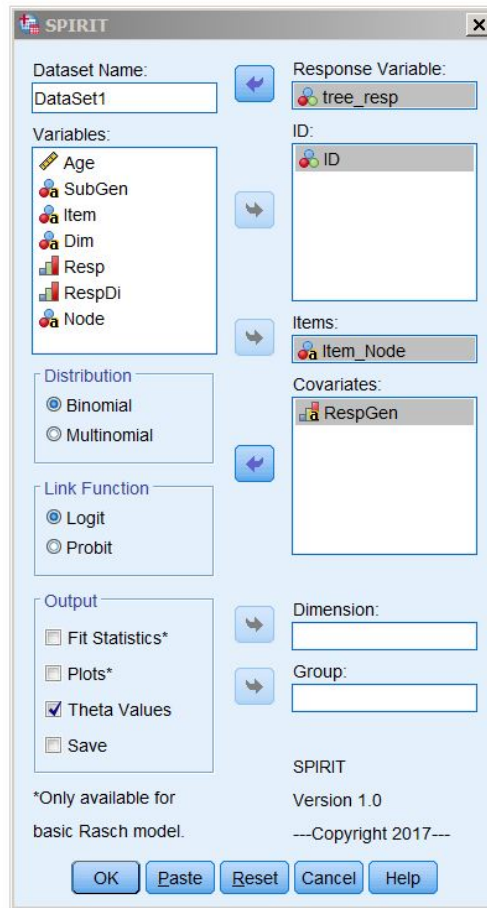


Figure 5 ■ How to specify the sequential IRTree model described here.



scribed in the next paragraph.

IRTrees are very flexible in the types of models that can be run and the types of questions that can be answered. In the above illustration, it is assumed that each “node” is dependent on one person trait: in other words, the probability of a child being scored as greater than “0” (node 1) is dependent on the same person trait that is related to nodes 2 and 3. This constraint can be alleviated, such that each respondent would actually have three different person traits: one for each node. These three dimensions would provide person-specific tendencies that correspond to each of the unique nodes. For example, each respondent would have a “node 1” trait, which can be interpreted as that reporter’s tendency to respond with greater than “0” relative to exactly “0,” and so on. This would be an example of a multidimensional model, where each node is a unique dimen-

sion. This extension of the sequential model can easily be specified on SPIRIT by putting the “Node” variable into the “Dim” option on the point-and-click interface.⁶

Discussion

The SPIRIT macro, which has the capacity to run a multitude of 1PL item response models in SPSS, has been introduced. Using an applied dataset containing generated ADHD scale responses, multiple IRT models have been presented that demonstrate how the SPIRIT macro can be used effectively. It should be emphasized that the IRT framework is not limited to the models illustrated here; SPIRIT allows for a flexible, broad array of potential models, such as basic 1PL models, models with item and/or person covariates, multidimensional IRT models, DIF models, multi-group models, rating scale models, IRTree models, and any

⁶This more complex model does not converge with the provided simulated dataset. Advanced multidimensional models such as this one often require a larger sample size.

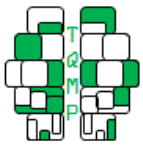


Figure 6 ■ Random effect person estimates calculated from the IRTree model described here.

Person Statistics		
ID		Theta
1		-.66
2		-2.50
3		-2.50
4		-2.96
5		1.94
6		1.33
7		1.54
8		-2.13
9		-.50
10		1.37
11		-.42
12		-1.07
13		-2.05
14		-.44
15		2.52
16		1.16
17		-3.40
18		1.94
19		1.65
20		-1.90
21		1.01
22		-3.40
23		-.86
24		-1.14
25		.22

combination of these. The macro's output provides item parameters, item and person covariate fixed effects, person random effect values (optional), estimated random effect covariance matrices, item and person fit statistics (optional), and other options like item characteristic curves and item information plots. The macro also includes a User Manual that gives detailed instruction for conducting these different analyses. The User Manual can easily be accessed by selecting the "Help" button in SPIRIT's point-and-click box.

One of the most obvious benefits of SPIRIT is that it provides users with an intuitive point-and-click interface within the familiar SPSS environment. The hope is that researchers who are new to IRT or who are worried about learning IRT-specialized software packages can readily perform IRT analyses using the SPIRIT macro. The macro can also be a great pedagogical outlet for instructors introducing IRT to students who are familiar with SPSS.

The macro also provides useful properties for more experienced users. The PQL estimation method that SPSS uses allows for very rapid estimation, meaning even mod-

els with many dimensions can be estimated quickly. For example, a multidimensional IRTree model took just seconds to run in SPIRIT, but took over eight minutes to run in lme4. It is also possible for the estimation of rating scale models for ordinal responses, something that other GLMM software like lme4 (using Laplace estimation) do not currently allow.

An inconvenient feature of the SPIRIT 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, these values are not included in the SPIRIT macro. Since the PQL estimation method is used, these likelihood-based model fit values may be imperfect and therefore not recommended for use (e.g. Van den Noortgate, De Boeck, & Meulders, 2003). A future direction for the macro could be a deeper investigation into the given AIC and BIC values, as well as an exploration into other possible criteria for IRT model fit, such as the M_2 statistic, RMSEA for IRT, etc. (Maydeu-Olivares, 2013).

SPIRIT can also be used to fit IRT models using SPSS syntax, rather than the point-and-click interface described



here. This option may be useful for researchers who would like to save the code used for specifying particular models for later use. Details of this approach and other matters related to SPIRIT can be found in the SPIRIT user manual.

Authors' note

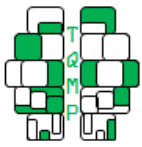
Data tools used in the preparation of this manuscript were obtained from the NIH-supported National Database for Autism Research (NDAR). NDAR is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in autism. Dataset identifier(s): DOI 10.15154/1367677. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDAR. Correspondence should be sent to: Jack DiTrapani, Department of Psychology, Ohio State University, 1827 Neil Avenue, Columbus, OH 43221, email: ditrapani.4@osu.edu. All authors have read and approved the manuscript in its final form. The authors have no conflicts of interest related to this study and did not benefit from any funding.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. doi:10.1177/0146621697211001
- Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised item response modelling software*. 4th ed. Camberwell, Victoria: Australian Council for Educational Research.
- Aggen, S. H., Neale, M. C., & Kendler, K. S. (2005). DSM criteria for major depression: evaluating symptom patterns using latent-trait item response models. *Psychological Medicine, 35*(4), 475–487. doi:10.1017/S0033291704003563
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. doi:10.1007/BF02293814
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.
- Böckenholt, U. & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: a review and tutorial. *British Journal of Mathematical and Statistical Psychology, 70*(1), 159–181. doi:10.1111/bmsp.12086
- Bodì, M. B., Muñoz-Santa, I., Armero, C., Doerr, S. H., Mataix-Solera, J., & Cerdá, A. (2013). Spatial and temporal variations of water repellency and probability of its occurrence in calcareous Mediterranean range-land soils affected by fires. *Catena, 108*(1), 14–25. doi:10.1016/j.catena.2012.04.002
- Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*(421), 9–25. doi:10.1080/01621459.1993.10594284
- Breslow, N. E. & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika, 82*(1), 81–91. doi:10.1093/biomet/82.1.81
- Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows. Version 2.1. Retrieved from <http://www.ssicentral.com/irt/>
- Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. doi:10.18637/jss.v048.i06
- De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533. doi:10.1007/s11336-008-9092-x
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1–28. doi:10.18637/jss.v039.i12
- De Boeck, P. & Partchev, I. (2012). IRTrees: tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*(1), 1–28. doi:10.18637/jss.v048.c01
- De Boeck, P. & Wilson, M. (2004). *Explanatory item response models*. New York, NY: Springer.
- DiTrapani, J. (2016). *IRT in SPSS: the development of a new software tool to conduct item response models* (Master's thesis, The Ohio State University, Columbus, Ohio).
- DiTrapani, J., Jeon, M., De Boeck, P., & Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence, 56*(1). doi:10.1016/j.intell.2016.02.012
- Dragow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x
- Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: a meta-analysis. *Clinical Psychology: Science and Practice, 7*(4), 435–453. doi:10.1093/clipsy.7.4.435
- DuPaul, G. J., Anastopoulos, A. D., Power, T. J., Reid, R., Ikeda, M. J., & McGoey, K. E. (1998). Parent ratings of attention-deficit/hyperactivity disorder symptoms:



- factor structure and normative data. *Journal of Psychopathology and Behavioral Assessment*, 20(1), 83–102. doi:[10.1023/A:1023087410712](https://doi.org/10.1023/A:1023087410712)
- Edelen, M. O. & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1), 5–18. doi:[10.1007/s11136-007-9198-0](https://doi.org/10.1007/s11136-007-9198-0)
- Embretson, S. E. & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37(6), 359–374. doi:[10.1016/0001-6918\(73\)90003-6](https://doi.org/10.1016/0001-6918(73)90003-6)
- Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: lessons learned and paths forward. *Organizational Research Methods*, 20(3), 1–22. doi:[10.1177/1094428116689708](https://doi.org/10.1177/1094428116689708)
- Fox, J. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software*, 20(5), 1–16. doi:[10.18637/jss.v020.i05](https://doi.org/10.18637/jss.v020.i05)
- Hagquist, C., Bruce, M., & Gustavsson, J. P. (2009). Using the Rasch model in nursing research: an introduction and illustrative example. *International Journal of Nursing Studies*, 46(3), 380–393. doi:[10.1016/j.ijnurstu.2008.10.007](https://doi.org/10.1016/j.ijnurstu.2008.10.007)
- Houts, C. R. & Cai, L. (2013). *flexMIRT® user's manual version 2: flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.
- Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: a Monte Carlo study. *Applied Psychological Measurement*, 6(3), 249–260. doi:[10.1177/014662168200600301](https://doi.org/10.1177/014662168200600301)
- IBM Corp. (Released 2013). *IBM SPSS Statistics for Windows, Version 22*. Armonk, NY: IBM Corp.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In *Explanatory item response models* (pp. 189–212). Springer.
- Jeon, M. & De Boeck, P. (2015). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48(3). doi:[10.3758/s13428-015-0631-y](https://doi.org/10.3758/s13428-015-0631-y)
- Jeon, M., Rijmen, F., & Rabe-Hasketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Measurement*, 38(5), 404–405. doi:[10.1177/0146621614524982](https://doi.org/10.1177/0146621614524982)
- Kiefer, T., Robitzsch, A., & Wu, M. (2016). *TAM: test analysis modules*. R package version 1.17-0.
- King, J., Bond, T., & Blandford, S. (2002). An investigation of computer anxiety by gender and grade. *Computers in Human Behavior*, 18(1), 69–84. doi:[10.1016/S0747-5632\(01\)00030-9](https://doi.org/10.1016/S0747-5632(01)00030-9)
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328.
- Linacre, J. M. (2016). *Winsteps Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.
- Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: the eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9). doi:[10.18637/jss.v020.i09](https://doi.org/10.18637/jss.v020.i09)
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, 11(3), 71–101. doi:[10.1080/15366367.2013.831680](https://doi.org/10.1080/15366367.2013.831680)
- Muthen, L. K. & Muthen, B. O. (1998-2011). *Mplus user's guide*. Sixth. Los Angeles, CA: Muthen & Muthen.
- Nijsten, T., Unaeze, J., & Stern, R. S. (2006). Refinement and reduction of the Impact of Psoriasis Questionnaire: Classical Test Theory vs. Rasch analysis. *British Journal of Dermatology*, 154(4), 692–700. doi:[10.1111/j.1365-2133.2005.07066.x](https://doi.org/10.1111/j.1365-2133.2005.07066.x)
- Partchev, I. & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, 40(1), 23–32. doi:[10.1016/j.intell.2011.11.002](https://doi.org/10.1016/j.intell.2011.11.002)
- Rabe-Hasketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM manual. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 160.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research.
- Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8(2), 185–205. doi:[10.1037/1082-989X.8.2.185](https://doi.org/10.1037/1082-989X.8.2.185)
- Rizopoulos, D. (2006). Ltm: an R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. doi:[10.18637/jss.v017.i05](https://doi.org/10.18637/jss.v017.i05)
- SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.
- Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, 50, 219–232.
- Seiffge-Krenke, I. & Kollmar, F. (1998). Discrepancies between mothers' and fathers' perceptions of sons' and daughters' problem behaviour: a longitudinal analysis of parent-adolescent agreement on internalising and externalising problem behaviour. *Journal of Child Psychology and Psychiatry*, 39(5), 687–697. doi:[10.1111/1469-7610.00368](https://doi.org/10.1111/1469-7610.00368)



- Small, D. M., Simons, A. D., Yovanoff, P., Silva, S. G., Lewis, C. C., Murakami, J. L., & March, J. (2008). Depressed adolescents and comorbid psychiatric disorders: are there differences in the presentation of depression? *Journal of Abnormal Child Psychology*, 36(7), 1015–1028. doi:[10.1007/s10802-008-9237-5](https://doi.org/10.1007/s10802-008-9237-5)
- Sollie, H., Larsson, B., & Mørch, W. T. (2012). Comparison of mother, father, and teacher reports of ADHD core symptoms in a sample of child psychiatric outpatients. *Journal of Attention Disorders*, 17(8), 699–710. doi:[10.1177/1087054711436010](https://doi.org/10.1177/1087054711436010)
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, 43(1), 39–55. doi:[10.1111/j.2044-8317.1990.tb00925.x](https://doi.org/10.1111/j.2044-8317.1990.tb00925.x)
- Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, 28(4), 369–386. doi:[10.3102/10769986028004369](https://doi.org/10.3102/10769986028004369)
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97–116. doi:[10.1111/j.1745-3984.1977.tb00031.x](https://doi.org/10.1111/j.1745-3984.1977.tb00031.x)
- Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Models*. ERIC.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Appendix A: SPIRIT Setup and Estimation

SPIRIT Model Implementation

- Data:** For SPSS to correctly run its GENLINMIXED function, data must be presented in long form. This is true for any GLMM 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. The VARSTOCASES function in SPSS can be used to restructure a dataset from wide to long form.
- Installation:** To install SPIRIT, the user must download the SPIRIT macro .spd file that contains the point-and-click dialog.⁷ The user will then be prompted to “install” the dialog once this file is opened. When it is installed, the user can find the SPIRIT interface within the “Analyze” dropdown from SPSS’s dropdown menu. Once the SPIRIT interface is installed, it will never have to be installed again into that computer.
- Output View:** Before any analysis can be run in SPIRIT, the user must turn off SPSS’s 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 useful output. This change should only need to be made once; “Pivot Tables” will remain the selected option moving forward, even if the user ends the current session. Once this change is made, models are ready to be specified using SPIRIT’s point-and-click approach.
- Specification:** Model specification is relatively straightforward. Within the point-and-click interface, there is an option to specify the response variable of the model, the item variable of the model, other person or item covariates that may be included, the person random effect (as well as other potential random effects), a variable that indicates the dimension of a particular response, a variable that indicates the group of a particular respondent, the link function of the model, and the distribution of the response variable. There are also options available for outputting predicted random effect values, fit statistics, and whether or not to save output values into the original dataset. All of these options can be seen in Figure 7. The response variable box is the only box that must be filled with a variable; the other boxes can be filled or left empty, depending on the user’s desired model.

Model Estimation

SPSS uses a penalized quasi-likelihood (PQL) approach to estimating GLMMs (Breslow & Clayton, 1993). Methods using quasi-likelihood techniques have been shown to have 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. However, this bias does not typically bring substantial practical differences in model interpretation when SPIRIT is used.

⁷This file can be obtained by emailing a request to the corresponding author.

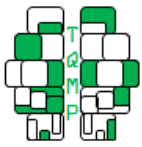
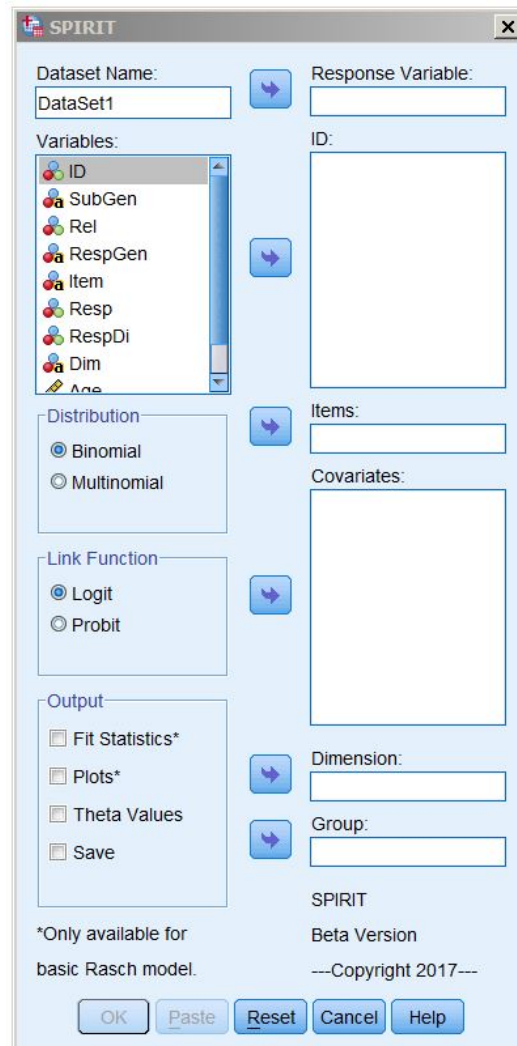


Figure 7 ■ The general point-and-click layout for the SPIRIT macro.



Appendix B: SPSS Syntax for Illustrations

Basic 1PL Model

Once the ADHD dataset is open (and the SPIRIT macro code has been run), the following SPSS syntax will execute the basic 1PL model.

```
spirit response = RespDi/items = Item/id = ID/theta=YES.
```

Explanatory Model

Once the ADHD dataset is opened, the following SPSS code will create the variables necessary to run the explanatory model described in the manuscript. The “spirit” line of code will then run the explanatory model, given that the SPIRIT macro code has already been run.

```
Compute Int = 1.
```



```
Compute DimH = 0.
IF Dim EQ 'H' DimH = 1.
```

```
Compute D = DimH*Age.
Variable Level D (Scale).
Variable Level DimH (Scale).
```

Execute.

```
spirit response = RespDi/cov = Int Age D RespGen DimH/id = ID Item/theta = NO.
```

IRTree Model

Once the ADHD dataset is opened, the following SPSS code will create the variables necessary to run the IRTree model described in the manuscript. The “spirit” line of code will then run the IRTree model, given that the SPIRIT macro code has already been run.

```
* Compute n1, n2, and n3 (the different node responses).
RECODE Resp (0=0) (1 thru Highest=1) INTO n1.
EXECUTE.
RECODE Resp (0=SYSMIS) (1=0) (2 thru Highest=1) INTO n2.
EXECUTE.
RECODE Resp (2=0) (3=1) (Lowest thru 1=SYSMIS) INTO n3.
EXECUTE.
```

```
* Restructure data so row for every item_node combo.
VARSTOCASES
  /MAKE tree_resp FROM n1 n2 n3
  /INDEX=Node(tree_resp)
  /KEEP=ID Age SubGen RespGen Item Dim Resp RespDi
  /NULL=KEEP.
```

```
* Create item_node identifier.
STRING Item_Node (A20).
COMPUTE Item_Node=CONCAT (RTRIM(Item) , "_", RTRIM(Node)).
EXECUTE.
```

```
spirit response = tree_resp/id = ID/items = Item_Node/cov = RespGen/theta = YES.
```

Open practices

📄 The *Open Material* badge was earned because supplementary material(s) are available on the [journal's web site](#).

Citation

DiTrapani, J., Rockwood, N., & Jeon, M. (2018). Explanatory IRT analysis using the SPIRIT macro in SPSS. *The Quantitative Methods for Psychology, 14*(2), 81–98. doi:[10.20982/tqmp.14.2.p081](https://doi.org/10.20982/tqmp.14.2.p081)

Copyright © 2018, DiTrapani, Rockwood, and Jeon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 02/11/2017 ~ Accepted: 12/02/2018