Modelling ordinal assessments: fit is not sufficient

David Andrich and Pender Pedler

The University of Western Australia

Corresponding Author

David Andrich

Graduate School of Education

The University of Western Australia

Crawley, Western Australia 6009

AUSTRALIA

Phone: +61 (0) 8 6488 1085

Email: david.andrich@uwa.edu.au

Acknowledgements

The research reported in this paper was supported in part by an Australian Research Council Linkage grant with Australian Curriculum, Assessment and Reporting Authority; The Schools Standards and Curriculum Authority of Western Australia as Industry Partners, and by Pearson plc. Barry Sheridan, Irene Styles and Tim Dunne provided valuable comments for the paper, and in particular, Tim Dunne drew our attention to the definition of a strictly log-concave (SLC) distribution. An unknown reviewer of a related paper noted the constraint on the generality of strict log-concavity for ordinal assessments.

Abstract

Assessments in ordered categories are ubiquitous in the social sciences. These assessments are assigned ordinal counts and analysed with probabilistic models. If the counts fit the model, it is assumed that no unaccounted for factors govern the distribution and that it is a random error distribution. However, because tests of fit utilise parameter estimates from the data, the data may fit the model even when the modelled distributions cannot be random error distributions. This paper applies the additional criterion of strict unimodality, common to all random error distributions, to decide if a modelled distribution is not a random error distribution. However, not only are common random error distributions strictly unimodal, they are also strictly log-concave, a stronger form of unimodality which ensures smooth transitions between probabilities of adjacent counts. The paper shows that the operation for determining the strict unimodality also ensures that the distribution is locally strictly log-concave around the measure of the entity of assessment.

Key Words: modelling ordinal assessments, modelling ordinal counts, strictly unimodal, strictly log-concave, random ordinal distributions, random ordinal errors.

1 Introduction and Summary

1.1 Ordinal assessments

Assessments in ordered categories, used when no measuring instrument of the kind used in the physical and natural sciences is available, are ubiquitous in the educational, behavioural and related sciences. The most elementary example is an assessment item with just two ordered categories, for example an item where responses are assessed as either *Incorrect* or *Correct*. Items with only two categories are termed *dichotomous*; those with more than two categories, and the focus of this paper, are termed *polytomous*. Educational assessment abounds with performances assessed with partial credit from no credit through to full credit. Many attitude questionnaires, with a Likert (1932) style response format of *Strongly Disagree, Disagree, Agree* and *Strongly Agree,* are of this kind. Fisher (1958, p. 290) gives an example in which serological readings from 12 blood cells are classified into five levels of reaction. An example from the health sciences is the Ashworth (1964) Scale, used to assess the degree of muscle tone in five ordered categories.

With all polytomous items, a clear ordering of the categories is implied representing distinct increasing levels on the assessment variable. The finite set of ordered categories are labelled with the successive integers 0, 1, 2, 3, ... *m*, enabling each response to the item to be assessed as an integer count. Note that each category label indicates a *position* in the sequence of categories, i.e. the 0th, 1st, 2nd, ... *m*th, the ordinal counts 0, 1, 2, 3, ... *m*. Because it is only the order relationships among these counts that have any substantive meaning, such assessments are referred to as *ordinal assessments* and characterised by *ordinal counts*.

Elementary data analyses treat these ordinal counts simply as measurements. More advanced analyses use a statistical model specifying that each observed count is modelled with a *distribution* of counts, a function of parameters of the item categories and the location of the entity of assessment, often a person. A successful modelling of the data implies that, given the assessment model and the values of the entity and item parameters, each distribution of ordinal counts is a random error distribution of ordinal assessments.

1.2 Modelling ordinal counts

There is a substantial literature on modelling ordinal counts in both contingency tables (e.g. Anderson, 1984; Andrich, 1979; Bock, 1975; Clogg & Shihadeh, 1994; Goodman, 1984; McCullagh, 1980) and psychometric contexts (e.g. Andersen, 1980; Andrich, 1978; Bock, 1972; Dossar & Mesbah, 2018; Drasgow, 1995; Ostini & Nering, 2006; Rasch, 1961; Samejima, 1969; Van der Linden & Hambleton, 1997; Wright & Masters, 1982).

In contrast to the study of properties of random errors of measurements in the physical sciences (Eisenhart, 1983; Stigler, 1986), there seems to have been no consideration of random errors of ordinal counts in the educational, behavioural and related social sciences literature, nor of any property that these distributions are required to satisfy. This omission is puzzling because each assessment item may be regarded as analogous to a measuring instrument. Although counts are not measurements (Wright, 1994), each ordinal count is a discrete analogue of a continuous measurement. Hence a distribution of random errors of ordinal counts is a discrete analogue of a distribution of random errors of measurements.

1.3 Summary

The focus of this paper is on the development of a criterion for distributions of ordinal counts to be random error distributions of ordinal assessments. By analogy to random error distributions of measurements, this paper first shows that random error distributions of ordinal counts are strictly unimodal (SU). We then show how empirical error distributions for each item can be inferred from a statistical analysis of a data set, and formalise the SU criterion for these distributions to be random error distributions. We illustrate how empirical error distributions can be interpreted from the category probability curves, and establish a necessary and sufficient condition for the item to satisfy the SU criterion. The necessary and sufficient condition is that the *thresholds*, points of equal probability of adjacent categories, are strictly ordered. We then show that the same condition not only ensures a SU distribution, but that it ensures a *locally* strictly log-concave (SLC) distribution, a form of unimodality that is stronger than strict unimodality (Ibragimov, 1956; Keilson & Gerber, 1971). We refer to a distribution which is locally, but not generally SLC, to be *almost* SLC.

If the ordering of thresholds is violated, then some empirical error distributions cannot be random error distributions and a systematic factor or factors, such as the operation of one or more categories, are interfering with the functioning of the item.

We stress that the paper is epistemological, concerned with the theory of assessment, distributions of random assessment errors, and the condition that ensures the distribution is SU and locally SLC. Determining whether an item satisfies such a distribution provides diagnostic information for reviewing the item. This review is independent of the discipline of the assessment variable, the form of the data set, whether in a contingency table or a psychometric context, and the statistical model and its class (Dossar & Mesbah, 2018). The SU and locally SLC properties are distinct from model fit. Because tests of fit utilise parameter estimates from the data, the data may fit the model even though the modelled

distributions are not random error distributions. While model fit remains a necessary criterion for modelling ordinal assessments, the SU criterion provides additional information distinct from model fit. Fit is not sufficient.

2. Random errors of ordinal counts and strict log-concavity

From the analogy between ordinal assessment and measurement, we show that distributions of random errors of ordinal counts are SU and formalise its properties.

2.1 Random errors of ordinal counts

Because ordinal assessments of an entity with an item are analogous to measurements of an object with an instrument, the analogy between random errors of measurements and random errors of ordinal counts is operationalized as follows. Distributions of random errors of measurements are Gaussian, unimodal with a single peak, for all objects measured (Eisenhart, 1983). Hence distributions of random errors of ordinal counts are required to be unimodal with a single peak for all entities assessed. However, the Gaussian distributions are continuous and we need to specify the particular form of discrete unimodality for distributions of ordinal counts to be random error distributions.

We begin with the discrete binomial model distribution, analogous to the continuous Gaussian, used to model frequency count data. When all systematic factors are accounted for, the binomial distribution is a discrete random error distribution for *all* values of its model parameters. The binomial distribution is unimodal, single-peaked, that is SU, a property characterised formally in the next section. Abstracting the criterion of random errors of frequency counts to random errors of ordinal counts, it follows that distributions of random errors of ordinal counts are also SU. We also note that other discrete random error distributions, such as the Poisson and the negative binomial, are also SU.

2.2 Statistical models for ordinal counts

We now describe the common features of statistical models for ordinal counts and show how their distributions describe both error distributions and random error distributions. A statistical model is a discrete random variable X specifying that each ordinal count x is modelled with a *distribution of counts* $p(x)$, $p(x) > 0$, $x \in \{0, 1, 2, \ldots m\}$, m finite. For simplicity, subscripts designating the entity, often a person, and item are omitted. For each ordinal count x let x^* be the unknown true count, a *mode* of the distribution $p(x)$, $p(x^*) \geq p(x)$ for all x. Let the integer $e = x - x^*$ be the *error* associated with the count x, a value of the random variable $E = X - x^*$. Note that when $x = x^*$, $e = 0$. The distribution $p(e)$ of E is the

discrete error distribution for the distribution $p(x)$ of X. Because $p(e) = p(x - x^*)$, the distribution $p(x)$ of counts X and the distribution $p(e)$ of errors E have the *same shape*. Thus when all systematic factors that govern the distribution are accounted for, the distribution $p(x)$ of counts X is a discrete *error distribution* which is random.

All statistical models for ordinal counts have the following features. First, for each polytomous item, the categories 0, 1, 2, 3, ... m, $m \geq 2$, m finite but possibly varying across items, are ordered so that the higher the category value, the higher the level of assessment on the variable. Second, each entity is characterised by a scalar parameter β with increasing β values implying increasing probabilities of assessments in the higher categories. Third, each item is characterised by a vector parameter $(\lambda) = (\lambda_1, \lambda_2, \ldots)$. Fourth, for each entity and each item, the assessed value of the entity with the item, the ordinal count x, is a value of a discrete random variable X. Fifth, each ordinal count x is modelled with the distribution of counts

$$p(x; \beta, (\lambda)), \quad p(x; \beta, (\lambda)) > 0, \quad x \in \{0, 1, 2, \ldots m\}, \qquad (1)$$

the distribution of X. Sixth, the set of random variables X are statistically independent for all parameter values. Hence provided all systematic factors are accounted for by the structure of the model and the values of the entity and item parameters, each distribution $p(x; \beta, (\lambda))$, $x \in \{0, 1, 2, \ldots m\}$, is *a random error distribution of ordinal counts*, SU for all parameter values.

2.3 Illustrating strict unimodality and strict log-concavity

The graph of the binomial distribution $p(x; 5, 0.6)$, $m = 5$, $p = 0.6$, in Figure 1a together with that of $ln\ p(x; 5, 0.6)$, the logarithms of these probabilities in Figure 1b, illustrate properties of the binomial distribution.

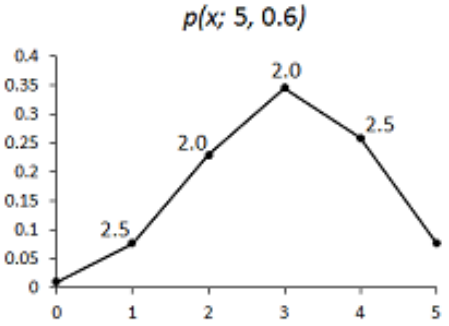

Figure 1a. Binomial distribution $p(x; 5, 0.6)$.

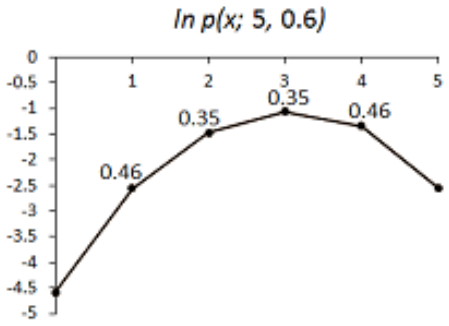

Figure 1b. Distribution $ln\ p(x; 5, 0.6)$.

Figure 1a illustrates the binomial distribution's unimodality. Its graph is SU with mode $x = 3$, and $p(0) < p(1) < p(2) < p(3) > p(4) > p(5) > p(m)$. Figure 1b illustrates the binomial

distribution's strict log-concavity. The graph of $ln\ p(x;\ 5,\ 0.6) = ln\ p(x)$ is discrete ∩-shaped, strictly concave curving downwards, or more simply *strictly concave*. The values included in Figure 1b are those of

$$ln\ p(x) - [ln\ p(x - 1) + ln\ p(x + 1)]/2, \qquad (2)$$

the extent to which the point $(x, ln\ p(x))$, $x \in \{1, 2, 3, 4\}$, is *above* the midpoint $(x, [ln\ p(x - 1) + ln\ p(x + 1)]/2)$ of the line joining the points $(x - 1, ln\ p(x - 1))$ and $(x + 1, ln\ p(x + 1))$. That each value is strictly greater than 0 confirms the strict concavity of $ln\ p(x;\ 5,\ 0.6)$, and hence the strict log-concavity of $p(x;\ 5,\ 0.6)$. The values included in Figure 1a are those of the ratio

$$p(x)^2/p(x - 1)p(x + 1) \qquad (3)$$

for each $x \in \{1, 2, 3, 4\}$. Because each value is strictly greater than 1, each value of the expression (2) is strictly greater than 0, confirming that $p(x;\ 5,\ 0.6)$ is SLC.

2.4 SU and SLC distributions of counts

We now formalise the relationship between strict unimodality and strict log-concavity.

Definition 1: A distribution $p(x)$ of a discrete random variable X modelling count data is strictly unimodal if for some count $x \in \{0, 1, 2, \ldots m\}$

$$p(0) < p(1) < \ldots < p(x-1) \leq p(x) > \ldots > p(m-1) > p(m). \qquad (4)$$

Thus the probabilities $p(x)$ of a SU distribution are either strictly increasing then strictly decreasing, always strictly increasing, or always strictly decreasing, with either a unique mode or a pair of adjacent equivalent modes. As a consequence, a SU distribution has no local minimum, that is there is no triplet $(x–1, x, x+1)$ for which

$$p(x) \leq p(x - 1) \text{ and } p(x) \leq p(x + 1). \qquad (5)$$

Definition 2: The distribution $p(x)$ of a discrete random variable X modelling count data $x \in \{0, 1, 2, \ldots m\}$, m finite, is strictly log-concave (e.g. Keilson & Gerber, 1971), if for all $x \in \{1, 2, \ldots m - 1\}$,

$$p(x)^2/p(x - 1)p(x + 1) > 1. \qquad (6)$$

We note the following three points. First, condition (6) for the binomial distribution $p(x;\ m,\ p)$, simplifies to

$$(m + 1 - x)(x + 1)/[(m - x)x] > 1,$$

verifying that $p(x; m, p)$ is SLC for all m, p, $0 < p < 1$. Second, in general a distribution $p(x)$ is SLC when the graph of $ln\ p(x)$ is strictly concave.

Finally the graph of $ln\ f(x; \mu, \sigma)$, the logarithm of the Gaussian distribution $f(x; \mu, \sigma)$, mean μ, standard deviation σ, is a $\cap$-shaped parabola and hence the continuous Gaussian distribution is also SLC. Strict log-concavity is a common property of both random count and random measurement error distributions (Ibragimov, 1956; Keilson & Gerber, 1971).

The Appendix proves that strict log-concavity is a stronger form of unimodality than strict unimodality. It ensures smoothness in the transitions between probabilities of adjacent counts that is not guaranteed by strict unimodality alone.

3. Empirical error distributions and the SU criterion

We now identify the empirical error distributions inferred from an analysis of a data set of ordinal counts, formalise the SU criterion, and show how the empirical error distributions can be interpreted from category probability curves (CPCs) and the ordering of their thresholds, the points of equal probability of adjacent ordinal counts. We then show that if the thresholds are in their natural order, then not only is the distribution SU, but if the location β of the entity is between the thresholds, then the probability $p(x)$ of the count defined by the thresholds is the mode of the distribution and the triplet of probabilities $p(x-1)$, $p(x)$, $p(x+1)$ satisfies the SLC ratio of Eq. (3). This justifies the terminology of *locally* SLC. This distribution has smoother transitions between probabilities than guaranteed only by strict unimodality, but not necessarily as smooth as a SLC distribution.

3.1 Empirical error distributions

Analysing a data set of ordinal counts with a statistical model X produces an *estimate* $\hat{\beta}$ of β for each entity, and *estimates* $(\hat{\lambda}) = (\hat{\lambda}_1, \hat{\lambda}_2, \dots)$ of $(\lambda) = (\lambda_1, \lambda_2, \dots)$ for each item. Inserting the estimated value $(\hat{\lambda})$ of the parameter (λ) for the item in Eq. (1) gives

$$p(x; \beta, (\hat{\lambda})) = p(x; \beta), \quad x \in \{0, 1, 2, \dots m\}, \tag{7}$$

the *empirical error distribution* of an entity β with the item $(\hat{\lambda})$ inferred from the analysis. For all β, the distribution $p(x; \beta)$ models the ordinal count x of the entity β with the item $(\hat{\lambda})$. By analogy to a distribution of replicated measurements, it is as if the distribution arose from independently replicated assessments of the same entity with the same item under identical conditions (Eisenhart, 1983).

3.2 The SU criterion

We now formalise the SU criterion, that if the empirical error distributions $p(x; \beta)$ of a polytomous item are random error distributions, they must satisfy the criterion for random assessment errors, expressed in two parts, as follows.

1. Irrespective of the data context or the statistical model used to analyse the data set of ordinal counts, and provided all systematic factors are accounted for by the structure of the model and the values of the model parameters, all empirical error distributions $p(x; \beta) = p(x; \beta, (\hat{\lambda}))$ of an entity β with the item $(\hat{\lambda})$ are random error distributions, SU for all entity β values.
2. Conversely, if an empirical error distribution $p(x; \beta) = p(x; \beta, (\hat{\lambda}))$ of an entity β with the item $(\hat{\lambda})$ is not SU, then it is not a random error distribution and a systematic factor or factors, such as the operation of one or more categories, are interfering with the functioning of the item.

We note the following three points. First, the SU criterion is an *a priori* requirement for the ordinal assessments to be consistent with assessment theory. Second, the SU criterion is an item criterion, distinct from model fit. Third, despite random errors sometimes being described as erratic and haphazard without any pattern, random error *distributions* are neither erratic nor haphazard but constrained to a specific pattern.

3.3 Category probability curves

Although category probability curves (CPCs) appear routinely in the literature on ordinal assessments, in both contingency table (e.g. Anderson, 1984; McCullagh, 1980) and psychometric contexts (e.g. Andrich, 2011; Samejima, 1969), no consideration has been given to interpreting properties of error distributions from them.

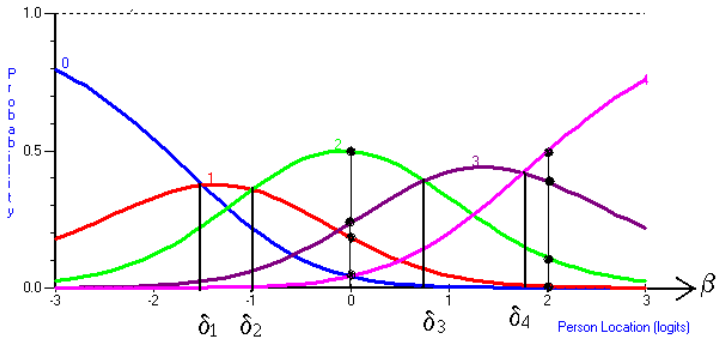

Figure 2. Category probability curves for item 1 with probabilities of ordinal counts highlighted (●) for the values $\beta = 0$, $\beta = 2$.

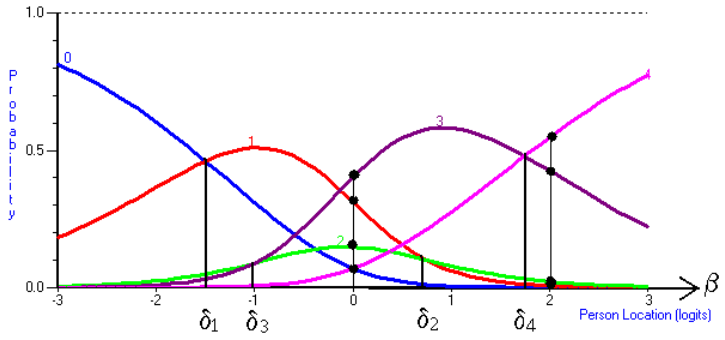

Figure 3. Category probability curves for item 2 with probabilities of ordinal counts highlighted (●) for the values $\beta = 0$, $\beta = 2$.

Given a polytomous item with estimates ($\hat{\lambda}$) and empirical error distributions $p(x; \beta)$, $x \in \{0, 1, 2, \ldots m\}$, the CPCs are the set of graphs, on the same pair of axes, of the probabilities $p(x; \beta)$ in Eq. (7) as functions of β. Figures 2 and 3 show the CPCs for two items, 1 and 2, each with the five categories 0, 1, 2, 3, 4. The intersections of a vertical line with the CPCs show the empirical distribution $p(x; \beta)$ for that β value. In Figures 2 and 3, these probabilities for $\beta = 0$ and $\beta = 2$ are highlighted as dots (●).

Figures 4a and 4b are the graphs of the distributions $p(x; 0)$ and $p(x; 2)$, $x \in \{0, 1, 2, 3, 4\}$, for item 1; Figures 5a and 5b for item 2. These graphs also include the values of the ratio $p(x)^2/p(x-1)p(x+1)$ of Eq. (3) for each $x \in \{1, 2, 3\}$. That corresponding values of the ratio are identical in Figures 4a and 4b, and in Figures 5a and 5b, is a property of the particular statistical model used to generate the CPCs. In Figures 4a and 4b, all values are strictly greater than 1, both distributions $p(x; 0)$ and $p(x; 2)$ are SLC and therefore, because strict log-concavity is a stronger form of unimodality than strict unimodality, also SU. In Figure 5a the bimodal distribution $p(x; 0)$ is clearly neither SU nor SLC with the ratio of Eq. (3) at $p(x; 2)$ having the value $0.2 \leq 1$. Although unimodal primarily because the location of the entity is close to the extreme, the distribution in Figure 5b is not SLC. Specifically the ratio of Eq. (3) at $p(x; 2)$ also has the value $0.2 \leq 1$.

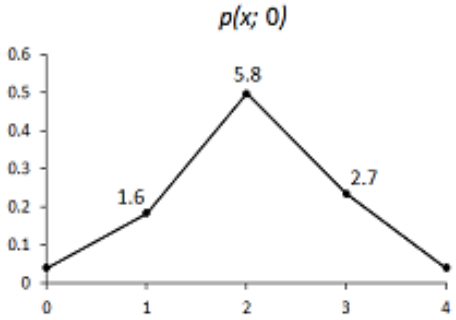

Figure 4a. Item 1: distribution $p(x; 0)$.

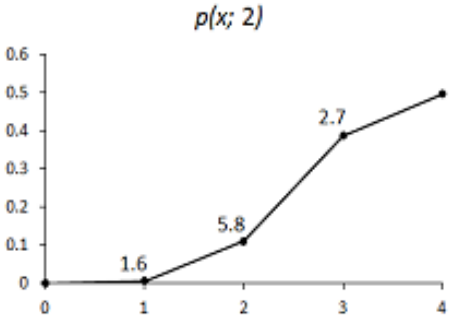

Figure 4b. Item 1: distribution $p(x; 2)$.

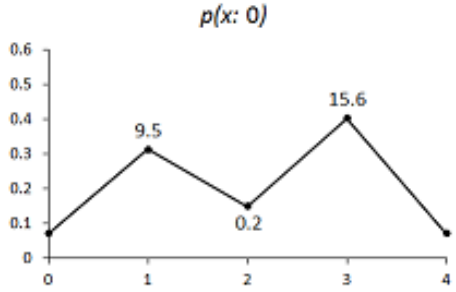

Figure 5a. Item 2: distribution $p(x; 0)$.

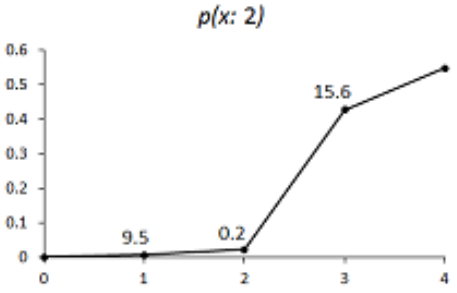

Figure 5b. Item 2: distribution $p(x; 2)$.

An item satisfies the SU criterion when its empirical error distributions $p(x; \beta)$ are SU for all β values. In Figure 2, despite both distributions $p(x; 0)$ and $p(x; 2)$ being SU, this is insufficient to conclude that item 1 satisfies the SU criterion. The criterion requires strict unimodality for *all* β values. In Figure 3, $p(2; 0)$ is a local minimum with $p(2, 0) < p(1,0)$ and $p(2, 0) < p(3,0)$. Item 2 does not satisfy the SU criterion.

The CPCs in Figures 2 and 3 also include the item *thresholds*, δ_1, δ_2, δ_3, δ_4, the β values for which the graphs of pairs of adjacent categories intersect. For example, in Figure 2 the curves $p(0; \beta)$ and $p(1; \beta)$ intersect when $\beta = -1.5 = \delta_1$. Note that in Figure 2, the thresholds are in their natural order, $\delta_1 < \delta_2 < \delta_3 < \delta_4$; in Figure 3, they are not, $\delta_1 < \delta_3 < \delta_2 < \delta_4$ with δ_2 and δ_3 reversed.

4. The SU criterion and reviewing polytomous items

We now establish the connection between CPC threshold order and whether the item satisfies the SU criterion by proving that strictly increasing item thresholds is both a necessary and sufficient condition for the item to satisfy the SU criterion. In addition, we show that if $\delta_x < \beta < \delta_{x+1}$, then $p(x; \beta)$ is the mode of the distribution and that it also satisfies the stronger

condition of the SLC ratio of Eq. (3). Examples from the literature illustrate its use for reviewing the operation of polytomous items.

4.1 Item thresholds

Consider a polytomous item with ordered categories 0, 1, 2, 3, ... m, m finite, parameters $(\lambda) = (\lambda_1, \lambda_2, \ldots)$, estimates $(\hat{\lambda}) = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots)$ and empirical distributions $p(x; \beta) = p(x; \beta, (\hat{\lambda}))$.

Definition 3: For each pair $(x - 1, x)$, $x \in \{1, 2, \ldots m\}$, of adjacent categories, the item threshold δ_x is the value of β for which the corresponding ordinal counts are equally likely. Specifically if $\beta = \delta_x$,

$$p(x - 1; \beta) = p(x; \beta) \qquad (8)$$

and δ_x is the β value for which the $x - 1$ and x category graphs in the CPCs intersect.

4.2 Threshold order and the SU criterion

We now prove that strictly increasing item thresholds is both a necessary and sufficient condition for the item to satisfy the SU criterion. Furthermore, if an item does not satisfy the criterion, the threshold disorder identifies which category or categories are malfunctioning.

Theorem: A necessary and sufficient condition for a polytomous item with parameters (λ), estimates $(\hat{\lambda})$ and thresholds $\delta_1, \delta_2, \ldots \delta_m$, to satisfy the SU criterion is for the thresholds to be in strictly increasing order. Specifically, if the thresholds are:

a. in strictly increasing order $\delta_1 < \delta_2 < \ldots < \delta_m$, then the item satisfies the SU criterion and there is no evidence that there are systematic factors not accounted for;

b. not in strictly increasing order, then the item does not satisfy the SU criterion and for each pair of reversed thresholds, $\delta_x \geq \delta_{x+1}$, $x \in \{1, 2, \ldots m - 1\}$, the distribution $p(x; \beta)$ is not SLC for all β values $\delta_{x+1} \leq \beta \leq \delta_x$ in the interval $[\delta_{x+1}, \delta_x]$, and there is evidence that there are systematic factors not accounted for and that category x in particular is malfunctioning.

Proof: If the item thresholds are in strictly increasing order, then $\delta_x < \delta_{x+1}$ for each $x \in \{1, 2, \ldots m - 1\}$ and *no* value of β satisfies

$$\beta \leq \delta_x \quad \text{and} \quad \beta \geq \delta_{x+1}.$$

First recall that if $\beta = \delta_x$, then $p(x\text{-}1; \beta) = p(x; \beta)$. Therefore if $\beta \leq \delta_x$, $p(x-1; \beta) \geq p(x; \beta)$, the ordinal count is equal or more likely to be $x - 1$ rather than x, and if $\beta \geq \delta_{x+1}$, $p(x; \beta) \leq p(x+1; \beta)$, the ordinal count is equal to or more likely to be $x + 1$ rather than x. It follows that, no value of β satisfies

$$p(x; \beta) \leq p(x-1; \beta) \text{ and } p(x ; \beta) \leq p(x+1; \beta).$$

and hence the distribution $p(x; \beta)$ has no count x which satisfies Eq. (5) and is a local minimum. Therefore the distribution $p(x; \beta)$ *is* strictly unimodal for *all* β values. This completes the proof of part a.

If the thresholds are not in strictly increasing order, then $\delta_x \geq \delta_{x+1}$ for at least one count $x \in \{1, 2, \ldots m-1\}$ and all entities β, $\delta_{x+1} \leq \beta \leq \delta_x$, in the interval $[\delta_{x+1}, \delta_x]$ satisfy

$$\beta \geq \delta_{x+1} \quad \text{and} \quad \beta \leq \delta_x.$$

Hence, $p(x; \beta) < p(x+1; \beta)$ *and* $p(x; \beta) < p(x\text{-}1; \beta)$, which satisfies Eq. (5) and establishes that count x is a local minimum. Therefore, the distribution $p(x; \beta)$ is not SU. This completes the proof of part b.

Corollary. If the item thresholds are in strictly increasing order, then $\delta_x < \delta_{x+1}$ for each $x \in \{1, 2, \ldots m-1\}$ and β satisfies $\delta_x < \beta < \delta_{x+1}$, then

a. x is the mode of the SU distribution;

b. the SLC ratio, $p(x)^2/p(x-1)p(x+1) > 1$ is satisfied for x

Proof: If $\delta_x < \beta < \delta_{x+1}$ and $\delta_x < \delta_{x+1}$ for each $x \in \{1, 2, \ldots m-1\}$, then $p(x; \beta) > p(x-1; \beta) > p(x-2; \beta) \ldots p(x=0; \beta)$ and $p(x; \beta) > p(x+1; \beta) > p(x+2; \beta) > \ldots p(m; \beta)$, which completes the proof of part a.

From $p(x; \beta) > p(x-1; \beta)$ and $p(x; \beta) > p(x+1; \beta)$, $p(x; \beta)^2 > p(x-1; \beta)p(x+1; \beta)$ which completes the proof of part b.

Note that the distribution $p(x; \beta)$ is SLC only for β values in the interval (δ_x, δ_{x+1}) and only for the particular count x. This justifies the terminology of *locally* SLC, also referred to as almost SLC.

Returning to Figure 2, the thresholds are in strictly increasing order, $\delta_1 < \delta_2 < \delta_3 < \delta_4$, item 1 satisfies both the SU criterion for all values of β and it appears that all systematic factors are accounted for. For $\beta = 0$, $\delta_1 < \beta < \delta_2$, the count $x = 2$ is the mode and

$p(2;\beta)^2 > p(1;\beta)p(3;\beta) = 5.8 > 1$, *which satisfies the SLC ratio of Eq.(6).* In Figure 3, the thresholds are not in strictly increasing order and hence item 2 does not satisfy the SU criterion for all values of β. Because $\delta_2 > \delta_3$, entities with a β value in the interval $[\delta_3, \delta_2]$ result in the count 2 being a local minimum indicating that the category 2 is malfunctioning.

A researcher can determine whether a polytomous item satisfies the SU criterion simply by checking the threshold order of its CPCs. If the thresholds are in strictly increasing order, the item satisfies the SU criterion for all locations β and the SLC ratio at $p(x;\beta)$ for any $\delta_x < \beta < \delta_{x+1}$, implying there is no evidence that not all systematic factors are accounted for. If the thresholds are not in strictly increasing order, the item does not satisfy the SU criterion, there is evidence that an unaccounted for factor is disturbing the assessments, and the threshold disorder provides diagnostic information identifying the malfunctioning category or categories.

4.3 Reviewing polytomous items

The following four examples from the literature illustrate the use of the SU criterion for reviewing the operation of polytomous items.

In a psychology text book, Embretson and Reise (2000, p. 109, Fig. 5.4) interpret data from a 12 item neuroticism questionnaire, each item with ratings in five ordered categories. The operation of item 3, highlighted by the authors but without comment, is summarised by its CPCs in Figure 6. We note that the threshold order, $\delta_1 < \delta_3 < \delta_2 < \delta_4$, has δ_2 and δ_3 reversed and hence item 3 does not satisfy the SU criterion. Because $\delta_2 \geq \delta_3$, we conclude that category 2 is malfunctioning and interfering with the item 3 assessment of neuroticism.

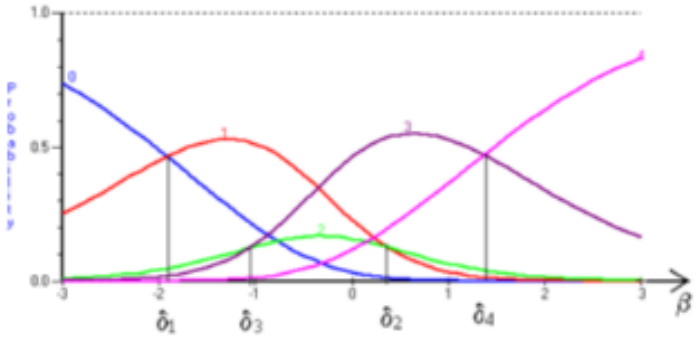

Figure 6. Category probability curves for the neuroticism item 3.

The second example is an analysis (Fisher, 1958, p. 290) of 152 serological readings from 12 blood cells, each classified into one of five levels of reaction and summarised in a 12 by 5

contingency table. Andrich (1996, p. 31) re-analysed the data summarising the level of reaction by its CPCs in Figure 7. The thresholds, $\delta_1 < \delta_2 < \delta_3 < \delta_4$, are in strictly increasing order and hence the item, the level of reaction, satisfies the SU criterion and the SLC ratio for entities between pairs of adjacent thresholds. Fisher used a least squares method to analyse the data and check that the levels of reaction, corresponding to the thresholds, were correctly in increasing order, noting as follows.

> *It will be observed that the numerical values...lie... in the proper order for increasing reaction. This is not a consequence of the procedure by which they have been obtained, but a property of the data examined* (Fisher, 1958, p. 294).

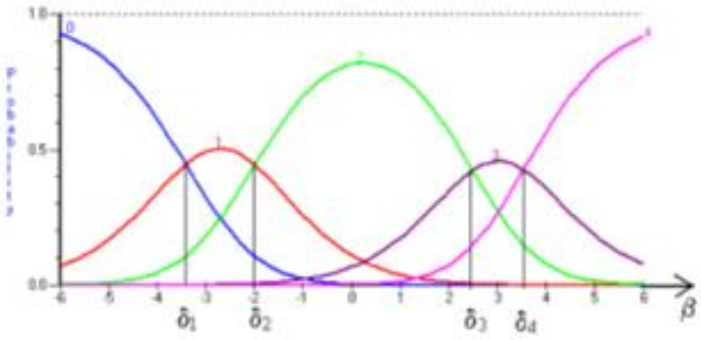

Figure 7. Category probability curves for the level of reaction.

The third example is an analysis (Anderson, 1984, p. 16) of the severity of 223 boys' dreams. The age of each boy, from 5 to 15, was classified into one of five age groups, while the severity of each dream was assessed on a four point scale of increasing severity, enabling the data to be summarised in a 5 by 4 contingency table. Andrich (1996, p. 28) re-analysed the data and summarised the severity of the dreams by its CPCs in Figure 8. The three thresholds, $\delta_2 = \delta_3 < \delta_1$, are not in strictly increasing order and hence the item, the severity of the dreams, does not satisfy the SU criterion. Because $\delta_1 \geq \delta_2$ and $\delta_2 \geq \delta_3$, both middle categories 1 and 2 are malfunctioning and interfering with the assessment. The source of the malfunctioning needs to be identified empirically or by a closer examination of the counts.

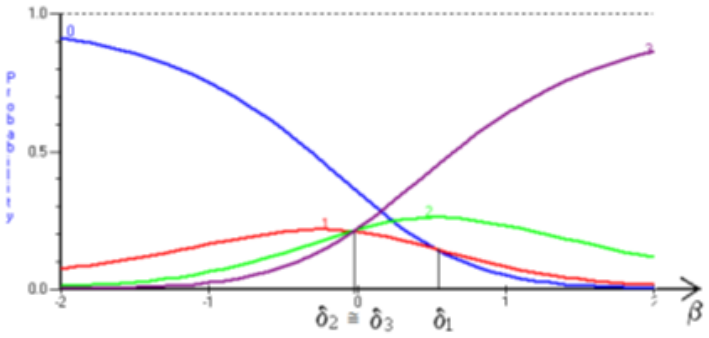

Figure 8. Category probability curves for the severity of boys' dreams.

The fourth example is from the health sciences. Andrich (2011) analysed the data from 656 people with multiple sclerosis whose eight lower limb muscles, four on each side, were each assessed for muscle tone. For this assessment, clinicians used the five category Ashworth Scale; *Limb Rigid* (*minimal movement), Increased Tone (restricting movement), Increased Tone (easily flexed), Catch,* and *Normal Tone*. The entities of assessment are the 5,248 lower limb muscles, eight for each of the 656 people; the Ashworth Scale is an eight-item instrument, one for each lower limb muscle. Despite the data fitting the model, none of the eight muscles satisfied the SU criterion. Each of the CPCs had thresholds ordered $\delta_1 < \delta_2 < \delta_4 < \delta_3$, with $\delta_3 \geq \delta_4$ as in Figure 9, the CPCs for muscle 5, left knee flexion. Because $\delta_3 \geq \delta_4$, it is category 3, muscle tone *Catch*, that is malfunctioning and interfering with the assessment for all eight muscles. When assessing a muscle with tone *Catch*, β values in the interval $[\delta_4, \delta_3]$, clinicians were more likely to assess it incorrectly as either *Increased Tone (easily flexed)* or *Normal Tone*. Their interpretation of *Catch* is not sufficiently distinct from its two adjacent categories. A solution to this problem will involve consulting with the clinicians and possibly developing and trialling a new specification of the Ashworth Scale categories.

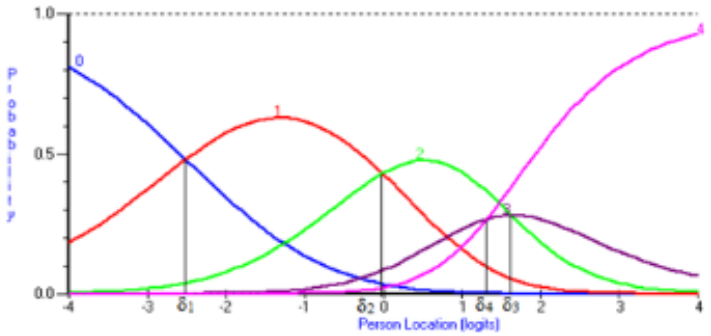

Figure 9. Category probability curves for muscle 5, left knee flexion.

4.4 Further implications of SU distributions

The paper has primarily focussed on the SU property as a criterion which ensures that an inferred error distribution of replications is random. Satisfying this criterion in turn implies that there is no evidence that an unaccounted for factor has affected the distribution. We now consider briefly two further implications, one statistical and one substantive.

In maximum likelihood (ML) estimation, commonly used to estimate parameters of models for ordinal assessments in both contingency table and psychometric contexts, there is a requirement that there is a likelihood that is a maximum. If thresholds are reversed, then for a series of locations of the entity, the distribution is at least double-peaked, and perhaps has multiple peaks, and although it can be calculated algebraically, the ML estimate of a single object parameter is rendered meaningless. In least squares estimation, and in the derivation of the Gaussian distribution to characterise the random error distribution of replicated measurements, it is assumed that the distribution is symmetric and single-peaked (Eisenhart, 1983). Elementary text books warn against characterising a distribution with a single location if the distribution is, for example, bimodal or has multiple peaks.

From a substantive perspective, it can be considered that the analysis provides a form of *quality control* of the operation of an item with ordered categories. In many cases, whether in a contingency table or psychometric context, individual or clinical decisions may be made based on the assessment. In most cases each person only responds once to an item. Although random errors are expected, it would be unacceptable if the most accurate assessment is not the mode with the highest probability in a single-peaked, SU distribution. Thus it is expected that the probability of an incorrect response should decrease with the size of the error, and thus reduce the risk of misclassification and in many cases misdiagnosis and mistreatment. Moreover, it would be expected that the mean of the inferred distribution of replications is close to the mode of a SU distribution. In contrast, if the distribution has two peaks, then for a region of the continuum the mean of the inferred replications is somewhere between the peaks, the probability of an error does not decrease with its size, and two assessments on either side of the mean are more likely than the mean. Depending on the seriousness of the implications of misclassification, this kind of distribution seems unacceptable in clinical decision making.

The question of the distance between successive thresholds that is acceptable may arise. An answer that is unacceptable is that the distance is statistically significantly greater than zero. The reason this is unacceptable is that significance depends on the sample size, and the

justification of the size of distance between thresholds should depend on the substantive implications of making an error, not on the sample size that happens to be available. Andrich (2016) provides an approach to deciding the size of the distance between thresholds relative to the substantive context.

5. Conclusion

An SU criterion for the random error distribution of ordinal assessments, augmented by a locally strictly log-concave distribution referred to as almost SLC, is developed and its use illustrated for reviewing the operation of polytomous items. The critical evidence for this review is in the item's CPCs and hence the review is independent of the discipline of the variable, the form of the data set, or the model used to analyse the data. A researcher determines whether an item satisfies the SU criterion and its augmented local SLC property simply by checking the CPC threshold order. If the thresholds are in strictly increasing order, the item's distributions satisfy the SU criterion and are locally SLC. If the thresholds are not in strictly increasing order, the item's distributions do not satisfy the SU criterion and the threshold disorder provides diagnostic information identifying the item's malfunctioning category or categories.

The focus of this paper is on the ubiquitous ordinal assessments and a criterion for distributions of ordinal counts of these assessments to be random error distributions. From the analogy with measurement, the ordinal assessment of an entity with a polytomous item is analogous to measuring an object with a measuring instrument. It follows that a distribution of random errors of ordinal counts is a discrete analogue of a distribution of random errors of measurements, which are not only SU, but the more strongly SLC. Therefore, by analogy to measurements, the distribution is required to be at least SU and locally SLC. These properties not only ensure a strictly unimodal distribution, but one with smoother transitions between probabilities of adjacent counts than provided only by strict unimodality. The paper shows how empirical error distributions can be inferred from a statistical analysis of a data set of ordinal counts and interpreted from the item CPCs. The requirement that these empirical distributions must be distributions of random assessment errors provides additional information distinct from model fit, establishing that fit is not sufficient.

Appendix

This Appendix proves that strict log-concavity is a stronger form of unimodality than strict unimodality.

Theorem. For a distribution $p(x)$ of a discrete random variable X, where the probabilities are strictly positive, $p(x)>0; x=0, 1, 2, \ldots m;$

(a) strict log-concavity implies strict unimodality; but
(b) strict unimodality does not imply strict log-concavity.

Proof: We prove the equivalent contrapositive, not strict unimodality implies not strict log-concavity. If the distribution $p(x)$ is not strictly unimodal, it has a local minimum at x for some $x; x=0, 1, 2, \ldots m-1.$ Then

$$0<p(x)\le p(x-1) \text{ and } 0<p(x)\le p(x+1),$$

from which

$$0<p(x)^2\le p(x-1)p(x+1),$$

and hence distribution $p(x)$ is not SLC, which completes the proof of part a.

In part b, we show a counterexample. Consider the SU distribution $p(0)=0.2, p(1)=0.3, p(2)=0.5$ with mode 2. Because

$$p(1)^2 = 0.09 \le 0.10 = p(0)p(2)$$

the distribution $p(x)$ is not strictly log-concave at x = 1 and hence not strictly log-concave, which completes the proof of part b.

Such a distribution, though SU, does not have transitions between probabilities that are as smooth as those which are SLC.

References

Andersen, E. B. (1980). *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Anderson, J. A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B, 46,* 1–30.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–574.

Andrich, D. (1979). A model for contingency tables having an ordered response classification. *Biometrics, 35*, 403–415.

Andrich, D. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye & C.C. Clogg (Eds.), *Analysis of categorical variables in developmental research* (pp. 3–35). Orlando, Florida: Academic Press.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research, 11*, 571–585.

Andrich, D. (2016). Rasch Rating-Scale Model. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume One: Models* (pp. 75–94). Boca Raton, Florida: Chapman and Hall/CRC.

Ashworth, B. (1964). Preliminary trial of carisoprodol in multiple sclerosis. *Practitioner, 192*, 540–542.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.

Clogg, C. C. & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, California: Sage Publications.

Dossar, P. & Mesbah, M. (2018). Cumulative or adjacent logits: Which choice for an ordinal logistic latent variable model? *Communications in Statistics - Theory and Methods. DOI: 10.1080/03610926.2015.1060342.*

Drasgow, F. (Ed.). (1995). Special Issue on Polytomous IRT models. *Applied Psychological Measurement, 19*(2).

Eisenhart, C. (1983). Law of error I: Development of the concept. In S. Kotz & N. L Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 4, pp. 530 – 547). Toronto: Wiley.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. London: Erlbaum.

Fisher, R. A. (1958). *Statistical methods for research workers* (13th ed.). New York: Hafner.

Goodman, L. A. (1984). *The analysis of cross classifications having ordered categories*. Cambridge, MA: Harvard University Press.

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Theory of probability and its applications*, Volume I, 255–260.

Keilson, J. & Gerber, H. (1971). Some results for discrete unimodality. *The Journal of the American Statistical Association, 66,* 386–389.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5–55.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

Ostini, R. & Nering, M. L. (2006). *Polytomous item response models*. Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 07–144. Thousand Oaks and London: Sage Publications.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, IV* (pp. 321–334). Berkeley CA: University of California Press. Reprinted in D. J. Bartholomew (Ed.), *Measurement Volume I. Benchmarks in Social Research Methods* (2006, pp. 319 –334). London: Sage Publications.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement No. 17, 34*(4, Pt. 2), 1-100.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Van der Linden, W. & Hambleton, R. (Eds.). (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

Wright B. D. (1994). Measuring and Counting. *Rasch Measurement Transactions, 8*(3), *371.*

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.