Health Serv Outcomes Res Method (2016) 16:117–131
DOI 10.1007/s10742-016-0145-9

A mixed-effects location-scale model for ordinal questionnaire data

Donald Hedeker[1] · Robin J. Mermelstein[2] · Hakan Demirtas[2] · Michael L. Berbaum[2]

Received: 30 November 2015 / Revised: 25 March 2016 / Accepted: 1 April 2016 /
Published online: 11 April 2016

Abstract In health studies, questionnaire items are often scored on an ordinal scale, for example on a Likert scale. For such questionnaires, item response theory (IRT) models provide a useful approach for obtaining summary scores for subjects (i.e., the model's random subject effect) and characteristics of the items (e.g., item difficulty and discrimination). In this article, we describe a model that allows the items to additionally exhibit different within-subject variance, and also includes a subject-level random effect to the within-subject variance specification. This permits subjects to be characterized in terms of their mean level, or location, and their variability, or scale, and the model allows item difficulty and discrimination in terms of both random subject effects (location and scale). We illustrate application of this location-scale mixed model using data from the Social Subscale of the Drinking Motives Questionnaire assessed in an adolescent study. We show that the proposed model fits the data significantly better than simpler IRT models, and is able to identify items and subjects that are not well-fit by the simpler models. The proposed model has useful applications in many areas where questionnaires are often rated on an ordinal scale, and there is interest in characterizing subjects in terms of both their mean and variability.

Keywords Proportional odds model · Variance modeling · Extreme response styles · Scaling model · IRT

Supported by National Cancer Institute grant P01 CA98262 (Mermelstein, PI) and National Heart Lung and Blood Institute grant R01 HL121330 (Hedeker & Dunton). Thanks to Oksana Pugach for aiding in carrying out the simulation studies.

✉ Donald Hedeker
hedeker@uchicago.edu

[1] Department of Public Health Sciences, University of Chicago, 5841 S. Maryland Ave., Room W254, MC2000, Chicago, IL 60637, USA

[2] University of Illinois at Chicago, Chicago, IL, USA

1 Introduction

In health studies, ordinal outcomes are often used to represent, for example, severity of illness, pain levels, or symptomatology (e.g., none, mild, moderate, severe). The ordinal logistic regression model, described as the proportional odds model by McCullagh (1980), is a popular model for analyzing such ordinal outcomes. For multilevel data, where observations are nested within clusters (e.g., classes, schools, clinics) or are repeatedly assessed across time, mixed-effects regression models are often used to account for the dependency inherent in the data (Hedeker and Gibbons 2006), and several authors have developed logistic and/or probit mixed-effects models for ordinal data (Ezzet and Whitehead 1991; Agresti and Lang 1993; Hedeker and Gibbons 1994). Tutz and Hennevogl (1996) and Johnson (2003) extend these mixed ordinal regression models by additionally allowing the model thresholds to be treated as random effects.

Logistic models for ordinal outcomes typically include the proportional odds assumption (or its equivalent under the probit link function) for model covariates. For an ordinal response with C categories, this assumption requires that the effect of an explanatory variable is the same across the $C - 1$ cumulative logits of the model, or proportional across the cumulative odds. In the context of an ordinal mixed model, Hedeker and Mermelstein (1998) extended this to allow for non-proportional odds for all or a subset of the explanatory variables. A similar extension is described in Saei and McGilchrist (1998), who allow non-proportional time effects in panel studies. In this approach, explanatory variables are allowed to have varying effects on the $C - 1$ cumulative logits. Thus, for a particular explanatory variable, $C - 1$ regression coefficients are estimated. These additional parameters reflect different location effects of the explanatory variables.

A somewhat different extension of the proportional odds model is described by Tosteson and Begg (1988). Here, in the context of receiver operating characteristic (ROC) analysis, the *scale* of the effects of explanatory variables is allowed to vary. In other words, the underlying variance of the logistic distribution can vary as a function of model covariates. McCullagh and Nelder (1989) refer to this extended model for ordinal data as a generalized "rational" model. Toledano and Gatsonis (1996) use this extension in describing generalized estimating equations (GEE) analysis of correlated ROC data, while Ishwaran and Gatsonis (2000) build upon this approach using Bayesian methods.

For cross-sectional data, Cox (1995) brought together these extensions of the proportional odds model into what he termed location-scale cumulative odds models. Hedeker et al. (2006) further developed this approach within a mixed model framework for longitudinal ordinal data. The inclusion of scale parameters within the mixed model is particularly advantageous because it allows modeling of both the within-subjects (WS) and between-subjects (BS) variances. Extending this, Hedeker et al. (2009) described a mixed model for variance modeling of longitudinal ordinal data that also included a random subject effect to the WS variance model.

In this paper, we build upon the model of Hedeker et al. (2009) to develop a mixed-effects location-scale model for ordinal questionnaire data, and propose a model with four types of item parameters: difficulty, discrimination, scale, and scale discrimination. Additonally, a random subject effect is included in the WS variance specification to allow the WS variance to vary at the subject level. Thus, the model includes two random subject effects that represent a subject's location and scale. These random effects are allowed to be correlated. Data from an adolescent study are used to illustrate application of the proposed 4-parameter location-scale model for ordinal questionnaire data.

The article is organized as follows. Section 2 describes the mixed-effects proportional odds model as applied to questionnaire data, and the extentions of it that lead to the proposed 4-parameter location-scale model. Estimation aspects are described in Sect. 3. Section 4 presents the results of a small simulation study. The adolescent dataset and questionnaire used to illustrate the model is presented in Sect. 5, as are the results of analysis using the proposed model. Section 6 gives details of how differential item functioning (DIF) can be incorporated into the model, including an examination of DIF by gender. Finally, in Sect. 7, we discuss and summarize features of the model and our application.

2 Mixed-effects proportional odds model

Consider the ordinal response Y_{ij} of subject i ($i = 1, 2, \ldots, N$) on item j ($j = 1, 2, \ldots, n_i$). Although the given questionnaire has n items, in what follows we do not assume that all subjects respond to all n items, but rather to a subject-specific number n_i. We do assume that all of the items have C ordered response categories, though this could be generalized to allow for different numbers of categories for each item. An ordinal proportional odds random-intercept model for the response of subject i to item j in category c can be written as:

$$\log\left[\frac{P_{ijc}}{1 - P_{ijc}}\right] = \alpha_c - \left[\boldsymbol{x}_j'\boldsymbol{\beta} + \boldsymbol{x}_j'\boldsymbol{\delta}\theta_i\right]. \quad (1)$$

Here, $P_{ijc} = \Pr(Y_{ij} \leq c \mid \theta)$ represent the (conditional) cumulative probabilities for the C categories of Y, $\boldsymbol{x}_j$ is an $n \times 1$ vector of item indicators, $\boldsymbol{\beta}$ is an $n \times 1$ vector of "item difficulty" parameters, $\boldsymbol{\delta}$ is an $n \times 1$ vector of "item discrimination" parameters, and $\alpha_1 < \alpha_2 < \ldots < \alpha_{C-1}$ are strictly increasing thresholds or intercepts. Here, because of the inclusion of all n item difficulty ($\boldsymbol{\beta}$) parameters, one of the thresholds must be set to zero for identification; we arbitrarily set $\alpha_1 = 0$. The random subject effect θ_i is distributed in the population following a standard normal distribution (i.e., as $N(0, 1)$).

Though not parameterized in exactly the same way, the above model is akin to Samejima's model for graded response data (Samejima 1969). In particular, in IRT models, it is common for the item difficulty parameters ($\boldsymbol{\beta}$) to be centered around zero, with negative (positive) values indicating easier (harder) items. As written above, a more positive β would indicate an item that is endorsed in the higher ordinal response categories. Similarly, in IRT models it is common to center the discrimination parameters ($\boldsymbol{\delta}$) multiplicatively around 1, but in the model above we do not impose this parameterization. Nonetheless, a greater δ value would indicate an item that is related more strongly to the random subject effect θ, or is better able to discriminate subjects in terms of the random effect. The item discrimination parameters can also be thought of as factor loadings, which indicate the degree to which the items load on the random subject effect. The differences in the common parameterization of the mixed and IRT models are more fully described in Hedeker et al. (2006).

2.1 Item scaling

As described in Skrondal and Rabe-Hesketh (2004), scale parameters can be added to allow the dispersion across the response categories to vary for each of the items.

$$\log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \frac{\alpha_c - (\boldsymbol{x}_j'\boldsymbol{\beta} + \boldsymbol{x}_j'\boldsymbol{\delta}\theta_i)}{\sigma_{\epsilon_j}}, \tag{2}$$

where the submodel for scaling follows that of heteroscadastic regression models (Harvey 1976; Aitkin 1987) and is given as

$$\sigma^2_{\epsilon_j} = \exp(\mathbf{x}_j'\boldsymbol{\gamma}). \tag{3}$$

Here, $\boldsymbol{\gamma}$ are item scale parameters indicating dispersion across the ordered response categories. For identification, one item must be specified as the reference item, and so the item indicator vector $\mathbf{x}_j$ and the item scale parameter vector $\boldsymbol{\gamma}$ are both of size $n-1$. The exponential function ensures a positive multiplicative factor, and so the resulting variances are positive. The $\boldsymbol{\gamma}$ represent within-subjects (WS) or error variance parameters. Thus, another way to think of these parameters is that they allow the error variance to vary across items, and therefore represent the degree of relative item fit/misfit to the 2-parameter IRT model (i.e., the model represented by the numerator of the equation). Higher values of γ_j would indicate items that are relatively worse-fitting to the 2-parameter model.

2.2 Subject scaling and item discrimination of scale

As described in Hedeker et al. (2009), the WS variance can also vary across subjects by including random subject scale effects ω_i:

$$\sigma^2_{\epsilon_{ij}} = \exp(\mathbf{x}_j'\boldsymbol{\gamma} + \omega_i). \tag{4}$$

These random scale effects ω_i are normally distributed in the population of subjects with mean 0 and variance σ^2_ω and, as such, represent log-normal subject-specific perturbations of WS variance. Akin to the item scale parameters, one can think of these as representing the degree of *subject* fit/misfit to the 2-parameter IRT model, additionally controlling for item fit/misfit.

Similar to their counterparts θ_i, the random scale effects ω_i can be represented in standardized form, namely,

$$\sigma^2_{\epsilon_{ij}} = \exp(\mathbf{x}_j'\boldsymbol{\gamma} + \sigma_\omega\zeta_i), \tag{5}$$

and so, θ_i and ζ_i are each standard normal "location" and "scale" random effects indicating how a subject differs in terms of the mean and variance of his or her data, respectively. The correlation of these, denoted as $\rho_{\theta\zeta}$, indicates the association of location and scale in the population of subjects.

Finally, one can also include item discrimination parameters of scale $\boldsymbol{\tau}$, namely,

$$\sigma^2_{\epsilon_{ij}} = \exp(\mathbf{x}_j'\boldsymbol{\gamma} + \boldsymbol{x}_j'\boldsymbol{\tau}\,\zeta_i) \tag{6}$$

which indicate how well the items distinguish subjects of varying scale. These are akin to the item discrimination parameters $\boldsymbol{\delta}$ in that they can also be thought of as factor loadings. However, while the item discrimination parameters $\boldsymbol{\delta}$ indicate the degree to which the items load on the random subject location effect θ, the item discrimination parameters of scale $\boldsymbol{\tau}$ indicate the degree to which the items load on the random subject scale effect ζ. Taken together, this leads to the proposed 4-parameter location-scale model for ordinal data in terms of the logit λ_{ijc} (of subject i, item j, and category c):

$$\lambda_{ijc} = \log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \frac{\alpha_c - (\boldsymbol{x}_j'\boldsymbol{\beta} + \boldsymbol{x}_j'\boldsymbol{\delta}\theta_i)}{\exp(\mathbf{x}_j'\boldsymbol{\gamma} + \boldsymbol{x}_j'\boldsymbol{\tau}\,\zeta_i)} \tag{7}$$

with thresholds α_c; item parameters $\boldsymbol{\beta}$ (difficulty), $\boldsymbol{\delta}$ (discrimination), $\boldsymbol{\gamma}$ (scale), and $\boldsymbol{\tau}$ (scale discrimination); and subject parameters θ_i (location) and ζ_i (scale). Again, for identification, only $C-2$ thresholds and $n-1$ item scale parameters are estimable (and so the vector $\mathbf{x}_j$ is of size $n-1$, whereas the vector $\boldsymbol{x}_j$ for the other item parameters is of size n).

3 Estimation

Parameter estimation can be solved using maximum likelihood (ML). Let $\boldsymbol{Y}_i$ denote the vector of responses from subject i, and let the vector $\boldsymbol{\theta}_i$ represent the two random effects (i.e., θ_i and ζ_i). The probability of any response pattern $\boldsymbol{Y}_i$ (of size n_i), conditional on the random effects $\boldsymbol{\theta}$, is equal to the product of the probabilities of the item responses:

$$\ell(\boldsymbol{Y}_i \mid \boldsymbol{\theta}_i) = \prod_{j=1}^{n_i}\prod_{c=1}^{C} \Pr\,(Y_{ij} = c \mid \boldsymbol{\theta}_i), \tag{8}$$

where

$$\Pr\,(Y_{ij} = c \mid \boldsymbol{\theta}_i) = \Psi(\lambda_{ijc}) - \Psi(\lambda_{ij,c-1}), \tag{9}$$

and $\Psi(\cdot)$ represents the logistic cumulative distribution function (cdf). Here, we assume that a subject's responses are independent given the random effects (i.e., the conditional independence assumption). The marginal density of $\boldsymbol{Y}_i$ in the population is expressed as the following integral of the conditional likelihood $\ell(\cdot)$

$$h(\boldsymbol{Y}_i) = \int_{\boldsymbol{\theta}} \ell(\boldsymbol{Y}_i \mid \boldsymbol{\theta}_i)\, f(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{10}$$

where $f(\boldsymbol{\theta})$ represents the distribution of the random effects, namely a bivariate normal density. The marginal log-likelihood from the sample of N subjects is then obtained as $\log L = \sum_i^N \log h(\boldsymbol{Y}_i)$. Maximizing this log-likelihood yields ML estimates, which are sometimes referred to as *marginal* maximum likelihood estimates (Bock and Aitken 1981) because they are obtained by maximizing the marginal likelihood. For the analyses presented in this article, `SAS PROC NLMIXED`, which utilizes adaptive quadrature for integration of the random effects, was used and is detailed in the "Appendix". Additionally, empirical Bayes estimates of the random subject effects can be obtained from the procedure.

4 Simulation study

To evaluate the performance of the proposed model, a small simulation study was conducted. The specifications and parameter values for the simulation study were based on the example and data analysis presented in the next section. Specifically, 1000 datasets were simulated, each with 500 subjects and 4 item responses with 5 ordinal categories, according to the proposed 4-parameter model. For each dataset, the parameters of three models were estimated: a 2-parameter IRT model (with difficulty and discrimination), a

3-parameter model adding item scale, and the proposed 4-parameter model adding subject scale and item scale discrimination. The results for the 4-parameter model are presented in Table 1, and the results for the 2- and 3-parameter models are in Table 2. In these tables, bias is the difference between the average parameter estimate and the true value, coverage is the proportion of 1000 datasets in which the 95 % confidence interval included the true value, width is the average width of the 95 % confidence interval, and root mean squared error (rmse) for a given parameter η equals $\sqrt{E[(\hat{\eta}-\eta)^2]}$.

As can be seen from Table 1, the parameters of the proposed model are well-estimated with low bias, good coverage, and low rmse. The confidence interval widths are generally larger for the scale parameters than for the location parameters (difficulty and discrimination). Conversely, from Table 2, one can see that the results for the simpler 2-parameter model exhibit a great deal of positive bias, poor coverage, and higher rmse. Thus, if the data follow the proposed 4-parameter model and one uses the simpler 2-parameter model, the results can certainly be misleading. The 3-parameter model does considerably better, especially for the difficulty and discrimination parameters. However, the item scale parameters are poorly estimated with low coverage, relatively high rmse, and diminished confidence interval width. This finding agrees with results reported in Leckie et al. (2014) for a location-scale model of continuous outcomes, who also noted poor coverage and spurious precision for scale (regression parameter) estimates if the random scale effect was

Table 1 Simulation results for 4-parameter location-scale model

Parameter	True value	Estimate	Bias	Coverage	rmse	Width
β_1	2.800	2.837	0.037	0.957	0.180	0.689
β_2	1.800	1.816	0.016	0.945	0.145	0.565
β_3	2.000	2.023	0.023	0.950	0.143	0.547
β_4	1.600	1.611	0.011	0.947	0.155	0.604
δ_1	1.000	1.009	0.009	0.944	0.116	0.455
δ_2	1.900	1.913	0.013	0.957	0.151	0.605
δ_3	1.400	1.406	0.006	0.954	0.111	0.436
δ_4	2.200	2.233	0.033	0.949	0.194	0.747
γ_2	−0.700	−0.669	0.031	0.961	0.196	0.826
γ_3	−1.100	−1.086	0.014	0.954	0.191	0.797
γ_4	−0.200	−0.205	−0.005	0.958	0.200	0.799
τ_1	0.300	0.315	0.015	0.960	0.202	0.799
τ_2	1.000	0.907	−0.093	0.948	0.299	1.107
τ_3	1.100	1.060	−0.040	0.946	0.271	1.052
τ_4	0.900	0.821	−0.079	0.947	0.299	1.114
α_2	1.500	1.523	0.023	0.950	0.115	0.438
α_3	2.600	2.636	0.036	0.956	0.164	0.632
α_4	4.500	4.551	0.051	0.949	0.241	0.923
$\rho_{\theta\zeta}$	0.200	0.186	−0.014	0.961	0.111	0.445

β = item difficulty, δ = item discrimination, γ = item scale, τ = item scale discrimination, α = threshold, ρ = correlation

Table 2 Simulation results for 2- and 3-parameter models

Parameter	True value	Estimate	Bias	Coverage	rmse	Width
2-parameter model						
β_1	2.800	3.305	0.505	0.056	0.529	0.544
β_2	1.800	2.107	0.307	0.456	0.341	0.584
β_3	2.000	2.377	0.377	0.169	0.400	0.518
β_4	1.600	1.851	0.251	0.676	0.303	0.652
δ_1	1.000	1.165	0.165	0.690	0.205	0.437
δ_2	1.900	2.221	0.321	0.438	0.355	0.594
δ_3	1.400	1.586	0.186	0.666	0.219	0.465
δ_4	2.200	2.732	0.532	0.159	0.565	0.727
α_2	1.500	1.839	0.339	0.022	0.352	0.338
α_3	2.600	3.159	0.559	0.000	0.572	0.421
α_4	4.500	5.253	0.753	0.001	0.771	0.560
3-parameter model						
β_1	2.800	2.854	0.054	0.939	0.183	0.668
β_2	1.800	1.814	0.014	0.953	0.144	0.579
β_3	2.000	2.045	0.045	0.939	0.147	0.544
β_4	1.600	1.586	−0.014	0.946	0.156	0.626
δ_1	1.000	0.980	−0.020	0.937	0.115	0.444
δ_2	1.900	1.954	0.054	0.944	0.161	0.600
δ_3	1.400	1.430	0.030	0.948	0.116	0.438
δ_4	2.200	2.278	0.078	0.944	0.204	0.738
γ_2	−0.700	−0.459	0.241	0.704	0.304	0.701
γ_3	−1.100	−0.792	0.308	0.504	0.350	0.624
γ_4	−0.200	−0.025	0.175	0.802	0.258	0.692
α_2	1.500	1.570	0.070	0.900	0.128	0.397
α_3	2.600	2.709	0.109	0.895	0.188	0.580
α_4	4.500	4.540	0.040	0.936	0.231	0.871

not included in the model. Thus, if one is interested in assessing item scale effects, it is important to allow for random subject scale effects in the model.

5 Adolescent study

Data from a natural history study of adolescent smoking (Dierker and Mermelstein 2010) are used to illustrate application of the mixed-effects ordinal location-scale model. Students included in this study were either in 9th or 10th grade at baseline, 55.1 % female, and self-reported on a screening questionnaire 6–8 weeks prior to baseline that they had smoked at least one cigarette in their lifetime. The majority (57.6 %) had smoked at least one cigarette in the past month at baseline. Although the focus of this study is primarily on smoking, information about other substance use was also collected.

A total of 1263 students completed the baseline measurement wave. Following baseline, subjects were re-assessed at 6-month, 15-month, 2-, 4-, 5-year follow-ups. Here, we focus

on responses at the 5-year follow-up to five items from the Social Subscale of the Drinking Motives Questionnaire (DMQ, Cooper (1994)). This timepoint was chosen because of the increased level of drinking. In all, there were 1025 subjects with item responses at this timepoint. The five items were answered on a 5-point Likert scale indicating a subject's level of endorsement (1=almost never, 2=some of the time, 3=half of the time, 4=most of the time, 5=almost always) of: (1) as a way to celebrate; (2) because it is what most of my friends do when we get together; (3) to be sociable; (4) because it is customary on special occasions; (5) because it makes a social gathering more enjoyable. Nearly all 1025 subjects answered all items, however one and three subjects did not answer items 4 and 3, respectively.

Figure 1 displays the category response proportions for the five items. As can be seen, item 1 (drinking as a way to celebrate) has the largest proportion of responses in the higher categories, whereas item 4 (drinking because it is customary on special occasions) has the lowest. Four models of increasing complexity were fit to these data: (1) 1-parameter IRT model including only item difficulty parameters ($\boldsymbol{\beta}$) and random subject location effects (θ_i); (2) 2-parameter IRT model adding item discrimination parameters ($\boldsymbol{\delta}$); (3) 3-parameter model adding item scale parameters ($\boldsymbol{\gamma}$); and (4) 4-parameter location-scale model adding item scale discrimination parameters ($\boldsymbol{\tau}$), random subject scale effects (ζ_i), and location scale correlation ($\rho_{\theta\zeta}$). Model (4) is the proposed model in Equation (7). Based on likelihood-ratio tests, each more complex model fit significantly better than the previous simpler model, and the proposed 4-parameter model fit best ($\chi^2_6 = 120, p < .001$ comparing the 4- to the 3-parameter model).

Table 3 lists the estimates and 95 % confidence intervals for the item parameters of the location-scale model. For the item scale estimates, the first item is the reference (unity) WS

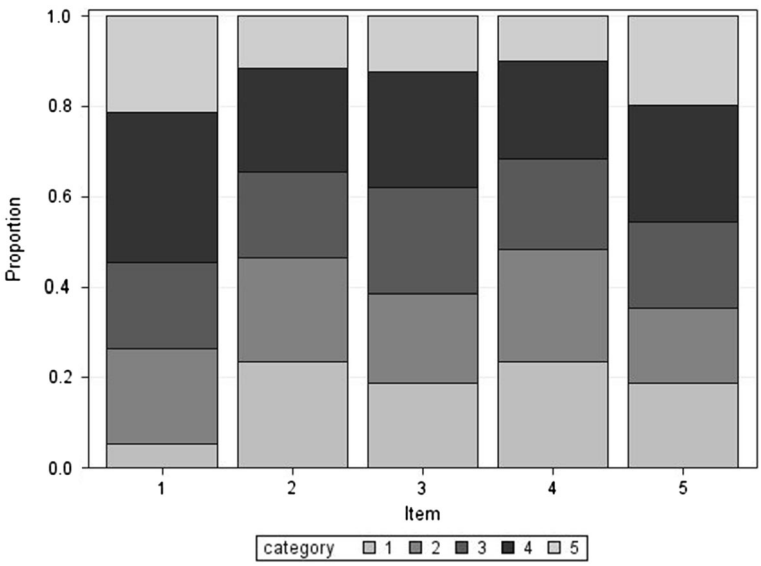

Fig. 1 DMQ social subscale items: category response proportions. Items: (1) as a way to celebrate; (2) because it is what most of my friends do when we get together; (3) to be sociable; (4) because it is customary on special occasions; (5) because it makes a social gathering more enjoyable. Categories: *1* = almost never, *2* = some of the time, *3* = half of the time, *4* = most of the time, *5* = almost always

Table 3 4-parameter location-scale model estimates (95 % confidence intervals)

Item	Difficulty	Discrimination	Scale*	Scale Discrimination
1. Way to celebrate	2.780	1.046	1.000	0.302
	(2.550, 3.010)	(0.886, 1.206)		(0.060, 0.543)
2. What friends do together	1.778	1.929	0.490	0.981
	(1.582, 1.974)	(1.722, 2.135)	(0.381, 0.629)	(0.658, 1.304)
3. To be sociable	2.040	1.830	0.322	1.112
	(1.841, 2.240)	(1.644, 2.015)	(0.245, 0.422)	(0.765, 1.458)
4. Customary special occasions	1.642	1.469	0.840	0.930
	(1.456, 1.828)	(1.287, 1.651)	(0.678, 1.040)	(0.633, 1.226)
5. Social gathering more enjoyable	2.291	2.326	0.277	1.053
	(2.064, 2.518)	(2.094, 2.558)	(0.196, 0.392)	(0.610, 1.496)

* exp(·) presented in this column; 1st item is reference (unity) WS variance

variance, and the remaining item estimates are exponentiated to represent ratios relative to the first item. In terms of item difficulty, the first item (way to celebrate) is the most frequently endorsed in the higher categories, whereas items 2, 3, and 4 are less endorsed. Item 5 is of medium difficulty. As expected, the item difficulties agree with the impressions provided by the observed response proportions in Fig. 1. The discrimination estimates also signal item 1 as a relatively poor discriminator of subject's location, whereas item 5 (social gathering more enjoyable) is the most discriminating item. For the item scale estimates, with item 1 as the reference item, all other items exhibit less scale than item 1, with the possible exception of item 4 (customary special occasions). As mentioned, these scale estimates give a sense of the degree to which the items fit the modeling in the numerator of Eq. (7), or in other words, the 2-parameter model. Thus, items 1 and 4 exhibit relatively worse fit, while items 2, 3, and 5 exhibit better fit. Finally, the scale discrimination estimates also single out item 1 as being a poor discriminator of subject scale. Taken together, item 1 is seen as a relatively poor item, whereas item 5 is perhaps the most useful item with moderate difficulty, good discrimination, low scale (error variance), and reasonable scale discrimination.

Turning to the random effects of the model (location θ_i and scale ζ_i); these are both distributed as standard normals in the population and allowed to be correlated. For this example, the random effect correlation is estimated at 0.20 (std err = 0.099) which suggests that subjects with higher (lower) location also have higher (lower) scale. Table 4 lists the empirical Bayes estimates of the random effects and the observed item responses for a few subjects with the highest, near-zero, and lowest estimates of the scale random effect. For comparison purposes, Table 4 also lists the location estimates that were obtained from the simpler 2- and 3-parameter models. Notice that the subjects with the highest scale all have rather extreme responses to the items, mostly 1s or 5s, and the location estimates vary quite a bit across the models. Conversely, those subjects with near-zero scale estimates have

Table 4 Subjects with high, near-zero, and low scale: Random effect estimates and observed item responses

Scale	2 parameter	3 parameter	4 parameter		observed item responses				
	Location $\hat{\theta}_i$	Location $\hat{\theta}_i$	Location $\hat{\theta}_i$	Scale $\hat{\zeta}_i$	Item 1	Item 2	Item 3	Item 4	Item 5
High	0.640	0.915	0.897	1.791	5	1	5	1	5
	−0.988	−0.766	−0.463	1.666	2	1	1	1	5
	−0.276	−0.106	0.186	1.504	5	1	2	1	5
Near 0	−0.411	−0.409	−0.448	0.002	4	2	1	2	3
	0.863	0.814	0.824	0.001	5	5	4	3	4
	0.636	0.457	0.442	0.001	5	4	4	4	3
Low	0.286	0.289	0.220	−1.122	4	3	3	3	4
	0.105	0.045	0.011	−1.234	4	3	3	3	3
	−0.546	−0.574	−0.624	−1.239	3	2	2	2	2

rather similar location estimates across models, as presumably their observed responses are more consistent with the simpler models. Finally, subjects with the lowest scale (most negative) give nearly the same response to all items, and their location estimates from the simpler models are slightly more positive.

6 Differential item functioning (DIF)

As detailed by Rijmen et al. (2003), in the mixed model framework it is straightforward to examine differential item functioning (DIF). One simply adds in interactions of the item parameters with subject variables. For example, for the difficulty parameters, $\boldsymbol{x}_j'\boldsymbol{\beta}$ becomes $\boldsymbol{x}_j'\boldsymbol{\beta} + S_i \times \boldsymbol{x}_j'\boldsymbol{\beta}^S$, where, say, S_i equals 0 for females and 1 for males. Here, to the 4-parameter location-scale model, we added such interactions with gender for all of the item parameters and observed a statistically improved fit ($\chi^2_{20} = 50, p < .001$). Table 5 lists the

Table 5 4-parameter location-scale model: Male gender DIF estimates (95 % confidence intervals)

Item	Difficulty	Discrimination	Scale*	Scale Discrimination
1. Way to celebrate	0.00	0.14	0.90	−0.16
	(−0.27, 0.27)	(−0.14, 0.41)	(0.68, 1.18)	(-0.63, 0.31)
2. What friends do together	0.43	−0.18	1.26	0.76
	(0.14, 0.73)	(−0.48, 0.13)	(0.79, 2.01)	(0.08, 1.45)
3. To be sociable	0.51	−0.28	1.04	0.15
	(0.24, 0.78)	(−0.55, -0.01)	(0.64, 1.67)	(−0.54, 0.83)
4. Customary special occasions	0.38	0.00	0.63	0.06
	(0.10, 0.66)	(−0.28, 0.29)	(0.44, 0.90)	(−0.56, 0.67)
5. Social gathering more enjoyable	0.68	−0.19	1.15	−0.51
	(0.36, 1.01)	(−0.53, 0.15)	(0.63, 2.09)	(−1.43, 0.41)

* exp(·) presented in this column

gender interaction estimates (and 95 % confidence intervals) for the 4 parameters and each of the items. Gender was coded with females as the reference cell, so these estimates indicate the degree to which males differed from females. Significant gender effects are indicated if the confidence intervals do not include zero, with the exception of the item scale estimates, which were exponentiated to represent ratios. For the latter, intervals not including one represent significant gender effects. As can be seen, most of the significant effects are in terms of the difficulty parameters, where males have significantly higher levels on all items except the first. Otherwise, DIF is limited to three other parameters: the third item (to be sociable) is less discriminating of subject location for males, the fourth item (customary special occasions) exhibits less item scale for males, and the second item (what friends do together) is more discriminating of subject scale for males.

7 Summary

The proposed model extends previous work on ordinal scale probit factor models (Skrondal and Rabe-Hesketh 2004) and the ordinal mixed location-scale model (Hedeker et al. 2009). Four types of item parameters are included in the model (difficulty, discrimination, scale, and scale discrimination), as well as random subject location and scale effects. These random effects are allowed to be correlated.

Estimates of the item scale parameters can be helpful in identifying items that are better or worse in terms of model fit. This can be particularly useful when one is constructing or refining quesitonnaires. Similarly, at the subject level, the random scale effects can be useful for detecting subjects with overly consistent or varying response styles. Covariates can be added to the model to examine whether such response styles are related to subject-level variables (e.g., demographic variables). Similarly, differential item functioning can be investigated by including interactions of covariates with the item parameters.

Results from a small simulation study indicated that the proposed model parameters were reasonably estimated using `SAS PROC NLMIXED`. Given that the data were simulated based on the proposed model, this was to be expected. However, estimates from the simpler 2-parameter model, in particular, were quite biased in the presence of item and subject scaling effects. The 3-parameter model estimates were more reasonable, however the item scale estimates were biased, exhibited poor coverage, and were overly precise in this case of random subject scale effects. This is perhaps not surprising as it is well-known that inferences for the (location) fixed effects are incorrect if a random subject (location) effect is not included in the model (Dorman 2008). Essentially, ignoring the random subject effect in the within-subject variance assumes that the errors across the items are independent of subjects, which is clearly a dubious assumption.

The quesionnaire considered in this paper was the Social Subscale of the Drinking Motives Questionnaire, which consisted of five items answered on a 5-point scale. Items varied in terms of all four item parameters, and in particular item 1 (way to celebrate) was seen to be a relatively poor discriminator of both subject's location and scale, and exhibited the greatest degree of misfit (item scale). The random effect scale estimates were able to identify both consistent and erratic responders. Additionally, the location estimates based on the simpler 2-parameter model were fairly different for such erratic responders, relative to the location estimates from the proposed model. Differential item functioning was examined in terms of gender, and was present primarily in terms of the item difficulty

parameters, where males were seen to more highly endorse all items except item 1 (way to celebrate).

The example and simulations presented here had a limited number of items. It is unclear how feasible this approach would be for a questionnaire with many items. We have fit the model to scales with more items, for example, a 10-item version of the Nicotine Dependence Syndrome Scale (Shiffman et al. 2004). However, for questionnaires with a large number of items the current approach may be computationally impractical or need to be improved upon. Certainly, `SAS PROC NLMIXED` can be slow to run, and users need to supply starting values, which can be challenging. Future work on application of this model to questionnaires with many items, say 50 or even 100, is necessary. Hopefully, this will make it possible to more easily apply this model to questionnaire data with many more items.

A related question concerns the sample size needed to estimate the proposed 4-parameter model. To address this, we repeated the simulation study but with subject sample sizes of 100, 200, and 300. Not surprisingly, results were poor for the $N = 100$ simulation, with a fair degree of bias and poor coverage (between .84 and .92) of the item discrimination of scale parameters $\boldsymbol{\tau}$. Results for the $N = 200$ simulation were clearly better, but still exhibited some bias and non-optimal coverage (.91–.94) for the item discrimination of scale parameters. The results for the $N = 300$ simulation were relatively similar to those presented in Table 1 for the $N = 500$ simulation. Thus, based on this limited simulation study, it would appear that sample sizes of at least 300 subjects are recommended, though this may need to be increased for questionnaires with more items.

While the simulation study assumed that the four-parameter was the true model, it is also of interest to examine results if the true model was either the 2- or 3-parameter model. For this, we repeated the simulation study using these two simpler models as the true models. In both cases, we obtained estimates close to zero for most of the additional unnecessary parameters of the 4-parameter model, and reasonable estimates for the 2- and 3-parameter models. The one excpetion is that the first item discrimination of scale estimate (τ_1) was badly biased (around 0.35 instead of 0.0). Also coverage was not ideal for the item discrimination of scale parameters, ranging from .80 to .88. However, comparing the 4-parameter model to the true 2-parameter model via likelihood ratio tests gave a rejection rate of 0.049, very close to the nominal 0.05 level. Similarly, comparing the 4-parameter model to the true 3-parameter model via likelihood ratio tests gave a rejection rate of 0.033. Finally, we also investigated whether misspecification of the mean function had an impact on the results. For this, we simulated data under the 2-parameter model, but also included two covariates, one each at the item and subjet level and both with effect sizes of 0.5 sd units. All three models (2-, 3-, and 4-parameter) performed poorly under this scenario, with no clear-cut difference between the models. Thus, it is important that the mean structure is correctly modeled to obtain reasonable item estimates.

Funding This study was funded by National Cancer Institute grant P01 CA98262 (Mermelstein, PI) and National Heart Lung and Blood Institute grant R01 HL121330 (Hedeker & Dunton).

Compliance with ethical standards

Conflict of interest All authors declare that he/she has no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Appendix: SAS PROC NLMIXED syntax

Below is a sample of syntax necessary to run the mixed-effects ordinal location-scale model described in this article. Uppercase letters are used for SAS specific syntax and lowercase letters are used for user defined entities. In terms of the variables used in this syntax, id is a subject identifier, y denotes the ordinal outcome and x1 to x3 are item indicators. Here, for simplicity, we illustrate the syntax consdering only three items and three response categories. The two cumulative logits are clogit1 and clogit2, and the two cumulative probabilities are cprob1 and cprob2. The random location effect is named theta, the random scale effect is zeta, and their correlation is corr. The location model is summarized by loc, while the scale model is given by scale.

```
PROC NLMIXED GCONV=1e-12;

PARMS beta1=-.5 beta2=.3 beta3=.1 alpha2=1

        delta1=1 delta2=1 delta3=1 gamma2=0 gamma3=0

        tau1=1 tau2=1 tau3=1 corr=0;

loc = beta1*x1 + beta2*x2 + beta3*x3 +

        (delta1*x1 + delta2*x2 + delta3*x3)*theta;

scale = EXP(gamma2*x2 + gamma3*x3 + (tau1*x1 + tau2*x2 + tau3*x3)*zeta);

clogit1 = (0 - loc) / SQRT(scale);

clogit2 = (alpha2 - loc) / SQRT(scale);

cprob1 = 1 / (1 + EXP(-clogit1));

cprob2 = 1 / (1 + EXP(-clogit2));

IF (y=1) THEN p = cprob1;

ELSE IF (y=2) THEN p = cprob2 - cprob1;

ELSE IF (y=3) THEN p = 1 - cprob2;

logl = LOG(p);

MODEL y ~ GENERAL(logl);

RANDOM theta zeta ~ NORMAL([0,0],[1,corr,1]) SUBJECT=id;
```

Users must provide starting values for all parameters on the PARMS statement. To do so, it is beneficial to run the model in stages using estimates from a prior stage as starting values and setting the additional parameters to zero or some small value. For example, one can start by estimating a random-intercepts ordinal model with item difficulty (beta1-beta3), item discrimination (delta1-delta3), and threshold parameter (alpha2). Estimates of these parameters can then be specified as starting values in a model that adds in the WS variance parameters (gamma1-gamma3). Finally, the full model with the additional parameters (tau1-tau3 and cor) can be estimated. In practice, this approach works well with PROC NLMIXED, which sometimes has difficulties in converging to a solution for complex models. Furthermore, for complex models, it is sometimes the case that the default convergence criteria is not strict enough. In the above syntax, the convergence criteria is specified as GCONV=1e-12 on the PROC NLMIXED statement.

References

Agresti, A., Lang, J.B.: A proportional odds model with subject-specific effects for repeated ordered categorical responses. Biometrika **80**, 527–534 (1993)

Aitkin, M.: Modelling variance heterogeneity in normal regression using GLIM. Appl. Stat. **36**, 332–339 (1987)

Bock, R.D., Aitken, M.: Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika **46**, 443–459 (1981)

Cooper, M.L.: Motivations for alcohol use among adolescents: development and validation of a four-factor model. Psychol. Assess. **6**, 117–128 (1994)

Cox, C.: Location-scale cumulative odds models for ordinal data: a generalized non-linear model approach. Stat. Med. **14**, 1191–1203 (1995)

Dierker, L., Mermelstein, R.: Early emerging nicotine-dependence symptoms: a signal of propensity for chronic smoking behavior in adolescents. J. Pediatr. **156**, 818–822 (2010)

Dorman, J.: The effect of clustering on statistical tests: an illustration using classroom environment data. Educ. Psychol. **28**, 583–595 (2008)

Ezzet, F., Whitehead, J.: A random effects model for ordinal responses from a crossover trial. Stat. Med. **10**, 901–907 (1991)

Harvey, A.C.: Estimating regression models with multiplicative heteroscedasticity. Econometrica **44**, 461–465 (1976)

Hedeker, D., Berbaum, M., Mermelstein, R.: Location-scale models for multilevel ordinal data: between- and within-subjects variance modeling. J. Probab. Stat. Sci. **4**, 1–20 (2006)

Hedeker, D., Demirtas, H., Mermelstein, R.J.: A mixed ordinal location-scale model for analysis of ecological momentary assessment data. Stat. Interface **2**, 391–402 (2009)

Hedeker, D., Gibbons, R.D.: A random-effects ordinal regression model for multilevel analysis. Biometrics **50**, 933–944 (1994)

Hedeker, D., Gibbons, R.D.: Longitudinal Data Analysis Longitudinal Data Analysis. Wiley, New York (2006)

Hedeker, D., Mermelstein, R.J.: A multilevel thresholds of change model for analysis of stages of change data. Multivar. Behav. Res. **33**, 427–455 (1998)

Hedeker, D., Mermelstein, R.J., Flay, B.R.: Models for intensive longitudinal data. In: Walls, T.A., Schafer, J.L. (eds.) Application of Item Response Theory Models for Intensive Longitudinal Data, pp. 84–108. Oxford University Press, New York (2006)

Ishwaran, H., Gatsonis, C.: A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. Can. J. Stat. **28**, 731–750 (2000)

Johnson, T.R.: On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. Psychometrika **68**, 563–583 (2003)

Leckie, G., French, R., Charlton, C., Browne, W.: Modeling heterogeneous variance-covariance components in two-level models. J. Educ. Behav. Stat. **39**, 307–332 (2014)

McCullagh, P.: Regression models for ordinal data (with discussion). J. R. Stat. Soc. Ser. B **42**, 109–142 (1980)

McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, New York (1989)

Rijmen, F., Tuerlinckx, F., De Boeck, P., Kuppens, P.: A nonlinear mixed model framework for item response theory. Psychol. Methods **8**, 185–205 (2003)
Saei, A., McGilchrist, C.A.: Longitudinal threshold models with random components. J. R. Stat. Soc. Ser. D (Stat.) **47**, 365–375 (1998)
Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychom. Monogr. **17**, 1–100 (1969)
Shiffman, S., Waters, A., Hickcox, M.: The nicotine dependence syndrome scale: a multidimensional measure of nicotine dependence. Nicotine Tob. Res. **6**, 327–348 (2004)
Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Chapman and Hall/CRC, New York (2004)
Toledano, A.Y., Gatsonis, C.: Ordinal regression methodology for ROC curves derived from correlated data. Stat. Med. **15**, 1807–1826 (1996)
Tosteson, A.N., Begg, C.B.: A general regression methodology for ROC curve estimation. Med. Decis. Mak. **8**, 204–215 (1988)
Tutz, G., Hennevogl, W.: Random effects in ordinal regression models. Comput. Stat. Data Anal. **22**, 537–557 (1996)