Ordinal Regression Models in Psychological Research: A Tutorial

Paul-Christian Bürkner[1] & Matti Vuorre[2]

[1] Department of Psychology, University of Münster, Germany

[2] Department of Psychology, Columbia University, USA

Author Note

Correspondence concerning this article should be addressed to Paul-Christian Bürkner, Department of Psychology, University of Münster, Fliednerstrasse 21, 48149 Münster, Germany. E-mail: paul.buerkner@gmail.com

Abstract

Ordinal variables are widely used in psychological research, especially in the form of Likert items. Such data are still almost exclusively analysed with statistical models that falsely assume the ordinal variables to be metric. This practice can lead to problems such as distorted effect size estimates and inflated error rates. Therefore, we argue for the application of more appropriate ordinal models that make reasonable assumptions about the ordinal variables under study. From both theoretical and applied perspectives, we explain the ideas behind three major ordinal model classes; the cumulative, sequential and adjacent category models. We then use data sets on stem cell opinions, confidence ratings, and marriage time courses to show how to fit ordinal models in a fully Bayesian framework with the R package brms. Ordinal models provide better theoretical interpretation and numerical inference from ordinal data, and we recommend their widespread adoption. To this end, we provide guidelines for the application of ordinal models in psychological research.

Keywords: ordinal models, Likert items, signal detection theory, brms, R

Ordinal Regression Models in Psychological Research: A Tutorial

1 Introduction

Whenever a variable's categories have a natural order, we speak of an *ordinal* variable (Stevens, 1946). In psychology, analyzing ordinal data has always been of high relevance, and ordinal data is ubiquitous: Almost all data gathered with questionaires using Likert-type scales are ordinal. However, assuming these variables to be *metric* is inherently problematic. As demonstrated by Liddell and Kruschke (2017), analysing ordinal data with statistical models that assume metric variables, such as t-tests and ANOVA, can lead to low correct detection rates, distorted effect size estimates, and inflated false alarm (type-I-error) rates – a problem that cannot be solved by simply averaging over multiple ordinal items. Historically, the possibilites of analysing ordinal data were rather limited, although simple analyses – such as the comparison between two groups – could be performed with non-parametric approachs (Gibbons & Chakraborti, 2011). However, for more complex analyses – regression-like methods, in particular – there were few alternatives to incorrectly treating ordinal data as either continuous or nominal. In practice, choosing a continuous or nominal model has led to over- or under-estimating (respectively) the information provided by the data.

Fortunately, recent advances in statistics and statistical software have provided researchers with many options for approriate models of ordinal data, in particular when it comes to modeling ordinal responses. Such methods are often summarized under the term *ordinal regression models*. Still, application of these methods has remained very limited, while the use of less appropriate linear regression for modeling ordinal data remains widespread (Liddell & Kruschke, 2017). Several reasons may underlie this persistence with linear models for ordinal data: For instance, researchers might not be aware of more appropriate methods, or they may hesitate to use them because of their perceived complexity. This applies both to model fitting and interpretation of the results. Moreover, since closely related (or even the same) ordinal models are called with very different names depending on the context in which they are introduced, it may be difficult for researchers to decide which ordinal model is most reasonable for their data. Finally, researchers may also feel compelled to use "standard" analyses, even if "standard" means less appropriate linear models for ordinal data, because journal editors and reviewers may be sceptical of any "non-standard" approaches. To summarize, there is need for better explanation and more examples of ordinal data and models to facilitate the use of ordinal models in psychological research. We hope that the present tutorial proves helpful in this regard.

The structure of this paper is as follows. In Section 2, we introduce three data sets serving as motivating examples for the use of ordinal models in psychology, followed by a detailed derivation of ordinal model classes in Section 3. We continue with fitting ordinal models on the sample data sets using the R statistical computing environment (R Core Team, 2017) in Section 4, and end with guidelines for using ordinal models and a conclusion in Section 5.

2 Motivating examples

Ordinal data is ubiquitous in psychological research. In this section, we present three representative real-world data sets from different areas of psychology that contain ordinal variables as the main dependent variable, and therefore would benefit from application of appropriate ordinal models.

2.1 Opinion about funding stem cell research

The first data set is part of the 2006 US General Society Survey (http://gss.norc.org/) and contains variables on the respondents' opinion about funding stem cell research, the fundamentalism / liberalism of their religious beliefs, and gender (Agresti, 2010). We wish to investigate to what extent fundamentalism and gender predict opinions about funding stem cell research. Here, opinion about funding stem cell research serves as the dependent variable. It was assessed on a four point Likert-scale with the anchors "definitely fund" (1), "probably fund" (2), "probably not fund" (3), "definitely not fund" (4). Clearly, this is an ordinal variable: We know the order of the categories, but we do not know if they are equidistant in the participants' minds, nor if the distances are the same across participants. Such variables – with typically about 3 to 7 response categories – are extremely common in psychology. They are usually analyzed with linear models (Liddell & Kruschke, 2017), possibly because of a *perceived* lack of alternative methods. However, the assumptions of linear models are violated, because we cannot assume ordinal variables to be continuous and certainly not normally distributed. An overview of the data is provided in Table 1.

Table 1

Frequencies of opinion about funding stem cell research

	male				female			
	1	2	3	4	1	2	3	4
fundamentalist	21	52	24	15	34	67	30	25
moderate	30	52	18	11	41	83	23	14
liberal	64	50	16	11	58	63	15	12

2.2 Recognition memory confidence ratings

The second data set comes from a recognition memory experiment, where participants rated their confidence in whether presented words were previously studied or not (Koen, Aly, Wang, & Yonelinas, 2013). We use this data set to illustrate the applicability of the ordinal regression modeling framework to Signal Detection Theoretic models (SDT; Macmillan and Creelman (2005)). SDT is a widely used cognitive model that allows separating participants' task abilities from response criteria. In the experiment, participants first studied a list of 200 words, and then completed a recognition test in two conditions: full attention and divided attention (Experiment 2 in Koen et al. (2013)). We focus on the full attention condition. In the recognition test, participants saw 100 old words from the previously studied list, and 100 new words, one at a time. For each word, they rated their confidence in whether the word was new or old (1 = *sure new*, 6 = *sure old*). These data are summarized in Table 2. It would be problematic to assume that the confidence ratings constitute a continuous and normally distributed variable and subsequently apply ordinary linear regression methods. Instead, the ratings are ordinal categories and are therefore naturally modeled in the ordinal regression framework. Importantly, as we explain below, this framework can be used to easily implement the useful equal and unequal variance SDT models.

Table 2

Recognition memory confidence ratings (Koen et al., 2013)

	1	2	3	4	5	6
new	1365	1335	871	454	356	379
old	309	422	389	384	634	2604

2.3 Years until divorce

The third example comes from the US National Survey of Family Growth 2013 - 2015 (NSFG; https://www.cdc.gov/nchs/nsfg), in which data were gathered about family life, marriage and divorce for over 10000 individuals (among other variables). For the purpose of the present tutorial, we will focus on a subsample of 1597 women, who had been married at least once in their life at the time of the survey. Inspired by Teachman (2011), who used the NSFG 1995 data, we are interested in predicting the duration (in years) of first marriage (`ma_years`), which ends either by divorce or continues beyond the time of the survey. We can understand this as time-to-event data, with the event of interest being divorce. As predictors we will use the participants' age at marriage (`age_at_ma`), whether the couple was already living together before marriage (`liv_together`) and whether the husband had been previously married (`hus_ma_before`). We illustrate the first ten rows of the data in Table 3. Most of the common methods for analysing time-to-event data such as Cox proportional hazard models (Cox, 1992) assume time to be continuous. However, since we only have information on a yearly basis, a continuous approximation may be problematic (Tutz &

Schmid, 2016). Accordingly, we will use a discrete time-to-event approach by means of ordinal models.

Table 3

Overview of marriage data from the NSFG 2013-2015 survey.

ID	age_at_ma	hus_ma_before	liv_together	divorced	ma_years
1	19	no	yes	TRUE	9
2	22	no	yes	FALSE	9
3	20	no	yes	FALSE	5
4	22	no	yes	FALSE	2
5	25	no	yes	FALSE	6
6	30	no	yes	FALSE	1
7	32	no	yes	FALSE	9
8	24	no	no	TRUE	14
9	37	yes	no	TRUE	1
10	18	yes	yes	TRUE	13

Below, we use these three data sets to illustrate ordinal modeling in practice. However, we remind the readers that ordinal data is not limited to the types of variables introduced here, but can actually be found in a wide variety of research areas, as noted by Stevens in a seminal paper (1946): “As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales” (p.679). But before our example analyses, we begin by a detailed derivation and theoretical motivation for the various ordinal models.

3 Derivations of the ordinal model classes

A large number of parametric ordinal models can be found in the literature. To the confusion of anyone seeking to apply these models, they all have their own names, and their interrelations are often left completely unclear. Fortunately, the vast majority of these models can be expressed within a framework of three distinct model classes (Mellenbergh, 1995; Molenaar, 1983; Van Der Ark, 2001). These are the *Cumulative Model* (CM), the *Sequential Model* (SM), and the *Adjacent Category Model* (ACM), which we introduce in this section. Throughout, we assume to have observed a total of N values of the ordinal response variable Y with $K+1$ categories from 0 to K.

3.1 Cumulative model

The CM, sometimes also called *graded response model* (Samejima, 1997), assumes that the observed ordinal variable Y originates from the categorization of a latent (i.e. not observable) continuous variable $\tilde{Y}$. That is, there are latent thresholds τ_k $(1 \leq k \leq K)$, which partition the values of $\tilde{Y}$ into the $K+1$ observable, ordered categories of Y. More formally

$$Y_n = k \Leftrightarrow \tau_k < \tilde{Y}_n \leq \tau_{k+1} \tag{1}$$

for each observation n and $-\infty = \tau_0 < \tau_1 < ... < \tau_K < \tau_{K+1} = \infty$. We write $\tau = (\tau_1, ..., \tau_K)$ for the vector of the thresholds. As explained above, it may not be valid to use linear regression on Y, because the differences between its categories are not known. However, linear regression is applicable to $\tilde{Y}$. Using η_n to symbolize the predictor term for the nth observation leads to

$$\tilde{Y}_n = \eta_n + \varepsilon_n, \tag{2}$$

where ε_n is the random error of the regression with $E(\varepsilon_n) = 0$. In the simplest case, η_n is a linear predictor of the form $\eta_n = X_n\beta = X_{n1}\beta_1 + X_{n2}\beta_2 + ... + X_{nm}\beta_m$, with m predictor variables $X_n = (X_{n1}, ..., X_{nm})$ and corresponding regression coefficients $\beta = (\beta_1, ..., \beta_m)$ (without an intercept). The predictor term η_n may also take more complex forms—for instance, multilevel structures or non-linear relationships. However, for the understanding of ordinal models, the exact form of η_n is irrelevant, and we can assume it to be linear for now.

To complete model (2), the distribution F of ε_n has to be specified. We might use the normal distribution because it is the default in linear regression, but alternatives such as the logistic distribution are also possible. As explained below, these alternatives are often more appealing then the normal distribution. Depending on the choice of F, the final model for $\tilde{Y}$ and also for Y will vary. At this point in the paper, we do not want to narrow down our modeling flexibility and therefore just assume that ε_n is distributed according to F:

$$\Pr(\varepsilon_n \leq z) = F(z). \tag{3}$$

Combining the assumptions (1), (2), and (3) leads to

$$\begin{aligned}\Pr(Y_n \leq k|\eta_n) &= \Pr(\tilde{Y}_n \leq \tau_{k+1}|\eta_n) = \Pr(\eta_n + \varepsilon_n \leq \tau_{k+1}) \\ &= \Pr(\varepsilon_n \leq \tau_{k+1} - \eta_n) = F(\tau_{k+1} - \eta_n).\end{aligned} \tag{4}$$

The notation $|\eta_n$ in the first two terms of (4) means the the probabilities will depend on the values of the predictors $X_1, ..., X_m$ for the nth observation. Equation (4) says that the probability of Y_n being in category k or less (depending on η_n) is equal to the value of the distribution F at the point $\tau_{k+1} - \eta_n$. In this context, F is also called a *response function* or processing function. In the present paper, we will use the term distribution and response function interchangeable, when talking about F. In case of the CM, F models the probability of the binary outcome $Y_n \leq k$ against $Y_n > k$, thus motivating the name "cumulative model".

The probabilities $\Pr(Y = k|X)$, which are of primary interest, can be easily derived from (4), since

$$\begin{aligned}\Pr(Y_n = k|\eta_n) &= \Pr(Y_n \leq k|\eta_n) - \Pr(Y_n \leq k-1|\eta_n) \\ &= F(\tau_{k+1} - \eta_n) - F(\tau_k - \eta_n).\end{aligned} \tag{5}$$

The CM as formulated in (5) assumes that the regression parameters β are constant across the response categories. It is plausible that a predictor may have, for instance, a higher impact on the lower categories of an item than on its higher categories. Thus, we could write β_k to obtain a single regression parameter per category for every predictor. For instance, if we had 4 categories while using 2 predictors, we would have $3 \times 2 = 6$ regression parameters instead of just 2. Admittedly, the β_k-model is not very parsimonious. Furthermore, estimating regression parameters as varying across response categories in the CM is not always possible, because it may result in negative probabilities (Tutz, 2000; Van Der Ark, 2001). Accordingly, we will have to assume β to be constant across categories when using the CM.

The threshold parameters τ_k, however, are estimated for each category separately, leading to a total of K threshold parameters. This does not mean that it is always necessary to estimate so many of them: We can assume that the distance between two adjacent thresholds τ_k and τ_{k+1} is the same for all thresholds, which leads to

$$\tau_k = \tau_1 + (k-1)\delta. \tag{6}$$

Accordingly, only τ_1 and δ have to be estimated. Parametrizations of the form (6) are often referred to as *Rating Scale Models* (RSM) (Andersen, 1977; Andrich, 1978a, 1978b) and can be used in many IRT and regression models not only in the CM. When several items each with several categories are administered, this leads to a remarkable reduction in the number of threshold parameters. Consider an example with 7 response categories. Under the model (5) we thus have 6 threshold parameters. Using (6) this reduces to only 2 parameters. The discrepancy will get even larger for an increased number of categories. More details about different parametrizations of the CM can be found, among others, in (Samejima, 1969, 1972, 1995, 1997). Note that in regression models, the threshold parameters are usually of subordinate interest as they only serve as intercept parameters. For this reason, restrictions to τ_k such as (6) are rarely applied in regression models.

The derivation and formulation of the general CM presented in this paper is from Tutz (2000), which was published in German language only. Originally, the CM was first proposed by Walker and Duncan (1967) but only in the special case where F is the standard logistic distribution, that is where

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}, \tag{7}$$

(see Figure 1, green line). This special model was later called *Proportional Odds Model (POM)* by McCullagh (1980) and is the most frequently used version of the CM (McCullagh, 1980; Van Der Ark, 2001). In many articles, the CM is directly introduced as the POM and the possibility of using response functions other than the logistic distribution is ignored (Ananth & Kleinbaum, 1997; Guisan & Harrell, 2000; Van Der Ark, 2001), thus hindering the general understanding of the CM's ideas and assumptions.

The name of the POM stems from the fact that under this model, the odds ratio of $\Pr(Y_n \leq k_1 | \eta_n)$ against $\Pr(Y_n \leq k_2 | \eta_n)$ for any $1 \leq k_1, k_2 \leq K$ is independent of η_n and only depends on the distance of the thresholds τ_{k_1} and τ_{k_2}, which is often called the proportional odds assumption[1]:

$$\frac{\Pr(Y_n \leq k_1 | \eta_n) / \Pr(Y_n > k_1 | \eta_n)}{\Pr(Y_n \leq k_2 | \eta_n) / \Pr(Y_n > k_2 | \eta_n)} = \exp(\tau_{k_1} - \tau_{k_2}). \tag{8}$$

Another CM version, the *Proportional Hazards Model (PHM)*, is derived when F is the extreme value distribution (Cox, 1972; McCullagh, 1980):

$$F(x) = 1 - \exp(-\exp(x)) \tag{9}$$

(see Figure 1, red line). This model was originally invented in the context of survival analysis for discrete points in time. It is also possible to use the standard normal distribution

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz. \tag{10}$$

[1] The proportional odds assumption can explicitly be tested by comparing the POM when β is constant across categories then when it is not (but consider the above described problems of category-specific parameters in the CM). The latter model is often called *partial* POM (Peterson & Harrell, 1990).

as a response function (see Figure 1, blue line). This is a common choice in signal detection theoretic models. Of course, one can use other distributions for F as well.

Following the conventions of generalized linear models, we will often use the name of the inverse distribution function F^{-1}, called the link-function, instead of the name of F itself. The link functions associated with the logistic, normal, and extreme value distributions are called *logit*-, *probit*, and *cloglog*-link, respectively.

Applying the CM with different response functions to the same data will often lead to similar estimates of the parameters τ and β as well as to similar model fits (McCullagh, 1980), so that the decision of F usually has only a minor impact on the results.

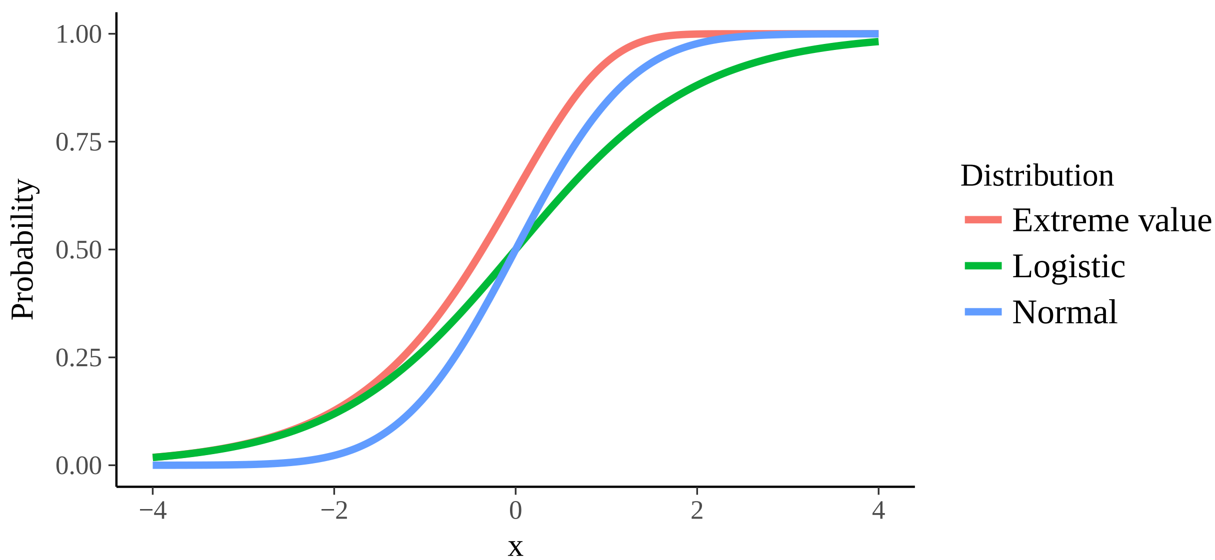

Figure 1. Illustration of various choices for the distribution function F.

The derivation of the CM advocated in the present paper demonstrates that this model is especially appealing when the ordinal data Y can be understood as a categorization of a continuous latent variable $\tilde{Y}$, because the thresholds τ_k have an intuitive meaning in this case. However, the CM is also applicable when this assumption seems unreasonable. In particular, the regression parameters β (and inferences about them) remain interpretable in the same way as before (McCullagh, 1980).

3.2 Sequential Model

For many ordinal variables, the assumption of a single underlying, continuous variable may not be fully appropriate. The depending variable Y in this example results from a counting process and is truly ordinal in the sense that in order to achieve a category k, one has to achieve all lower categories 0 to $k-1$, first. The *Sequential Model* (SM) in its generality proposed by Tutz (1990) explicitly incorporates this structure into its assumptions (see also, Tutz, 2000). For every category $k \in \{0, ..., K-1\}$ there is a latent continuous variable $\tilde{Y}_k$ mediating the transition between the kth and the $k+1$th category. The variables $\tilde{Y}_k$ may have different meanings depending on the research question. We assume that $\tilde{Y}_k$ depends linearly on the predictors $X_1, ..., X_M$, i.e.

$$\tilde{Y}_{nk} = \eta_n + \varepsilon_n. \tag{11}$$

for each observation n. As for the CM, ε_n has mean zero and is distributed according to F:

$$\Pr(\varepsilon_n \leq z) = F(z). \tag{12}$$

The sequential process itself is thought as follows: Beginning with category 0 it is checked whether $\tilde{Y}_{n0}$ surpasses the first threshold τ_1. If not, i.e. if $\tilde{Y}_{n0} \leq \tau_1$, the process stops and the result is $Y_n = 0$. If $\tilde{Y}_{n0} > \tau_1$, at least category 1 is achieved (i.e. $Y_n \geq 1$) and the process continuous. Then, it is checked whether $\tilde{Y}_{n1}$ surpasses threshold τ_2. If not, the process stops with result $Y_n = 1$. Else, the process continues with $Y_n \geq 2$. Extrapolating this to all categories $k \in \{0, ..., K-1\}$, the process stops with result $Y_n = k$, when at least category k is achieved, but $\tilde{Y}_{nk}$ fails to surpass the $k+1$th threshold. This event can be written as

$$Y_n = k | Y_n \geq k. \tag{13}$$

Combining assumptions (11), (12), and (13) leads to

$$\begin{aligned}\Pr(Y_n = k|Y_n \geq k, \eta_n) &= \Pr(\tilde{Y}_{nk} \leq \tau_{k+1}|\eta_n) \\ &= \Pr(\eta_n + \varepsilon_n \leq \tau_{k+1}) \\ &= \Pr(\varepsilon \leq \tau_{k+1} - \eta_n) \\ &= F(\tau_{k+1} - \eta_n). \qquad (14)\end{aligned}$$

Equation (14) we can equivalently be expressed by

$$\Pr(Y_n = k|\eta_n) = F(\tau_{k+1} - \eta_n) \prod_{j=1}^{k} (1 - F(\tau_j - \eta_n)). \qquad (15)$$

Because of its derivation, this model is sometimes also called the *stopping model*. A related sequential model was proposed by Verhelst, Glas, and De Vries (1997) in IRT notation focusing on the logistic response function only. Instead of modeling the probability (14) of the sequential process to *stop* at category k, they suggested to model the probability of the sequential process to *continue* beyond category k. In our notation, this can generally be written as

$$\Pr(Y_n \geq k|Y_n \geq k-1, k > 0, \eta_n) = F(\eta_n - \tau_k) \qquad (16)$$

or equivalently

$$\Pr(Y_n = k|\eta_n) = (1 - F(\eta_n - \tau_{(k+1)})) \prod_{j=1}^{k} F(\eta_n - \tau_j). \qquad (17)$$

In the following, model (15) is called SMS and model (17) is called SMC. When F is symmetric, SMS and SMC are identical, because of the relation $F(-x) = 1 - F(x)$ holding for symmetric distributions. Both, the normal and logistic distribution (10) and (7) are symmetric. Thus, there is only one SM for these distributions. The SM combined with the logistic distribution is often called *Continuation Ratio Model* (CRM) (Fienberg, 1980, 2007).

An example of an asymmetric response function is the extreme value distribution (9). In this case, SMS and SMC are different from each other, but surprisingly, SMS is equivalent to CM (Läärä & Matthews, 1985). That is, the PHM (Cox, 1972) arises from both, cumulative and sequential modeling assumptions.

Despite their obvious relation, SMS and SMC are discussed independently in two adjacent chapters in the handbook of Linden and Hambleton (1997; Tutz, 1997; see also, Verhelst et al., 1997), leading to the impression of two unrelated models and, possibly, some confusion. This underlines the need of a unified wording and notation of ordinal models, in order to facilitate their understanding and practical use.

In the same way as for the CM, the regression parameters β may depend on the categories when using the SM. In contrast to the CM, however, estimating different regression parameters per category is usually less of an issue for the SM (Tutz, 1990, 2000). However, such a model may still be unattractive due to the high number of parameters. Of course, restrictions to the thresholds τ_k such as the rating scale restriction (6) are also applicable. Although the SM is particularly appealing when Y can be understood as the result of a sequential process, it is applicable to all ordinal dependent variables regardless of their origin.

3.3 Adjacent Category Model

The *Adjacent Category Model* (ACM) is somewhat different than the CM and SM, because, in our opinion, it has no satisfying theoretical derivation. For this reason, we discuss the ideas behind the ACM after introducing its formulas. The ACM is defined as

$$\Pr(Y_n = k | Y \in \{k-1, k\}, k > 0, \eta_n) = F(\eta_n - \tau_k) \tag{18}$$

(Agresti, 1984, 2010), that is it describes the probability that category k rather than category $k-1$ is achieved. This can equivalently be written as

$$\Pr(Y_n = k|\eta_n) = \frac{\prod_{j=1}^{k} F(\eta_n - \tau_j) \prod_{j=k+1}^{K} (1 - F(\eta_n - \tau_j))}{\sum_{r=0}^{K} \prod_{j=1}^{r} F(\eta_n - \tau_j) \prod_{j=r+1}^{K} (1 - F(\eta_n - \tau_j))}, \tag{19}$$

with $\prod_{j=1}^{0} F(\eta_n - \tau_j) := 1$ for notational convenience. To our knowledge, the ACM has almost solely been applied with the logistic distribution (7). This combination is the *Partial Credit Model* (PCM; also called Rasch Rating Model)

$$\Pr(Y_n = k|\eta_n) = \frac{\exp\left(\sum_{j=1}^{k} (\eta_n - \tau_j)\right)}{\sum_{r=0}^{K} \exp\left(\sum_{j=1}^{r} (\eta_n - \tau_j)\right)} \tag{20}$$

(with $\sum_{j=1}^{0} (\eta_n - \tau_j) := 0$), which is arguably the most widely known ordinal model in psychological research. It was first derived by Rasch (1961) and subsequently by Andersen (1973), Andrich (1978a), Masters (1982), and Fischer (1995) each with a different but equivalent formulation (Adams, Wu, & Wilson, 2012; Fischer, 1995). Andersen (1973) and Fischer (1995) derived the PCM in an effort to find a model that allows the independent estimation of person and item parameters – a highly desirable property – for ordinal variables. Thus, their motivation for the PCM was purely mathematical and no attempt was made to justify the it theoretically.

On the contrary, Masters (1982) advocated an heuristic approach to the ACM (formulated as the PCM only) by presenting it as the result of a sequential process. In our opinion, his arguments rather lead to the SMC than the ACM: The only step that Masters (1982) explains in detail is the last one between category $K-1$ and K. For this step, the SMC and the ACM are identical because

$$(Y_n \geq K) = (Y_n = K) \quad \text{and} \quad (Y_n \geq K-1) = (Y_n \in \{K-1, K\}). \tag{21}$$

Generally modeling the event $Y_n = k|Y \in \{k-1, k\}$ (instead of $Y_n \geq k|Y_n \geq k-1$) not only excludes all lower categories 0 to $k-2$, but also all higher categories $k+1$ to K. When thinking of a sequential process, however, the latter categories should still be achievable after the step to category k was successful. In his argumentation, Masters (1982) explains the last step *first* and then refers to the other steps as similar to the last step, thus concealing (deliberately or not) that the PCM is not in full agreement with the sequential process he describes.

Andrich (1978a) and Andrich (2005) presented yet another derivation of the PCM. When two dichotomous processes are independent, four results can occur: $(0, 0), (1, 0), (0, 1), (1, 1)$. Using the Rasch model for each of the two processes, the probability of the combined outcome is given by the *Polytomous Rasch Model* (PRM) (Andersen, 1973; Wilson, 1992; Wilson & Adams, 1993). When thinking of these processes as steps between ordered categories, $(0, 0)$ corresponds to $Y_n = 0$, $(1, 0)$ corresponds to $Y_n = 1$, and $(1, 1)$ corresponds to $Y_n = 2$. The event $(0, 1)$, however, is impossible because the second step cannot be successful when the first step was not. For an arbitrary number of ordered categories, Andrich (1978a) proved that the PRM becomes the PCM when considering the set of possible events only. While this finding is definitely interesting, it contains no argument that ordinal data observed in scientific experiments may be actually distributed according to the PCM.

Similar to the SM, the threshold parameters τ_k are not necessarily ordered in the ACM, that is the threshold of a higher category may be smaller than the threshold of a lower category. Andrich (1978a) and Andrich (2005) concluded that this happens when the categories themselves are disordered so that, for instance, category 3 was in fact easier to achieve than category 2. In a detailed logical and mathematical analysis, (Adams et al., 2012) proved the view of Andrich to be *incorrect*. Instead, this phenomenon is simply a property of the ACM that has no implication on the ordering of the categories.

Despite our criticism, we do not argue that the ACM is worse than the other models. It may not have a satisfying theoretical derivation, but has good mathematical properties especially in the case of PCM. In addition, the same relaxations to the regression and threshold parameters β and τ can be applied and they remain interpretable in the same way as for the other models, thus making the ACM a valid alternative to the CM and SM.

3.4 Generalizations of ordinal models

An important extention of the ordinal model classes described above is achieved by incorporating a multiplicative effect $\alpha_n > 0$ to the terms within the response function F. In the cumulative model, for instance, this results in the following model:

$$\Pr(Y_n = k | \eta_n, \alpha_n) = F(\alpha_n(\tau_{k+1} - \eta_n)) - F(\alpha_n((\tau_k - \eta_n)) \tag{22}$$

Such an parameter influences the slope of the response function, which may vary across observations (hence the index n). The higher α_n, the steeper the function. It is used in item response theory (IRT) to generalize the 2-Parameter-Logistic (2PL) Model to ordinal data, while the standard ordinal models are only generalizations of the 1PL or Rasch model (Rasch, 1961). In this context, we call α_n the *discrimination* parameter. Similarily, we can use α_n (or more precisely its inverse) in signal detection theory to model unequal variances for the noise and signal distributions. To make sure α_n ends up being positive, we often specify its linear predictor η_{α_n} on the log-scale so that

$$\alpha_n = \exp(\eta_{\alpha_n}) > 0. \tag{23}$$

We will learn more about it in the next section using hands on examples.

4 Fitting ordinal models in R

Although there are a number of software packages in the R statistical programming environment (R Core Team, 2017) that allow modelling ordinal responses, here we will use the *brms* package (Bürkner, 2017b, 2017a) for several reasons. First, it can estimate all three ordinal model classes introduced above in combination with multilevel structures, category specific effects (except for the cumulativel model), and predictors on distributional parameters (e.g., the discrimination α_n). To our knowledge, no other R package to date includes these features. Second, it is fully Bayesian, which provides considerably more information about the model and its parameters (Gelman et al., 2013; McElreath, 2016), allows more natural quantification of uncertainty (Kruschke, 2014), and is able to estimate models for which more traditional maximum likelihood based methods fail (Eager & Roy, 2017). For a general introduction to brms see Bürkner (2017b) and Bürkner (2017a).

In the tutorial below, we assume that readers know how to load data sets into R, and execute other basic commands. Readers unfamiliar with R may consult free online R tutorials[2]. The complete R code for this tutorial, including the example data used here, can be found at (https://osf.io/cu8jv/). To follow the tutorial, users first need to install the required brms R package. Packages should only be installed once, and therefore the following code snippet should be run only once:

```
install.packages("brms")
```

Then, in order to have the brms functions available in the current R session, users must load the package at the beginning of every session:

```
library(brms)
```

4.1 Opinion about funding stem cell research

We start with the first data set, with which we will investigate the relationship between the opinion about funding stem cell research (variable `rating`) and the fundamentalism / liberalism of one's religious beliefs (`belief`), stratified by gender (`gender`). In other words, we wish to predict `rating` from `belief` and `gender`. It is reasonable to assume that the stem cell opinion ratings result from categorization of a latent continuous variable–the opinion about stem cell research. Therefore, the application of the cumulative model is theoretically motivated and justified. This model can easily be fitted using the `brm()` function:

[2] A brief introduction to R basics can be found at http://blog.efpsa.org/2016/12/05/introduction-to-data-analysis-using-r/ (Vuorre, 2016). For a comprehensive, book-length tutorial, we recommend https://r4ds.had.co.nz (Wickham & Grolemund, 2016).

```r
fit_sc1 <- brm(
  rating ~ 1 + gender + belief,
  data = stemcell, family = cumulative()
)
```

In the above code snippet, we specified the model with the standard R modeling syntax, where dependent variables are written on the left-hand side of `~` and the predictors on the right-hand side, separated with `+`s. In addition, we provided the `data` and the `family` arguments. The former takes a data frame from the current R environment. The latter is commonly used in many R model fitting functions for defining the distribution of the response variable. Inside the parenthesis in `cumulative()`, we may specify the link function; omitting it leads to the default logit-link function.

The model (which we saved into the `fit_sc1` variable) is readily summarized via `summary(fit_sc1)`. See Table 4 for a summary of regression coefficients. The `Estimate` column provides the posterior mean of the parameters, while `2.5%ile` and `97.5%ile` provide the bounds of the 95% credible intervals (i.e., Bayesian confidence intervals). To get different CIs, use the `prob` argument (e.g. `summary(fit_sc1, prob = .99)` for a 99% CI.) Because we did not tell R otherwise, it used dummy coding for `belief` and chose `fundamentalist` as the reference category. Accordingly, the coefficients `beliefmoderate` and `beliefliberal` indicate how the ratings of moderate and liberal people differ from those with fundamentalist beliefs. We see that the corresponding estimates are negative and that the CIs do not include zero. Thus, we can conclude with at least 95% probability that moderate and liberal people prefer lower response categories and thus hold more positive opinon regarding the funding of stem cell research (remember that "definitely fund" was coded as 1 and "definitely not fund" as 4). More specifically, the model predicts that – on the latent scale – individuals with liberal beliefs hold -0.98 units more positive opinions on stem cell funding than do individuals with fundamentalist beliefs.

We may also summarize the results visually by plotting the estimated marginal

Table 4

Summary of regression coefficients for the cumulative model fitted to the stemcell data.

	Estimate	2.5%ile	97.5%ile
Intercept[1]	-1.38	-1.65	-1.12
Intercept[2]	0.61	0.35	0.87
Intercept[3]	1.71	1.41	2.02
gendermale	-0.04	-0.31	0.21
beliefmoderate	-0.42	-0.74	-0.10
beliefliberal	-0.98	-1.30	-0.66

relationship between `belief` and `rating`. On the left-hand side of Figure 2, we see the mean rating varying with religous belief and it is quite clear that fundamentalists have stronger opinion *against* funding stem cell research. However, this plot has the drawback of assuming equidistant response categories. Thus, on the right-hand side of Figure 2, we additionally see the predicted probabilities of every response category, separately.

```
marginal_effects(fit_sc1, "belief")
marginal_effects(fit_sc1, "belief", ordinal = TRUE)
```

Next, we want to investigate whether `belief` has category specific effects. That is, we ask if `belief`'s effect on funding opinion varies across response categories. To achieve this in brms, we can simply wrap predictors in `cs()`. Further, as descibred in Section 3, fitting category specific effects in cumulative models is problematic, so we use an adjacent category model instead. To specify an adjacent category model, we use `family = acat()` instead of `family = cumulative()`, as an argument to the `brm()` function:

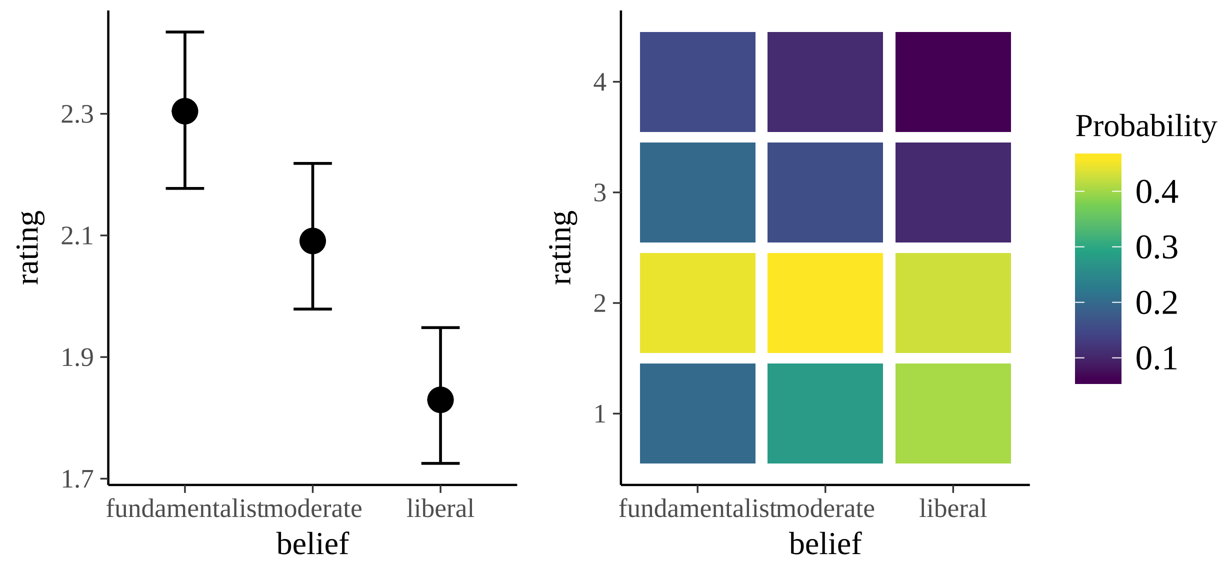

Figure 2. Marginal effects of religious belief on opinion about funding stem cell research based on model `fit_sc1`.

```
fit_sc2 <- brm(
  rating ~ 1 + gender + cs(belief),
  data = stemcell, family = acat()
)
```

As shown in Table 5, liberals and moderates tend to use lower response categories than fundementalists, but the strength of this effect varies substantially between categories. In particular, the difference between liberals and fundamentalists is strong (and, with 95% credibility, plausibly nonzero) for the transition between the first two categories (b = -0.85, 95%-CI = [-1.26, -0.45]) and to some extent also between the second and the third category (b = -0.51, 95%-CI = [-1.02, -0.01]). It can be difficult to interpret the size of these coefficients directly, because they are on the logit-scale within an adjacent category model. Thus, to obtain a better understanding of the magnitude of the effects, we recommend plotting the model's predicted values, for instance with `marginal_effects()`.

It remains unclear, however, whether category specific effects actually improve model

Table 5

Summary of regression coefficients for the category-specifc adjacent category model fitted to the stemcell data.

	Estimate	2.5%ile	97.5%ile
Intercept[1]	-0.78	-1.11	-0.47
Intercept[2]	0.79	0.47	1.12
Intercept[3]	0.30	-0.10	0.73
gendermale	-0.01	-0.16	0.13
beliefmoderate[1]	-0.13	-0.56	0.28
beliefmoderate[2]	-0.41	-0.88	0.07
beliefmoderate[3]	-0.19	-0.82	0.47
beliefliberal[1]	-0.85	-1.26	-0.45
beliefliberal[2]	-0.51	-1.02	-0.01
beliefliberal[3]	0.00	-0.69	0.67

fit. One approach to assess the latter is approximate leave-one-out cross-validation (LOO; (Vehtari, Gelman, & Gabry, 2017)), which provides a score that can be interpreted as typical information criteria such as AIC (Akaike, 1998) or WAIC (Watanabe, 2010)[3] in the sense that smaller values indicate better fit. To make sure differences between `fit_sc1` and `fit_sc2` are not simply the result of using another ordinal family, we also fit the adjacent category model without category specific effects.

```
fit_sc3 <- brm(
  rating ~ 1 + gender + belief,
  data = stemcell, family = acat()
)
```

[3] Actually AIC and WAIC can be interpreted as approximations of LOO.

The comparison between the three ordinal models using approximate leave-one-out cross-validation is done via

```
LOO(fit_sc1, fit_sc2, fit_sc3)
```

Table 6

LOO differences between the three ordinal models fitted to the stemcell data.

	LOOIC	SE
fit_sc1 - fit_sc2	-4.07	4.01
fit_sc1 - fit_sc3	-4.95	2.72
fit_sc2 - fit_sc3	-0.88	5.98

As can be seen in Table 6, the cumulative model (`fit_sc1`) has a somewhat better fit (smaller `LOOIC` value) than either adjacent category model, although the differences are not very large (up to 1 or 2 times the corresponding standard error). More importantly, both adjacent category models show very similar `LOOIC` values, which implies that estimating category specific effects does not improve model fit in a relevant manner, at least not when using leave-one-out cross-validation as the criterion. In the context of model selection, we may interpret a LOO difference greater than twice its corresponding standard error as suggesting that the model with a lower LOO value fits the data meaningfully better, at least when the number of observations is large enough[4]. Therefore, if forced to choose, we would prefer `fit_sc1` based on Table 6. However, we remind readers that model selection–based on any metric, be it a p-value, Bayes factor, or information criterion–is a controversial topic, and therefore suggest replacing hard cutoff values with context-dependent reasoning. For the

[4] LOO values and their differences are approximately normally distributed. Hence, for models based on enough observations, we may construct a frequentist confidence interval around the estimate. For instance, a 95%-CI around ΔLOO can be constructed via $[\Delta\text{LOO} - 1.96 \times \text{SE}(\Delta\text{LOO}), \Delta\text{LOO} + 1.96 \times \text{SE}(\Delta\text{LOO})]$.

current example, we favor the CM not only because of its best fit (as indicated by smallest LOO), but also because it is parsimonious and theoretically best justified.

In the above example, we only had data for one item per person. However, in many studies the participants provide responses to multiple items. For such data with multiple items per person, we can fit a multilevel ordinal model that takes the items and participants into account. This allows incorporating all information in the data into the model, while controlling for dependencies between ratings from the same person and between ratings of the same item. For this purpose, the data needs to be in long format, such that each row is an individual rating, with columns for the value of the rating, and identifiers for the participants and items. Suppose that we had measured opinion about funding stem cell research with multiple items and that we call the identifier columns `person` and `item`, respectively. Then, we could write the model formula as follows:

```
rating ~ 1 + gender + belief + (1|person) + (1|item)
```

The notation `(1|group)` implies that the intercept (1) varies over the levels of the grouping factor (`group`). In ordinal models, we have multiple intercepts (recall that they are called thresholds in ordinal models), and `(1|group)` allows these thresholds to vary by the same amount across levels of `group`. To model threshold-specific variances, we would write `(cs(1) | group)`. For instance, if we wanted all thresholds to vary differently across items so that each item receives its own set of thresholds, we could have added `(cs(1) | item)` to the model formula.

In summary, this example illustrated the use of CM and ACM (with and without category-specific effects) in the context of a Likert item response variable. We illustrated how to fit these three models to data using concise R syntax, enabled by the `brm()` function, and how to print, interpret, and visualize the model's estimated parameters. Paired with effective visualization (Figure 2), the models' results are readily interpretable and rich in information due to fully Bayesian estimation. We also found that, in this example, category-specific effects did not meaningfully improve model fit, and that the CM proved a better fit than either ACM.

4.2 Signal detection theoretic model of confidence ratings

In Section 2.2, we introduced the confidence rating data from a recognition memory experiment (Koen et al., 2013). Although software exists for modeling these type of data in a signal detection theoretic (SDT) framework (see Koen, Barrett, Harlow, & Yonelinas, 2017 for a MATLAB package), it is useful to recognize that the commonly used SDT models are equivalent to the cumulative model (CM) described above. Among other benefits, a regression framework makes it easy to include predictors and hierarchical structures for modeling multiple conditions and participants simultaneously. Although estimating the model with the brms R package is as straightforward as with the stem cell opinion data above, we take some space here to introduce the SDT models to highlight their similarity to the CM (DeCarlo, 2010).

In the context of word recognition memory, SDT assumes that when a word is presented, participants have some degree of familiarity with it, and that this familiarity may differ as a function of whether the presented item is new or old (Macmillan & Creelman, 2005). If familiarity is relatively weak, participants respond "new" (in binary new/old response tasks), or give a relatively low confidence rating that the word is old (in confidence rating tasks, such as the one discussed here). The participants' confidence ratings categorize the latent familiarity variable: In binary new/old tasks, there is a single threshold τ (commonly called a *criterion*, c) and if familiarity on a trial exceeds it, participants respond "old", otherwise they respond "new". In rating tasks, there are multiple thresholds $\tau = (\tau_1, ..., \tau_K)$, which divide the internal familiarity distribution to K + 1 confidence rating categories. Importantly, the SDT model includes an additional parameter for memory ability, which is commonly called *d'*. This parameter measures the extent to which old items elicit greater familiarity than do new items, and can be included in ordinal regression by adding item type (new/old) as a predictor.

The unobserved familiarity variable is commonly assumed to be normally distributed with a standard deviation of 1, in which the case model is (4) with a normal response distribution (10; i.e. $F = \Phi$). This model is known as the Gaussian equal variance SDT model (EVSDT). Importantly, the EVSDT assumptions can be changed, leading to different SDT models: For example, we could assume a logistic or extreme value distribution for the familiarity variable (DeCarlo, 1998, 2010). We could also allow the new and old item familiarity distributions to have different variances, leading to the unequal variance SDT model (UVSDT). A robust finding in the literature is that the old-item variance is greater than the new-item variance (Koen et al., 2013; Ratcliff, Sheu, & Gronlund, 1992), suggesting that the UVSDT model is particularly useful.

It is important to note that these variants of the SDT model are equivalent to various versions of the CM discussed above. In fact, the UVSDT model is also known as an ordered probit model with heteroscedastic error (DeCarlo, 2010). Below, we fit the EVSDT and UVSDT models as ordinal regression models using brms. An important benefit from using a regression modeling framework for fitting the SDT models is that it is easy to fit the model simultaneously to multiple participants' data by using a multilevel (also known as hierarchical or mixed effects) model. Multilevel modeling is an increasingly popular strategy for analyzing data with repeated measures and within-participant manipulations. However, an in-depth discussion of multilevel models is outside the scope of this tutorial, so we refer readers to textbooks on the topic (Gelman & Hill, 2007; McElreath, 2016). The benefits of multilevel modeling in the context of SDT are discussed in (Rouder & Lu, 2005; Rouder et al., 2007).

To fit the ordinal regression model, the data must be formatted with each observation (trial) in it's own row (i.e. the data must be in the long format). The current data comprises three columns, one that uniquely identifies each participant (`id`), a factor for the item type (`item`; new vs. old), and the 1-6 confidence rating (`rating`). We then fit the EVSDT model with these data using the R package brms (Bürkner, 2017b, 2017a). The syntax is identical to the CM fitted to the stem cell data in our previous example, with two important changes: Because we fit a multilevel model, we specify the model's parameters (thresholds and the effect of item type) as varying between participants by using brms' group-specific effect syntax. Second, we use a probit link function, instead of the default logit link function, because the familiarity distribution is commonly assumed to be Gaussian (brms allows examining other distributional forms, if desired). Additionally, we adjust the sampling parameters by increasing the number of iterations. The syntax for this model is as follows:

```
fit_evsdt <- brm(
  rating ~ 1 + item + (1 + item | id),
  data = sdt, family = cumulative("probit"),
  iter = 3000
)
```

In the above code snippet, the second line specified a regression model that predicts `rating` from a population-level intercept (`1`) and effect of `item`[5]. R includes intercepts in regression models by default, but they can be explicitly represented by adding the term `1`, as we did here. Recall that in the context of ordinal models, we do not refer to a single intercept, but instead to a vector of thresholds. Second, we used `(1 + item | id)` to specify that the thresholds and effects of `item` should vary between participants (`id`)[6]. See `?brmsformula` for more information on brms' multilevel modeling syntax.

We then focus on the model's estimated population-level parameters, which are summarized in the right panel of Table 7. To print the summary of the estimated model, use `summary(fit_evsdt)`. First, the five intercepts summarize the posterior distributions of the thresholds. These indicate, in standard normal deviates (i.e. z-scores), the five thresholds

[5] Population-level effects are also often known as "fixed" effects in the frequentist literature.

[6] Varying effects are sometimes known as "random" effects in the frequentist literature.

between the six confidence rating categories. As population-level effects, they can be interpreted as thresholds for the average person. For example, the second threshold describes the *z*-score of the probability of responding with a confidence rating of 2 or lower, when a new item was presented. The effect of old items—the memory ability parameter *d'*—is given on the last line (`itemold`): The average increase in familiarity for old vs. new items (*d'*) was 1.38 (95%-CI [1.25, 1.50]). The model's estimated marginal confidence ratings are shown in Figure 3.

Table 7

Summary of estimated ordinal models of memory recognition data

	UVSDT			EVSDT		
	Estimate	2.5%ile	97.5%ile	Estimate	2.5%ile	97.5%ile
Intercept[1]	-0.59	-0.71	-0.49	-0.49	-0.59	-0.39
Intercept[2]	0.20	0.09	0.31	0.21	0.12	0.31
Intercept[3]	0.70	0.59	0.81	0.65	0.55	0.75
Intercept[4]	1.04	0.94	1.15	0.94	0.84	1.04
Intercept[5]	1.50	1.38	1.61	1.31	1.21	1.42
itemold	1.89	1.62	2.18	1.38	1.25	1.50
disc_old	-0.39	-0.55	-0.24			

Next, we fit the UVSDT model with brms. This model is similar to EVSDT, but has one additional parameter to allow a different standard deviation for the old-item familiarity distribution. In brms, the SD parameter is called *disc* (short for *discrimination*), following conventions in item response theory. Predicting auxiliary parameters in brms is accomplished by passing multiple regression formulas to the `brm()` function, by first wrapping these formulas in another function, `bf()`. Because the SD parameter is by definition 1 for the baseline (new item) distribution, we must ensure that *disc* is only

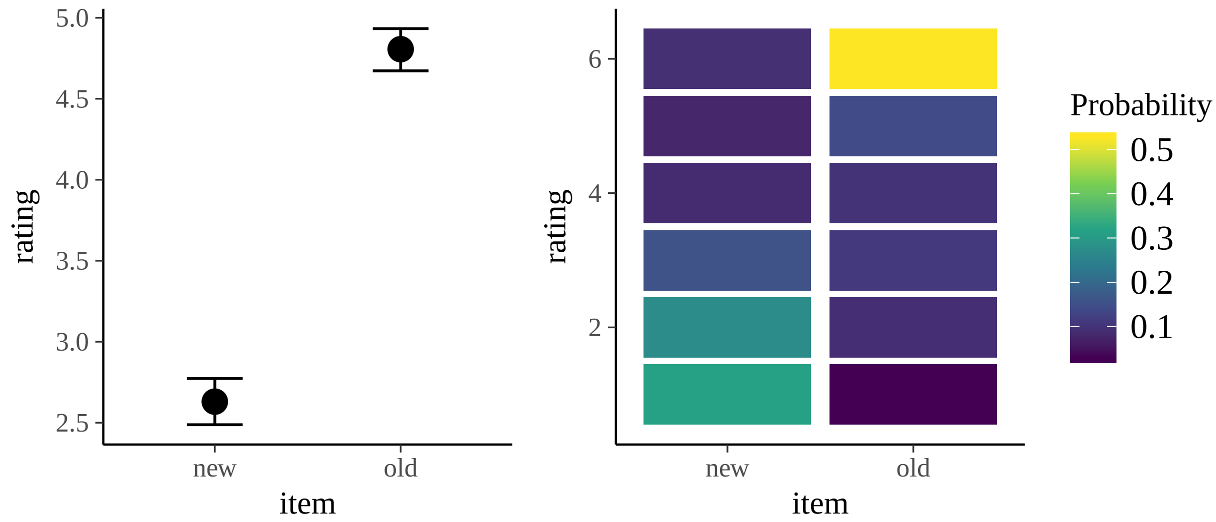

Figure 3. Marginal effects of item type on confidence ratings based on the EVSDT model.

estimated for the old items. To do so, we omit the intercept from the model by writing `0 + ...` on the right-hand side of the regression formula and add a contrast variable `old`, which is coded as `'new' = 0` and `'old' = 1`[7]. Further, the *disc* parameter is modeled on the log-scale by default, because it must be strictly positive. With this in mind, the UVSDT model is specified as EVSDT above, but with an additional formula for *disc*.

```
fit_uvsdt <- brm(
  bf(rating ~ 1 + item + (1 + item |i| id),
     disc ~ 0 + old + (0 + old |i| id)),
  data = sdt, family = cumulative("probit"),
  iter = 3000, inits = 0
)
```

There is an additional change to the group-specific effect syntax in the UVSDT model: We modeled the participant-specific effects on `rating` and `disc` as correlated across participants (see Rouder et al. (2007)). We did this by specifying `|i|` in the varying effects

[7] Instead of using the factor variable `item`, R requires using a numerical indicator variable to allow dropping the model's intercept without causing automatic cell-mean coding.

formulas passed to `bf()` (`i` is arbitrary, but because it is the same symbol across the formulas, brms will model these in a joint covariance matrix).

The UVSDT model's estimated population-level parameters are summarized in the left panel of Table 7. The main change across the two models is that the UVSDT model reports a plausibly nonzero *disc* parameter (appended with `_old` to indicate *disc* for items where `old`=1). However, *disc* reports ln(-σ_{old}) (see equation 23), so to convert it to a standard deviation we take its negative and exponentiate, leading to an estimated $\sigma_{old} = \exp(-disc_{old})$ = 1.48 (95%-CI = [1.27, 1.73])[8]. By exponentiating *disc* we find the ratio of the noise to signal distribution SD as 0.68 (95%-CI [0.58, 0.79]).

We also briefly highlight an additional benefit of estimating the SDT models of confidence rating data in a multilevel ordinal regression framework, as presented here. Researchers interested in comparing models' fits to participants' data sometimes compute fit metrics from models that are independently fit to each participant's data. Then, in a second step, these metrics are compared across models using descriptive or inferential statistics calculated from the participant-specific fit metrics. This approach may be suboptimal, especially when the data is not balanced across participants, because it ignores the uncertainty in the participant-specific fit metrics. However, model comparison across two multilevel models appropriately accounts for participant-level uncertainty, and provides a single metric for each model. Therefore, we investigate whether allowing for a different old-item variance improves model fit by comparing the multilevel EVSDT and UVSDT models using LOO. Confirming previous findings, the UVSDT had a smaller LOO value (LOO difference = 728.44, SE = 55.20), indicating a decisively better fit to these data.

Our discussion of the confidence rating data was more involved than the stem cell example above because we wished to higlight the connection between SDT models and the more general ordinal regression framework. The key point was that we obtained all common

[8] Because each parameter is estimated by a sample of random draws from its posterior distribution, it is straightforward to obtain other SDT metrics with their associated uncertainty estimates.

SDT metrics (with their uncertainty estimates), such as the memory ability parameter *d'*, the thresholds, and the new-old item variance ratio, from the CM. Viewing this common cognitive model in a regression modeling framework is useful because it allows easily adding further predictor variables to the model. For instance, if we had two different groups of participants, we could easily model how group membership affects the model's parameters by including the variable in the regression equation (e.g. `rating ~ item * group` would estimate the difference in *d'* between groups). Alternatively, a within-subject manipulation with repeated measures could also be easily included (e.g. `rating ~ item * condition + (item * condition | id)`). Furthermore, because multilevel models partially pool information across participants, there is no need for correcting for hit / false alarm rates of zero or one–a common nuisance in SDT modeling. We hope that the example presented here motivates a more widespread use of (multilevel) ordinal regression methods wherever applicable, including confidence rating data.

4.3 Years until divorce

In the third example, we are predicting years until divorce (`ma_years`) of the first marriage using three couple related variables, namely womens' age at marriage (`age_at_ma`), whether couples were already living together before marriage (`liv_together`), and whether the husband was married before (`hu_ma_before`). We can think of the years of marriage as a sequential process: Each year, the marriage may continue or end by divorce, but the latter can only happen if it did not happen before. This clearly calls for use of the sequential model and since we seek to predict the time until divorce (i.e., the time until marriage *stops*) we will use the stopping formulation specified in (14). In a first step, we will only consider actually divorced couples. Further, we assume an extreme-value response function (corresponding to the *cloglog* link), as it is the most common choice in discrete time-to-event / survival models. The model is readily set-up via

```r
prior_ma <- prior(normal(0, 5), class = "b") +
  prior(normal(0, 5), class = "Intercept")
fit_ma1 <- brm(
  ma_years ~ 1 + age_at_ma + liv_together + hus_ma_before,
  data = subset(marriage, divorced), family = sratio("cloglog"),
  prior = prior_ma, inits = 0
)
```

We used weakly informative `normal(0, 5)` priors[9] for all regression coefficients to improve model convergence, and to illustrate how to specify prior distributions with brms. After fitting this model, we then print a summary of the results with `summary(fit_ma1)`. As depicted in Table 8 (we omitted the thresholds from this table for clarity), women who marry later appear to have shorter marriages. The other predictors, on the other hand, show little relationship with marriage duration.

However, this model omits an important detail in the data: We only included couples who actually got divorced, and excluded couples who were still married at the end of the study. In the context of time-to-event analysis, we call this (right) censoring, because divorce did not happen up to the point of the end of the study, but may well happen later on in time. Both excluding this information altogether (as we did in the analysis above) or falsely treating these couples as having divorced right at the end of the study may lead to bias in the results of unknown direction and magnitude.

For these reasons, we must find a way to incorporate censored data into the model. In the standard version of the sequential model explained in Section 3, each observation must have an associated outcome category. However, for censored data, the outcome category was unobserved. Hence, we will need to expand the standard sequential model, which requires a

[9] This prior is weakly informative for the present model and variable scales. Be aware that for other models or other variable scales, such a prior may very well be informative.

Table 8

Summary of regression coefficients for the sequential model fitted to the marriage data.

	Estimate	2.5%ile	97.5%ile
age_at_ma	-0.04	-0.06	-0.02
liv_togetheryes	0.01	-0.15	0.17
hus_ma_beforeyes	-0.03	-0.26	0.22

little bit of extra work, to which we now turn.

In the field of time-to-event analysis, the so called *hazard rate* plays a crucial role (Cox, 1992). For discrete time-to-event data, the hazard rate $h(t)$ at time t is simply the probability that the event occurs at time t given that the event did not occur until time $t-1$. In our notation, the hazard rate of observation n at time t can be written as

$$h_n(t) = F(\tau_t - \eta_n) \tag{24}$$

Comparing this with equation (14), we see that the stopping sequential model is just the product of $h_n(t)$ and $1 - h_n(t)$ terms for varying values of t. Each of these terms defines the event probabilty of a bernoulli variable (0: still married beyond time t; 1: divorce at time t) and so the sequential model can be understood as a sequence of conditionally independent bernoulli trials. Accordingly, we can equivalently write the sequential model in terms of binary regression[10] by expanding each the outcome variable into a sequence of 0s and 1s[11].

[10] Binary regression might be better known as *logistic* regression, but since we do not apply the *logit* link in this example, we prefer the former term.

[11] This is generally possible, not just in the present example. That is, if desired, ordinal sequential models can be expressed as generalized liner models (GLMs) and thus fitted with ordinary GLM software. However, this is often much less convenient than directly using the ordinal sequential model, because the data has to be expanded in the above described way. We only recommend using the GLM formulation, if the standard formulation is not applicable, for instance when dealing with censored data.

More precisely, for each couple, we create a single row for each year of marriage with the outcome variable being 1 if divorce happend in this year and 0 otherwise. The expanded data is examplified in Table 9.

Table 9

Marriage data from the NSFG 2013-2015 survey expanded for use in binary regression.

ID	age_at_ma	hus_ma_before	liv_together	divorced	discrete_time
1	19	no	yes	0	1
1	19	no	yes	0	2
1	19	no	yes	0	3
1	19	no	yes	0	4
1	19	no	yes	0	5
1	19	no	yes	0	6
1	19	no	yes	0	7
1	19	no	yes	0	8
1	19	no	yes	1	9
2	22	no	yes	0	1
2	22	no	yes	0	2
2	22	no	yes	0	3
2	22	no	yes	0	4
2	22	no	yes	0	5
2	22	no	yes	0	6
2	22	no	yes	0	7
2	22	no	yes	0	8
2	22	no	yes	0	9

In the expanded data set, `discrete_time` is treated as a factor so that, when included in a model formula, its coefficients will represent the threshold parameters. This can be done in at least two ways. First, we could write `... ~ 0 + discrete_time + ...`, in which case the coefficients can immediately interpreted as thresholds. Second, we could write `... ~ discrete_time + ...` so that the intercept is the first threshold, while the $K - 1$ coefficients of `discrete_time` represent differences between the respective other thresholds and the first threshold (dummy coding). Note that these representations are equivalent in the sense that we can transform one into the other. However, the second option usually leads to improved sampling, because it allows brms to do some internal optimization. We are now ready to fit a binary regression model to the expanded data set.

```
fit_ma2 <- brm(
  divorce ~ 1 + discrete_time + age_at_ma + liv_together + hus_ma_before,
  data = marriage_long, family = bernoulli("cloglog"),
  prior = prior_ma, inits = 0
)
```

Estimated coefficients are summarized in Table 10. Again, we did not include the threshold estimates in order to keep the table readable. Marginal model predictions are visualized in Figure 4. When interpreting results of the second model, we have to keep in mind that we predicted the probabilty of divorce and not the time of marriage as in the first model. Accordingly, if including the censored data did not change something drastically, we would expect signs of the regression coefficients to be inverted in the second model as compared to the first model. Interestingly, age at marriage (`age_at_ma`) has the same sign in both models, leading to opposite conclusions: While the first model predicted longer lasting marriages (lower probability of divorce) for women marrying at lower age, the opposite seems to be true for the second model (probability of divorce was lower for women marrying at older age). This is plausible insofar as censoring is confounded with age at marriage: Women marrying at older ages are more likely to still be married at the time of the survey. Moreover, in contrast to the first model, the second model reveals that couples living together before marriage have considerably lower probability of getting divorced. This underlines the importance of correctly including censored data in (discrete) time-to-event models. The present example has demonstrated how to achieve this in the framework of the ordinal sequential model.

Table 10

Summary of regression coefficients for the extended sequential model fitted to the marriage data.

	Estimate	2.5%ile	97.5%ile
age_at_ma	-0.06	-0.08	-0.04
liv_togetheryes	-0.31	-0.48	-0.15
hus_ma_beforeyes	0.00	-0.23	0.21

Lastly, we briefly discuss time-varying predictors in discrete time-to-event data. Since the survey took place at one time and asked questions retrospectively, we do not have reliable time-varying predictors for years of marriage, but we can easily think of some potential ones. For instance, we can imagine that the probability of divorce changes over the years with changes in the socio-economic status of the couple. Such time-varying predictors cannot be modeled in the standard sequential model, because all information of a single marriage process has to be stored within the same row in the data set. Fortunately, time-varying predictors can be easily added to the expanded data set shown in Table 9 and then treated in the same way as other predictors in the binary regression model.

5 Conclusion

In this tutorial, we introduced three important ordinal model classes, namely the cumulative, sequential, and adjacent category model both from a theoretical and an applied perspective. The models were formally derived from their underlying assumptions and

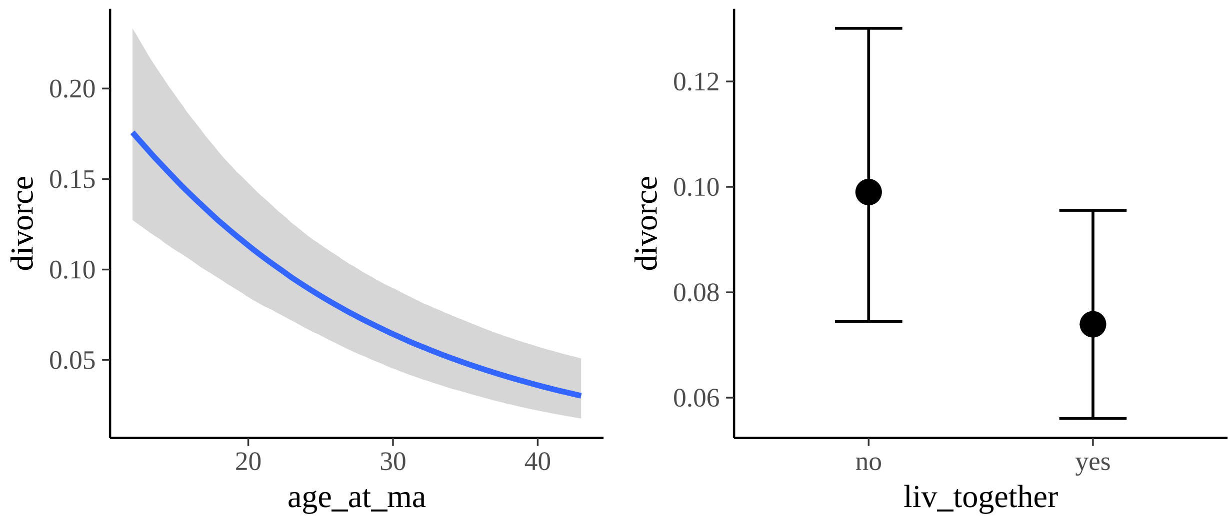

Figure 4. Marginal effects of *woman's age at marriage* and *living together before marriage* on the probability of divorce in the 7th year of marriage.

applied to three real-world data sets covering different psychological fields and research questions. We did not engage in demonstrating ourselves (e.g., via simulations) that using ordinal models for ordinal data is superior to other approaches such as linear regression, because we think this has already been sufficiently covered elsewhere (Liddell & Kruschke, 2017). Nevertheless, we briefly mention some further arguments in favor of ordinal models.

5.1 Why should researchers use ordinal regression?

Although we have highlighted the theoretical justification, and practical ease, of applying ordinal models to ordinal data, one might still object to using these models. We wish to point out here that some of these objections are not sound. First, one might oppose ordinal models on the basis that their results are more difficult to interpret and communicate than those of corresponding linear regressions. The main complexity of ordinal models, in contrast to linear regression, is in the threshold parameters. However, equivalent to intercept parameters in linear regression, these parameters rarely are the target of main inference. Furthermore, brms' helper functions make it easy to to calculate (see `?fitted.brmsfit`) and visualize (`?marginal_effects.brmsfit`) the model's fitted values (i.e. the predicted marginal proportions for each response category).

Second, it is sometimes the case that one's substantial conclusions do not strongly depend on whether an ordinal or a linear regression model was used. We wish to point out that even though the actionable conclusions may be similar, a linear model will have a lower predictive utility by virtue of assuming a theoretically incorrect outcome distribution. Perhaps more importantly, linear models for ordinal data can lead to effect size estimates that are distorted in size or certainty, and this problem is not solved by averaging multiple ordinal items (Liddell & Kruschke, 2017).

5.2 Choosing between ordinal models

Equipped with the knowledge about three ordinal model classes, researchers might wonder how to choose between them for a given data set containing ordinal data. From a theoretical perspective, we recommend the cumulative model, if the response can be understood as the categorization of a latent continuous construct, and the sequential model, if the response can be understood as being the result of a sequential process (sometimes both processes are reasonable at the same time). If category-specific effects are of interest, we recommend using the sequential or adjacent category model. If sampling speed matters, it might be wise to go with the cumulative model, because it is computationally much less intensive. Lastly, we want to point out that the most important step is using *any* ordinal model, instead of methods that falsely assume responses to be metric or nominal. Accordingly, if you apply ordinal models to ordinal data going forward, this tutorial was already a succes.

References

Adams, R. J., Wu, M. L., & Wilson, M. (2012). The Rasch rating model and the disordered threshold controversy. *Educational and Psychological Measurement*, *72*(4), 547–573. doi:10.1177/0013164411432166

Agresti, A. (1984). *Analysis of ordinal categorical data*. John Wiley & Sons.

Agresti, A. (2010). *Analysis of ordinal categorical data*. Chichester: John Wiley & Sons. doi:10.1002/9780470594001

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199–213). Springer.

Ananth, C. V., & Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology*, *26*(6), 1323–1333.

Andersen, E. B. (1973). CONDITIONAL inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*(1), 31–44. doi:10.1111/j.2044-8317.1973.tb00504.x

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, *42*(1), 69–81.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. doi:10.1007/BF02293814

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594.

Andrich, D. (2005). The Rasch model explained. In *Applied Rasch measurement: A book of exemplars* (pp. 27–59). Springer. doi:10.1007/1-4020-3076-2_3

Bürkner, P.-C. (2017a). Advanced bayesian multilevel modeling with the r package brms. *arXiv Preprint*, 1–15. Retrieved from https://arxiv.org/abs/1705.11123

Bürkner, P.-C. (2017b). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *34*, 187–220.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in statistics* (pp. 527–541). Springer.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. doi:10.1037/1082-989X.3.2.186

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*(3), 304–313. doi:10.1016/j.jmp.2010.01.001

Eager, C., & Roy, J. (2017). Mixed Effects Models are Sometimes Terrible. *arXiv*. Retrieved from http://arxiv.org/abs/1701.04858

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data*. MIT press.

Fienberg, S. E. (2007). *The analysis of cross-classified categorical data*. Springer Science & Business Media.

Fischer, G. H. (1995). The derivation of polytomous Rasch models. In *Rasch models* (pp. 293–305). Springer. doi:10.1007/978-1-4612-4230-7_16

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition*. Boca Raton: Chapman and Hall/CRC.

Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference. In *International encyclopedia of statistical science* (pp. 977–979). Springer.

Guisan, A., & Harrell, F. E. (2000). Ordinal response regression models in ecology. *Journal of Vegetation Science*, *11*(5), 617–626.

Koen, J. D., Aly, M., Wang, W.-C., & Yonelinas, A. P. (2013). Examining the causes of memory strength variability: Recollection, attention failure, or encoding variability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1726–1741. doi:10.1037/a0033671

Koen, J. D., Barrett, F. S., Harlow, I. M., & Yonelinas, A. P. (2017). The ROC Toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings. *Behavior Research Methods*, *49*(4), 1399–1406. doi:10.3758/s13428-016-0796-z

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial Introduction with R* (2nd Edition.). Burlington, MA: Academic Press.

Läärä, E., & Matthews, J. (1985). The equivalence of two models for ordinal data. *Biometrika*, *72*(1), 206–207.

Liddell, T., & Kruschke, J. K. (2017). Analyzing ordinal data with metric models: What could possibly go wrong? *Open Science Framework*. doi:10.17605/OSF.IO/9H3ET

Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory. New York*.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 109–142.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*(1), 91–100.

Molenaar, I. (1983). Item steps. *Heymans Bulletin HB-83-630-EX). Groningen: University of Groningen, Vakgroep Statistiek En Meettheorie.*

Peterson, B., & Harrell, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 205–217.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In

Proceedings of the fourth berkeley symposium on mathematical statistics and probability (Vol. 4, pp. 321–333). University of California Press Berkeley, CA.

Ratcliff, R., Sheu, C.-f., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*(3), 518–535. doi:10.1037/0033-295X.99.3.518

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. doi:10.3758/BF03196750

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal Detection Models with Random Participant and Item Effects. *Psychometrika*, *72*(4), 621–642. doi:10.1007/s11336-005-1350-6

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph Supplement*.

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, *60*(4), 549–572.

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*(2684), 677–680. doi:10.1126/science.103.2684.677

Teachman, J. (2011). Modeling repeatable events using discrete-time data: Predicting marital dissolution. *Journal of Marriage and Family*, *73*(3), 525–540.

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

Tutz, G. (1997). Sequential models for ordered responses. In *Handbook of modern item* *response theory* (pp. 139–152). Springer.

Tutz, G. (2000). *Die Analyse Kategorialer Daten: Anwendungsorientierte Einführung in Logit-Modellierung und Kategoriale Regression*. Oldenbourg: Oldenbourg Verlag.

Tutz, G., & Schmid, M. (2016). *Modeling discrete time-to-event data*. Springer.

Van Der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement*, *25*(3), 273–282. doi:10.1177/01466210122032073

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Verhelst, N. D., Glas, C., & De Vries, H. (1997). A steps model to analyze partial credit. In *Handbook of modern item response theory* (pp. 123–138). Springer.

Vuorre, M. (2016, December 5). *Introduction to Data Analysis using R. JEPS Bulletin*. Retrieved December 29, 2016, from http://blog.efpsa.org/2016/12/05/introduction-to-data-analysis-using-r/

Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167–179. doi:10.2307/2333860

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(Dec), 3571–3594.

Wickham, H., & Grolemund, G. (2016). *R for Data Science*. Retrieved from http://r4ds.had.co.nz/

Wilson, M. (1992). The ordered artition model: An extension of the partial credit model. *Applied Psychological Measurement*, *16*(4), 309–325. doi:10.1177/014662169201600401

Wilson, M., & Adams, R. J. (1993). Marginal maximum likelihood estimation for the ordered partition model. *Journal of Educational and Behavioral Statistics*, *18*(1), 69–90. doi:10.2307/1165183