Please see the online version of this article for supplementary materials.

Modeling and Evaluating Repeatability and Reproducibility of Ordinal Classifications

Jeroen DE MAST

Institute for Business and Industrial Statistics
University of Amsterdam
Plantage Muidergracht 12, 1018 TV Amsterdam
The Netherlands
(*j.demast@uva.nl*)

Wessel N. VAN WIERINGEN

Department of Epidemiology and Biostatistics
VU University Medical Center
P.O. Box 7075, 1007 MB Amsterdam
The Netherlands
and
Department of Mathematics
VU University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam
The Netherlands

This paper argues that currently available methods for the assessment of the repeatability and reproducibility of ordinal classifications are not satisfactory. The paper aims to study whether we can modify a class of models from Item Response Theory, well established for the study of the reliability of categorical measurements in psychometrics and education, for use in business and industry, and whether the resulting approaches offer a satisfactory solution. The fitted models can be presented graphically, but also allow the calculation of probabilities of correct ordering and consistent classification. In addition, the model-based approach allows refined diagnostics, giving the user insight into the workings of a classification procedure, which is vital information for a user willing to improve a poor classification procedure. The approach is illustrated from a real-life example, and the proposed analysis is contrasted to two popular alternative analyses, based on Goodman and Kruskal's gamma and Kendall's coefficient of concordance. The datasets and mathematical proofs are available as online supplemental materials.

KEY WORDS: Agreement; Attribute data; Categorical data; Concordance; Gauge capability; Gauge repeatability and reproducibility; Item response theory; Measurement system analysis; Ordinal data.

1. INTRODUCTION

Classifications can be nominal or ordinal. In the former case, there is no particular order in the classes of the used scale. With ordinal classifications the classes are ordered. Ordinal classifications are omnipresent in business and industry, as in the following examples:

- The classification of manufacturing faults into "minor," "major," "critical."
- Quality ratings using a scale such as "reject," "critical," "acceptable," "good," "very good."
- Sorting of produce or natural materials in grades such as I, II, III, and IV.

Ordinal classification is measurement on an ordinal scale, and as with numerical measurements, its reproducibility and repeatability (R&R) are important characteristics (especially in view of the often critical application of these measurements). This makes the development and study of methods for the assessment of the R&R of ordinal measurements a highly relevant subject.

Standard methods for the assessment of the quality of *numerical* measurement systems include gauge repeatability and reproducibility (gauge R&R) studies (Burdick, Borror, and Montgomery 2003). Gauge R&R studies (and similar approaches such as the intraclass correlation coefficient) express a measurement system's R&R in terms of standard deviations or (Pearson) correlations, both of which are not defined for ordinal data.

Methods that are offered in literature for ordinal data include approaches based on the concept of agreement, such as the kappa index (de Mast 2007; de Mast and van Wieringen 2007). Only considering consistency in absolute value, these methods treat ordinal data as nominal data (i.e., the order among the classes is not taken into account). Suppose, for instance, that an appraiser rates 10 objects on a four-point ordinal scale $\{a, b, c, d\}$ and the results are $c, b, d, d, b, b, d, b, d, c$. Suppose a second appraiser rates the same objects but finds $b, a, c, c, a, a, c, a, c, b$. Agreement methods treat these data as though they are nominal, and would yield that the agreement between these sequences is nil. Such evaluation overlooks that these appraisers are in fact quite consistent. Namely, appraiser 1 rates the first object larger than the second object, as does appraiser 2. This holds for all pairs of objects; that is, the appraisers are consistent in ordering objects relative to each other. Agreement methods for nominal data are incapable of reflecting this sort of consistency in order, and thus an important aspect of the quality of ordinal classifications is ignored.

An ad hoc and rather arbitrary (and therefore unsatisfactory) modification of agreement methods for ordered scales is the weighted kappa (Cohen 1968). Other methods are based on measures of association defined for ordinal scales (Haberman 1988 gives an overview). Goodman and Kruskal's gamma

TECHNOMETRICS, FEBRUARY 2010, VOL. 52, NO. 1
DOI 10.1198/TECH.2009.08052

(Goodman and Kruskal 1954), for example, is based on the notion of concordance, as is Kendall's tau (Kendall and Gibbons 1990, chapter 1). Concordance refers to the extent to which appraisers are consistent in ordering objects relative to each other; the two abovementioned sequences are fully concordant. Goodman and Kruskal's gamma expresses R&R as a difference between the probability of observing a concordant pair, and the probability of observing a discordant pair (details of this and the other indices discussed here are given in the last section). A different approach is taken in Kendall's coefficient of concordance W, which is a generalization of Spearman's rho (Kendall and Gibbons 1990, p. 117). Here, the idea is to transform ratings into rankings, and treat these rankings as though they were on an interval or ratio scale (with equidistant classes), and apply ANOVA-like techniques (such as sums of squares).

In our view, none of these approaches provides a satisfactory method for expressing the result of R&R studies for ordinal data. Some of these methods treat ordinal data as though they are nominal data (kappa), numerical data (weighted kappa), or rankings that in turn are treated as numerical data (Kendall's W). All of them capture the behavior of the measurements in a single number in between 0 (implying completely random and thus uninformative classifications) and 1 (implying perfect R&R). Whereas these extremes 0 and 1 have clear interpretations, the intermediate values are hard to give a tangible meaning. It is hard to substantiate that they convey more information to the user than that the ratings are "somewhere" in between perfectly repeatable and purely random. Such a single number may be useful for comparing measurement systems relative to each other, but it is hard to see its practical value for the user in evaluating a single measurement system. They provide little insight, and make the question as to how large the index should be to indicate an acceptable R&R hopelessly arbitrary.

In measurement in the social sciences the last decades witnessed the development of Item Response Theory (IRT). Lord (1980) is the classic introduction, while Embretson and Reise (2000) give a recent overview. In IRT, measurements are not evaluated based on a single index; instead, advanced statistical models are fitted to the data. The theoretical development of IRT has been substantial, and IRT methods are routinely applied to study a range of psychological and educational measuring instruments and tests.

Searching for an approach to characterize the R&R of ordinal measurements in a way that provides more insight than the single-index approaches discussed above, this paper aims to study whether IRT methods can be modified for use with the types of ratings that are common in business and industry, and whether the resulting approaches should be considered a better alternative to current methods. Where the original applications of IRT involve test items and tested persons, R&R experiments involve objects, appraisers and repeated measurements per appraiser. We propose an estimation method for fitting the proposed models (including suggestions for model diagnostics), and we demonstrate how the fitted models can be interpreted to build understanding of the R&R of classifications. A real-life example from industry serves as the basis for our discussion that evaluates the advantages and disadvantages of IRT modeling compared to the currently available single-index approaches.

2. EXPERIMENTAL DESIGN AND MODELING

We consider an ordinal measurement procedure, which classifies objects on an ordered scale $\{1, 2, \ldots, H\}$ (for example, {"reject," "critical," "acceptable," "good"}). This classification is intended to order the objects according to a certain property (such as *quality*), which is not directly observable. Following considerations in Goodman and Kruskal (1954), we reason that the fact that a scale's categories are ordered suggests this latent, underlying property is a continuum. We will denote the latent value of an object on this continuum as X; it would be referred to as *the measurand* in ISO's *Guide to the Expression of Uncertainty of Measurements* (ISO 1995), but we will refer to it as an object's *true value*. To assess the R&R of a classification procedure one takes I objects, which are classified J times by each of K appraisers into one of the classes $h = 1, \ldots, H$. The data are denoted Y_{ijk}, with $i = 1, \ldots, I$ indexing objects, $j = 1, \ldots, J$ indexing appraisers, and $k = 1, \ldots, K$ indexing repeated measurements per appraiser. We assume that the I objects are a sample representative for the process in which the classification procedure is used.

Of interest for R&R studies is the joint distribution of the $\{Y_{ijk}\}_{j=1,\ldots,J;k=1,\ldots,K}$, and in particular the association structure between the repeated measurements (a lack of association implying a poor R&R). We choose to model the Y_{ijk} using a latent variable model. The main alternative are log-linear models; although these are powerful means to analyze association structures (see Agresti 1988), the advantage of latent variable models is that the cause of the association among repeated measurements—the objects' true values—is modeled explicitly. Consequently, the variation in the measurements is explicitly attributed to a systematic part (variation among true values) and a random part (measurement variation), a practice which resembles the typical manner in which gauge R&R studies for numerical measurements are modeled.

The objects' true values X_i are assumed stochastically independent, and have a density f_X. Let

$$q_j(h|x) := P(Y_{ijk} = h|X_i = x). \tag{1}$$

For each appraiser j, and for $h = 1, \ldots, H$ the $q_j(h|x)$ could be seen as a function (from $\mathbb{R}$ to [0, 1]) in x. These functions, known as *characteristic curves*, determine the probability of observing a certain response category h, given the object's true value x. As announced in the Introduction, this paper studies a class of models borrowed from IRT. In particular, we will use Masters' (1982) Partial Credit Model, in the generalized form proposed by Muraki (1992). This model was developed for ordered polytomous responses, and it assumes that

$$q_j(h|x) = \frac{\exp(\sum_{m=1}^{h-1} \alpha_j(x - \delta_{jm}))}{\sum_{n=1}^{H} \exp(\sum_{m=1}^{n-1} \alpha_j(x - \delta_{jm}))}. \tag{2}$$

This model has the following parameters:

- The threshold points $\delta_{jh}, j = 1, \ldots, J$ and $h = 1, \ldots, H-1$, which are the points of intersection of the curves $q_j(h|x)$ and $q_j(h+1|x)$, namely, $P(Y_{ijk} = h|X_i = \delta_{jh}) = P(Y_{ijk} = h+1|X_i = \delta_{jh})$. Loosely said, the $\delta_{j1}, \ldots, \delta_{jH}$ determine how appraiser j relates response categories $h = 1, \ldots, H$ to the latent continuum of true values x, and we will refer to them as appraiser j's *category boundaries*.

- Discrimination parameters $\alpha_j > 0, j = 1, \dots, J$. These determine the width of the curves $q_j(h|x)$. Larger α_j imply that the curves are narrower, and (as will be shown later) that repeatability is better (i.e., the appraiser's ratings are more discriminative).

The origin and scale of the continuum on which the true values X_i vary are arbitrary and inconsequential. As a result, the δ_{jh} and x_i are fitted only up to a linear transformation. To simplify the notation in the estimation algorithm, we fix the origin and scale of the x continuum at μ_X and σ_X, whence we can take the X_i to have zero mean and unit variance [in the estimation section we will assume the X_i to be N(0, 1) distributed]. Without these or similar restrictions, the model suffers from an identifiability problem.

The experimental model described by (1) and (2) is completed by the assumption of conditional independence. This standard assumption in latent variable models states that conditional on X_i and for fixed appraisers $j = 1, \dots, J$, the $Y_{i11}, \dots, Y_{iJK}$ are independent.

Figure 1 shows characteristic curves $q_j(h|x)$ for a 4-point scale. In this example, $\alpha_j = 3$, $\delta_{j,1} = -1.0$, $\delta_{j,2} = -0.5$, $\delta_{j,3} = 1.5$. Note that for each x we have $\sum_{h=1}^{H} q_j(h|x) = 1$. If the $\delta_{j,h}$, $h = 1, \dots, H-1$, are ordered, category h is the most likely response if $\delta_{j,h-1} < x_i < \delta_{j,h}$.

There is no a priori justification for model (2); its usefulness must prove itself in application. Model (2) with all α_j identical (i.e., all appraisers have equal repeatability) is Masters's original Partial Credit Model, and it is a so-called Rasch model. The important property of Rasch models is known as *conjoint additivity* (Wright 1997), and its effect is that there are sufficient estimators for the threshold points δ_{jh} on the x-continuum, and for the true values x_i (whence they can be estimated independently). Masters (1982) derived the model by considering measurement on a polytomous scale as a sequence of dichotomous decisions. Thus, rating an object as "3" is considered as deciding positive, positive and negative on the sequence of dichotomies "2 instead of 1?," "3 instead of 2?," and "4 instead of 3?." For each of these dichotomies, Masters applies Rasch's (1960) model for dichotomous responses, which results in model (2). The main alternative to the Generalized Partial Credit Model is Samejima's (1969) Graded Response Model. This model is computationally more awkward, and it is less flexible (in the sense that in Samejima's model, for each h there is an interval on the x axis where the corresponding characteristic curve is larger than the remaining curves; our model avoids this assumption, thus accommodating a wider variety of R&R behavior). Given the widely accepted use of the Generalized Partial Credit Model in IRT, it is the most natural choice for our present purpose, pending a validation of its usefulness based on a large number of applications.

3. ESTIMATION AND MODEL DIAGNOSTICS

In order to fit his model to a set of data, Muraki (1992) makes the assumption that the δ_{jh}, $h = 1, \dots, H-1$, are equidistant (for all j). We avoid this assumption and fit the general model. The parameters of model (2) are estimated from the experimental data by means of the maximum likelihood (ML) method. To this end, it is more convenient to represent the experimental data $\{Y_{ijk}\}_{i=1,\dots,I;j=1,\dots,J;k=1,\dots,K}$ in the form of response patterns $\{R_{ijh}\}_{i=1,\dots,I;j=1,\dots,J;h=1,\dots,H}$, where $R_{ijh} = \{\#k|Y_{ijk} = h\}$. Conditional on x the $\mathbf{R}_{ij} = (R_{ij1}, \dots, R_{ijH})$ follow a multinomial distribution with parameters $q_j(1|x), \dots, q_j(H|x)$. Thus,

$$P(\mathbf{R}_{ij} = \mathbf{r}_{ij}|X_i = x) = \frac{K!}{\prod_{h=1}^{H} r_{ijh}!} \prod_{h=1}^{H} (q_j(h|x))^{r_{ijh}},$$

and the unconditional probability equals

$$P(\mathbf{R}_{ij} = \mathbf{r}_{ij}) = \int_{-\infty}^{\infty} \frac{K!}{\prod_{h=1}^{H} r_{ijh}!} \prod_{h=1}^{H} (q_j(h|x))^{r_{ijh}} \phi(x)\, dx,$$

where, as mentioned in the previous section, we take the X_i to be independently, standard normally distributed. The likelihood of the experimental outcome is

$$\begin{aligned} L = P(\mathbf{R} = \mathbf{r}) &= \prod_{i=1}^{I} P(\mathbf{R}_i = \mathbf{r}_i) \\ &= \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{j=1}^{J} P(\mathbf{R}_{ij} = \mathbf{r}_{ij}|X_i = x) f_X(x)\, dx \\ &= \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{K!}{\prod_{h=1}^{H} r_{ijh}!} \prod_{h=1}^{H} (q_j(h|x))^{r_{ijh}} \phi(x)\, dx, \end{aligned}$$

where we used the assumption of conditional independence introduced in the previous section.

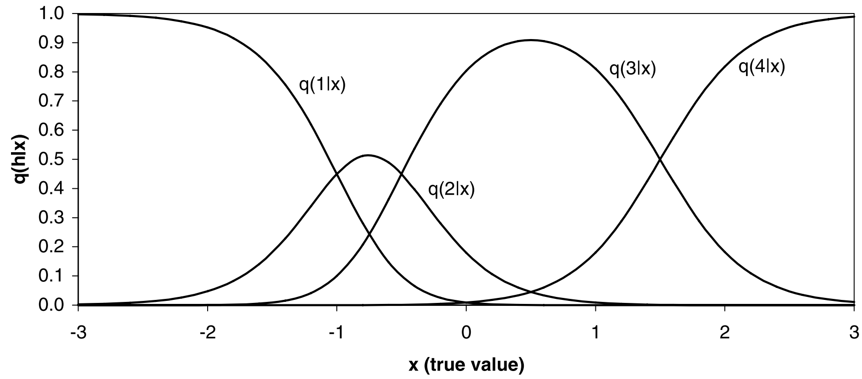

Figure 1. Characteristic curves; number of classes $H = 4$, $\alpha_j = 3$, $\boldsymbol{\delta}_j = \{-1.0, -0.5, 1.5\}$.

To maximize the likelihood with respect to the parameters we need to evaluate the integral in the likelihood. It is approximated by means of the Gauss–Hermite quadrature (Stroud and Secrest 1966):

$$L \simeq \prod_{i=1}^{I} \sum_{g=1}^{G} h(x_g) \prod_{j=1}^{J} \frac{K!}{\prod_{h=1}^{H} r_{ijh}!} \prod_{h=1}^{H} (q_j(h|x_g))^{r_{ijh}} =: \tilde{L},$$

where the x_g are the abscissas of the quadrature and $h(x_g)$ the corresponding weights. In effect, the latent variable is discretized having values $x_1, \ldots, x_G$ with probabilities $h(x_1), \ldots, h(x_G)$ which satisfy $\sum_{g=1}^{G} h(x_g) = 1$. The integral can be approximated to any desired degree of accuracy by increasing the number of quadrature points (we have used 35 points throughout our analyses).

Writing $\boldsymbol{\alpha}$ for $\{\alpha_j\}_{j=1,\ldots,J}$ and $\boldsymbol{\delta}$ for $\{\delta_{jh}\}_{j,h}$, we determine the ML estimates from

$$\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\delta}} \log(\tilde{L}).$$

The maximum of $\tilde{L}$ is found by application of the Newton–Raphson algorithm. In the algorithm we use the first and second order partial derivatives of the log-likelihood.

The estimation procedure outlined above is, when there is very little spread in the data, sensitive to the choice of the initial values used in the Newton–Raphson maximization. As a resolution, we propose an iterative maximum penalized likelihood procedure based on a sequence of penalty parameters $\lambda_0 > \lambda_1 > \cdots > \lambda_U = 0$. The idea is that for $u = \ldots, 3, 2, 1, 0$, the corresponding λ_u approach ∞, while for $u = \ldots, U-3, U-2, U-1$, the λ_u approach 0, with $\lambda_U = 0$. We work with the sequence $\lambda_u = (5^{U-u} - 1)/500$, with $U = 15$. For each iteration u the maximum penalized likelihood estimates are determined from

$$\hat{\boldsymbol{\alpha}}^u, \hat{\boldsymbol{\delta}}^u = \arg\max_{\boldsymbol{\alpha},\boldsymbol{\delta}} \log(\tilde{L}) - \lambda_u \sum_{j=1}^{J} (\log \alpha_j)^2. \tag{3}$$

For $u = 0$ the value λ_0 approaching ∞ ensures that $\hat{\alpha}_j^0 = 1$ for all j; in fact, we fit Masters's original Partial Credit Model. This estimation step appears to be insensitive to the choice of initial values for the Newton–Raphson procedure. Subsequent iterations ($u = 1, 2, \ldots, U$) consist of reapplying the Newton–Raphson procedure to solve (3), with $\hat{\boldsymbol{\alpha}}^{u-1}$ and $\hat{\boldsymbol{\delta}}^{u-1}$ as initial values. These iterations result in a sequence of estimates $(\hat{\boldsymbol{\alpha}}^u, \hat{\boldsymbol{\delta}}^u)_{u=0,1,\ldots,U}$. The final estimates are $\hat{\boldsymbol{\alpha}}^{u_0}$, $\hat{\boldsymbol{\delta}}^{u_0}$ with

$$u_0 = \arg\max_{u=0,\ldots,U} \log(\tilde{L}_{\hat{\boldsymbol{\alpha}}^u, \hat{\boldsymbol{\delta}}^u}),$$

that is, the maximum penalized likelihood estimate resulting in the highest unpenalized likelihood. Typically, $\lambda_{u_0} = 0$, but in situations where the R&R are near perfect we observed some nonzero values ($\lambda_{u_0} < 0.1$ in most cases). In these situations the differences between estimated parameters based on $\lambda = 0$ versus $\lambda_{u_0} > 0$ were very small.

Confidence intervals of the parameter estimates could be constructed by inverting the observed Fisher information matrix, which is the matrix of second-order partial derivatives of $\log(\tilde{L})$ evaluated at the ML estimates. This approach assumes that the confidence intervals are symmetric around the point estimates. This may be reasonable for large I, but not necessarily for the small sample situation under study. We bootstrap the confidence intervals, following De Menezes (1999), who uses resampling techniques for the construction of confidence intervals in the context of latent class models. New experimental data are generated by randomly drawing (with replacement) I samples (objects) from the original experimental data. The parameters are estimated for the new data. The process (resampling and estimation) is repeated a large number of times, say B times. The limits of the 95% confidence interval of the parameter estimates are then given by the 0.025 and 0.975 quantiles of each set of B estimated parameters (in the analyses later in this paper, we used $B = 1000$).

As for model diagnostics, we present a number of options. First we propose an approach for validating the normality assumption for the true values X_i. The true value X_i of each object i is predicted, given the response patterns $\mathbf{R}_i = (R_{i11}, \ldots, R_{ijh}, \ldots, R_{iJH})$, as

$$\begin{aligned}
\hat{x}_i &= E(X_i | \mathbf{R}_i = \mathbf{r}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}) \\
&= \int_{-\infty}^{\infty} x P(X_i = x | \mathbf{R}_i = \mathbf{r}_i; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}})\, dx \\
&= \int_{-\infty}^{\infty} x \frac{f_X(x) P(\mathbf{R}_i = \mathbf{r}_i | x; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}})}{\int_{-\infty}^{\infty} f_X(u) P(\mathbf{R}_i = \mathbf{r}_i | u; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}})\, du}\, dx \\
&= \left(\int_{-\infty}^{\infty} x \prod_{j=1}^{J} \frac{K!}{\prod_{h=1}^{H} r_{ijh}!} \prod_{h=1}^{H} (q_j(h|x; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}))^{r_{ijh}} \phi(x)\, dx \right) \\
&\qquad \Big/ \left(\int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{K!}{\prod_{h=1}^{H} r_{ijh}!} \prod_{h=1}^{H} (q_j(h|x; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\delta}}))^{r_{ijh}} \phi(x)\, dx \right),
\end{aligned} \tag{4}$$

which can be determined using the Gauss–Hermite quadrature. The validity of the normality assumption is assessed by plotting the predicted values $\hat{x}_i$ in a normal probability plot.

Second, residual analysis should be performed to check for observations that have disproportionate influence on the estimates. In latent variable modeling, it is customary to work with standardized or Freeman–Tukey variance stabilized residuals (Formann 2003). Both report response patterns whose observed frequency deviates substantially from their expected frequency given the fitted model. It turns out to be difficult to translate these flagged response patterns to interpretable indications about anomalous observations, as it is hard to relate them to specific objects or appraisers. Instead, we propose a method for reporting unusual observations on the level of objects. Analogous to common practice in fitting linear models, we define the results for an object i as "unusual" if its response pattern $\mathbf{R}_i$ is in the set of 5% least likely responses given the fitted model and predicted response. This mirrors the practice in normal regression analysis of labeling observations "unusual" if their absolute standardized residual is larger than 2.

Let $RP = \{0, \ldots, K\}^{JH}$ be the set of all possible response patterns, which contains $L = (K+1)^{JH}$ different patterns. For each potential response $\mathbf{rp}_\ell = (rp_{\ell,11}, \ldots, rp_{\ell,jh}, \ldots, rp_{\ell,JH}) \in RP$

we define the usualness, given the fitted model and predicted true value, for object i as

$$\begin{aligned} U_i(\mathbf{rp}_\ell) &= P_{\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\delta}}}(\mathbf{R}_i = \mathbf{rp}_\ell | X_i = \hat{x}_i) \\ &= \prod_{j=1}^{J} \frac{K!}{\prod_{h=1}^{H} rp_{\ell jh}!} \prod_{h=1}^{H} (\hat{q}_j(h|\hat{x}_i))^{rp_{\ell jh}}. \end{aligned}$$

Let, for part i, $\mathbf{rp}_{(\ell)}$ be the response patterns ordered by degree of usualness, that is, $U_i(\mathbf{rp}_{(1)}) > \cdots > U_i(\mathbf{rp}_{(L)})$. For object i, the collection RPU^i of 5% least likely response patterns consists of $\mathbf{rp}_{(\ell_{95})}, \mathbf{rp}_{(\ell_{95}+1)}, \ldots, \mathbf{rp}_{(L)}$, where ℓ_{95} is the smallest ℓ_0 for which $\sum_{\ell=1}^{\ell_0} U_i(\mathbf{rp}_{(\ell)}) > 0.95$. Object i is reported as "unusual" if $\mathbf{R}_i \in RPU^i$.

4. INTERPRETATION OF THE RESULTS

4.1 Intraappraiser Analysis

First we study how to assess the repeatability of a single appraiser j. In IRT the information function is frequently used. Each item in a test has an information function, which enables the test designer to distinguish good (i.e., discriminating) items from poor test items, and to select a set of items such that their combination is discriminating on the relevant range of the true value axis (see Lord 1980, p. 65). For our purpose, where we are not dealing with tests consisting of a set of items, the information function is of limited value. The information function does express for each true value x the repeatability of the measurements, but it does so in the form of an abstract value which is difficult to interpret.

Ordinal scales have two properties, namely, distinctiveness and order. We find it important that both properties are taken into account when evaluating the R&R. As explained in the Introduction, merely taking into account agreement on absolute values would effectively treat the data as nominal instead of ordinal. We propose two metrics, namely, the probabilities of correct ordering ρ, and of consistent classification π. The first one is a measure of concordance. Our model based approach allows us to go further than Goodman and Kruskal's gamma statistic, which is based on concordance between the observations of two appraisers. Our metric is based on concordance between observed order and true order.

We define the probability ρ_j^w (superscript "w" for *within*) of correct ordering of appraiser j as

$$\begin{aligned} \rho_j^w &:= P(Y_{ijk} \le Y_{ujk} | X_i \le X_u) \\ &= P(Y_{ijk} \le Y_{ujk}, X_i \le X_u)/P(X_i \le X_u) \\ &= 2\int_{x=-\infty}^{\infty} \int_{w=x}^{\infty} P(Y_{ijk} \le Y_{ujk} | X_i = x, X_u = w) \\ &\quad \times \phi(x)\phi(w)\,dw\,dx \\ &= 2\int_{x=-\infty}^{\infty} \int_{w=x}^{\infty} \sum_{h=1}^{H} \sum_{g=h}^{H} q_j(h|x) q_j(g|w) \\ &\quad \times \phi(x)\phi(w)\,dw\,dx. \qquad (5) \end{aligned}$$

If the discrimination parameter α_j approaches infinity (i.e., measurements approaching perfectly consistent ratings), $q_j(h|x)$ converges to 1 (uniformly for $\delta_{j,h-1} < x < \delta_{j,h}$; see the Appendix, available as online supplemental material), and $\lim_{\alpha_j \to \infty} \rho_j^w = 1$. The Appendix (online supplemental material) also shows that for $\alpha_j \downarrow 0$ (i.e., measurements approaching random ratings), $q_j(h|x)$ converges uniformly to $1/H$, and $\lim_{\alpha_j \downarrow 0} \rho_j^w = (H+1)/2H =: \rho_0$ (which is approximately 1/2 for larger H). Note that the lower bound ρ_0 depends on the number of classes of the ordinal scale. If one prefers a repeatability index whose values are interpretable independent of the number of classes H, one could work with $\tilde{\rho}_j^w = (\rho_j^w - \rho_0)/(1 - \rho_0)$, which is an index that is similar in form to the κ (kappa) index for nominal measurements (de Mast and van Wieringen 2007). The extremes $\tilde{\rho}_j^w = 0$ and $\tilde{\rho}_j^w = 1$ represent the situations of purely random and perfectly repeatable classifications. The disadvantage of this index is that it is more abstract.

The second metric is the probability of consistent classification. A classification Y_{ijk} by an appraiser j is consistent if it agrees with his own category bounds δ_{jh}. The probability of consistent classification for appraiser j and class h is

$$\begin{aligned} \pi_j^w(h) &= P(Y_{ijk} = h | \delta_{j,h-1} < X_i < \delta_{j,h}) \\ &= \frac{\int_{x=\delta_{j,h-1}}^{\delta_{j,h}} q_j(h|x)\phi(x)\,dx}{\int_{x=\delta_{j,h-1}}^{\delta_{j,h}} \phi(x)\,dx}. \end{aligned}$$

The limit behavior of $\pi_j^w(h)$ is similar to that of ρ_j^w. Approaching random ratings, we have $\lim_{\alpha \downarrow 0} \pi_j^w(h) = 1/H := \pi_0$, while $\lim_{\alpha \to \infty} \pi_j^w(h) = 1$. The probability of correct classification for appraiser j is

$$\begin{aligned} \pi_j^w &= \sum_{h=1}^{H} \pi_j^w(h) P(\delta_{j,h-1} < X_i < \delta_{j,h}) \\ &= \int_{x=-\infty}^{\infty} q_j^*(x)\phi(x)\,dx, \qquad (6) \end{aligned}$$

with $q_j^*(x) = \sum_{h=1}^{H} \mathbf{1}_{\{\delta_{j,h-1} < x \le \delta_{j,h}\}} q_j(h|x)$. Also π_j^w can be rescaled to the $[0, 1]$ interval, and we define $\tilde{\pi}_j^w = (\pi_j^w - \pi_0)/(1 - \pi_0)$.

Besides repeatability (expressed as a probability of correct ordering or consistent classification) it is important to verify for each appraiser whether the used ordinal scale is valid. The ordinal scale implies that there is a particular order in which the classes $1, \ldots, H$ are intended. There is a validity problem if the data show that there are appraisers who do not apply the classes in their intended order. This becomes apparent if the thresholds $\delta_{j,1}, \ldots, \delta_{j,H-1}$ of an appraiser j are not ordered; let us say $\delta_{j,2} < \delta_{j,1}$ (see, e.g., Figure 2, in which $\delta_{j,1} = -0.5$, while $\delta_{j,2} = -1.5$). Note that even in such case the intended order of the classes is preserved in the sense that the log-odds of choosing class $h + 1$ over class h,

$$\log \frac{q_j(h+1|x)}{q_j(h|x)} = \alpha_j(x - \delta_{jh}),$$

is an increasing function in x (confirming that lower classes correspond to lower true values x, and higher classes to higher x values). But a more strict sense of order is violated in this case. For x values smaller than $\delta_{j,2}$ we have $P(Y_{ijk} = 1|x) > P(Y_{ijk} = 2|x) > P(Y_{ijk} = 3|x)$, and for x values greater than $\delta_{j,1}$

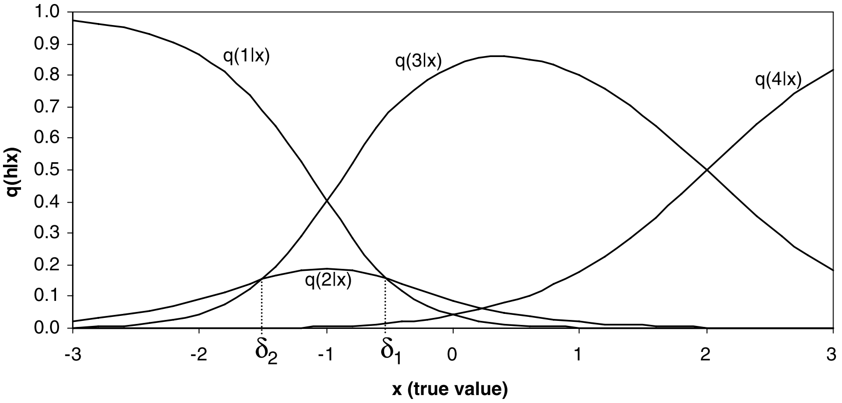

Figure 2. Characteristic curves; $H = 4$, $\alpha_j = 1.5$, $\boldsymbol{\delta}_j = \{-0.5, -1.5, 2.0\}$.

we have $P(Y_{ijk} = 1|x) < P(Y_{ijk} = 2|x) < P(Y_{ijk} = 3|x)$ (which implies the correct order), but for x values in between $\delta_{j,2}$ and $\delta_{j,1}$ we have $P(Y_{ijk} = 1|x) > P(Y_{ijk} = 3|x) > P(Y_{ijk} = 2|x)$ or $P(Y_{ijk} = 3|x) > P(Y_{ijk} = 1|x) > P(Y_{ijk} = 2|x)$, which implies a conflict with the intended order. Note also that the curve corresponding to $h = 2$ is dominated for all x values by other curves, implying that there exist no true values x for which $h = 2$ is the most probable response—put differently, the appraiser is reluctant to use this class.

If a particular curve is dominated by the other curves (such as $h = 2$ is dominated by the other curves in Figure 2) for all appraisers, it is probably best to drop that class from the scale (that is, the 4-point scale in Figure 2 is replaced with the 3-point scale $\{1, 3, 4\}$).

4.2 Interappraiser Analysis

Differences among appraisers can pertain to the discrimination parameters α_j (i.e., appraisers classify with different repeatability) and to the threshold parameters δ_{jh} (i.e., appraisers act to different boundaries between the categories $h = 1, \ldots, H$). For the first, it is useful to give a table of the repeatability ρ_j^w and π_j^w per appraiser, and the mean repeatability $\bar{\rho}^w$ and $\bar{\pi}^w$. Especially differences among the δ_{jh} are understood as an issue of reproducibility. Differences among each appraiser's δs could be differences with a fixed offset:

$$\delta_{jh} = \bar{\delta}_{\cdot h} + \tau_j \quad \text{with } \bar{\delta}_{\cdot h} = \frac{1}{J}\sum_{j=1}^{J} \delta_{jh}.$$

The boundaries of appraiser j are all shifted by an amount τ_j on the x continuum compared to the average over all appraisers. In this case, we have a simple calibration problem. If the differences in each appraiser's δs have a more complex structure, this is indicative for unclear definitions of the scale's classes $h = 1, \ldots, H$ (and especially their boundaries). In both cases, we think the user is best helped by a table or plot displaying how the appraisers' δs compare among each other.

For comparing the relative contributions of intraappraiser and interappraiser inconsistencies, the intraappraiser probabilities ρ_j^w of correct ordering could be compared to their interappraiser variant ρ^b ("b" for *between*). We define the pair-wise probability that observations from different appraisers result in correct ordering as

$$\rho^b_{j_1,j_2} = 2\int_{x=-\infty}^{\infty}\int_{w=x}^{\infty}\sum_{h=1}^{H}\sum_{g=h}^{H} q_{j_1}(h|x)q_{j_2}(g|w) \times \phi(x)\phi(w)\,dw\,dx. \quad (7)$$

The total interappraiser probability of correct ordering is

$$\rho^b = \frac{1}{J(J-1)}\sum_{j_1 \neq j_2 = 1}^{J} \rho^b_{j_1,j_2},$$

and the rescaled version is $\tilde{\rho}^b = (\rho^b - \rho_0)/(1 - \rho_0)$.

Where the intraappraiser probability of consistent classification π_j^w reflects the degree to which an appraiser's classifications are consistent with his own category bounds, the interappraiser version gives the probability that, given an object with true value X, the category boundaries of two randomly selected appraisers are consistent (in the sense that both sets of category boundaries classify the object in the same category). In symbols:

$$\begin{aligned}\pi^b_{j_1,j_2} &= \sum_{h=1}^{H} P\big(\delta_{j_1,h-1} < X < \delta_{j_1,h} \wedge \delta_{j_2,h-1} < X < \delta_{j_2,h}\big) \\ &= \sum_{h=1}^{H} \max\big\{0, \Phi\big(\min\{\delta_{j_1,h}, \delta_{j_2,h}\}\big) \\ &\qquad - \Phi\big(\max\{\delta_{j_1,h-1}, \delta_{j_2,h-1}\}\big)\big\} \qquad (8)\end{aligned}$$

and

$$\pi^b = \frac{1}{J(J-1)}\sum_{j_1 \neq j_2 = 1}^{J} \pi^b_{j_1,j_2}.$$

5. PUTTING THE PROPOSED APPROACH TO THE TEST: MEASURING SOLDERED JOINTS QUALITY

We illustrate our approach with a real-life example, and compare our analysis with alternative indices proposed in literature. A project at an electronics manufacturer aimed at redesigning the process for soldering printed circuit boards (PCBs).

The project's objective was to deliver a lead-free soldering process yielding an acceptable quality of soldered joints. Soldered joints quality was judged by means of a visual inspection, for which brief guidelines were given. Quality was rated on a 4-point scale (1 = "reject," 2 = "critical," 3 = "acceptable," 4 = "good").

To establish the repeatability and reproducibility of the visual inspections, the project leader set up an experiment, in which 45 PCBs were inspected by three operators, and three weeks later the same operators rated the same PCBs a second time (the "Initial" experiment). Table 1 gives the results. Not conforming to good practice in experimental design, the PCBs were not presented in randomized order, but in the order given in the table. Note that the table suggest that the project leader roughly sorted the PCBs by quality. They are a random sample, though, and therefore can be considered representative for the quality of the soldering process at that time; bear in mind that this concerns a soldering process under development, whence the frequency of rejects is rather large.

Table 1. Results of the Initial (left) and Follow-up experiment studying the R&R of quality ratings for soldered joints

	Appraiser						Appraiser					
Part	A		B		C		A		B		C	
1	1	1	1	1	2	2	3	3	2	3	3	3
2	1	1	1	1	2	2	4	4	4	4	4	4
3	2	1	2	1	2	2	2	2	2	2	2	2
4	1	1	1	1	2	2	2	2	2	2	2	2
5	2	1	1	1	2	2	4	4	4	4	4	4
6	2	1	2	2	2	2	1	1	1	1	1	1
7	1	1	1	1	2	2	2	2	3	3	2	2
8	2	1	1	1	2	2	3	3	3	3	3	4
9	2	2	1	1	2	2	3	3	3	3	3	3
10	1	1	1	1	3	2	2	2	2	3	2	2
11	3	2	2	2	3	3	2	3	3	4	3	3
12	3	2	2	2	3	2	3	3	3	3	3	3
13	3	2	2	2	3	2	1	1	1	1	1	1
14	3	2	2	1	3	2	2	3	3	3	2	3
15	3	3	2	3	3	2	2	2	2	2	2	2
16	3	3	3	2	3	2	4	4	4	4	4	4
17	3	3	3	3	3	2	4	4	4	4	4	4
18	3	3	3	3	3	3	4	4	4	4	4	4
19	3	3	3	3	3	3	3	3	3	3	3	3
20	4	3	3	2	4	3	3	3	3	3	3	3
21	3	3	3	3	3	3	3	4	4	4	4	4
22	3	3	1	1	3	3	1	1	1	1	1	1
23	3	3	3	2	3	3	4	4	4	4	4	4
24	2	3	3	2	3	2	4	4	4	4	4	4
25	3	3	1	1	3	2	2	2	2	2	2	2
26	3	3	3	2	3	3	2	2	2	2	2	2
27	3	3	3	2	4	3	2	2	3	3	3	3
28	2	3	1	1	3	2	2	2	2	2	2	2
29	2	3	2	2	3	3	3	3	3	3	3	3
30	2	3	1	2	3	3	2	2	3	3	2	2
31	3	4	3	2	4	4						
32	3	4	3	2	3	3						
33	3	4	2	2	4	4						
34	3	3	2	2	4	3						
35	3	3	2	3	3	3						
36	2	3	2	3	3	3						
37	3	3	3	3	3	3						
38	3	3	3	3	3	3						
39	3	4	3	3	3	3						
40	4	4	3	3	4	4						
41	4	3	1	1	4	3						
42	4	3	1	1	4	4						
43	4	3	1	3	3	4						
44	3	3	4	4	4	3						
45	3	3	3	3	3	3						

The project leader did a fairly elementary analysis, just counting the number of PCBs for which all six ratings agreed. The very low number (6 out of 45 PCBs) made her conclude that she had a serious problem with this inspection procedure. She discussed some of the PCBs in the sample with the operators, thus establishing clearer inspection guidelines. She also made photos showing "border cases," that is, photos which define the border between the "good" and "acceptable," "acceptable" and "critical," and "critical" and "reject" categories. To confirm the effectiveness of the new inspection guidelines, a new experiment was set up (the "Follow-up" experiment), involving the same three operators but 30 new PCBs (results given in Table 1). The improved results (21 out of 30 PCBs with full agreement) led her to accept the new inspection procedure.

5.1 IRT Analysis

Tables 2 and 3 and Figures 3 and 4 present the analysis as proposed in this paper, executed in the R environment (R Development Core Team 2008). For the Initial experiment, the results imply that the repeatability of the first and third operator is fair, while the second operator's repeatability seems poor. The per appraiser probabilities of consistent classification [calculated from (6)] are $\pi_j^w = 0.721, 0.540, 0.755$. Note that the probability of consistent classification would be $\pi_0 = 0.250$ for purely random ratings, which should be taken as an offset for the estimated probabilities. The probabilities of correct ordering are $\rho_j^w = 0.951, 0.846$, and 0.952 (with $\rho_0 = 0.625$).

Reproducibility is poor (the probability of correct ordering is $\rho^b = 0.864$). Figure 3 tells part of the story of what goes wrong in the classification procedure: the third operator avoids the category "1," while the second operator underuses the category "4." The category boundaries used by operator 2 are very different from the boundaries used by the other two. The inconsistent category boundaries are reflected in the interappraiser probability of consistent category boundaries, which is $\pi^b = 0.498$. The pairwise probabilities of consistent category boundaries [calculated from (8)] are $\pi_{1,2}^b = 0.421$, $\pi_{1,3}^b = 0.825$, $\pi_{2,3}^b = 0.247$, demonstrating that appraisers 1 and 3 are highly consistent, but appraiser 2 has deviating category boundaries.

As for the results of the Follow-up experiment, we see that matters have improved substantially, and all repeatabilities are quite good now (probabilities of correct ordering are $\rho_j^w = 0.989, 0.978$, and 0.993; probabilities of consistent classification are $\pi_j^w = 0.919, 0.830, 0.957$). Also the interrater consistency has greatly improved, as can be seen from Figure 4 and from the estimated probability of correct ordering $\rho^b = 0.980$ and the interappraiser probability of consistent category boundaries $\pi^b = 0.795$. Note that the relatively wide confidence intervals in Tables 2 and 3 indicate that the relatively small sample sizes somewhat limit the strength of our analyses.

Table 2. Estimated model parameters for the Initial experiment, including probabilities of correct ordering and consistent classification (with 95% confidence intervals)

j	α_j	$\delta_{j,1}$	$\delta_{j,2}$	$\delta_{j,3}$	ρ_j^w	95% C.I.	π_j^w	95% C.I.
1	3.2	−1.1	−0.5	1.3	0.951	(0.917, 0.977)	0.721	(0.665, 0.816)
2	1.0	−0.3	0.3	3.5	0.846	(0.747, 0.945)	0.540	(0.428, 0.743)
3	3.2	−22.3	−0.5	1.1	0.952	(0.893, 0.982)	0.755	(0.660, 0.876)
					ρ_0:	0.625	π_0:	0.250
Reproducibility ρ^b:					0.864	(0.813, 0.903)	π^b:	0.498

Table 3. Estimated model parameters for the Follow-up experiment, including probabilities of correct ordering and consistent classification (with 95% confidence intervals)

j	α_j	$\delta_{j,1}$	$\delta_{j,2}$	$\delta_{j,3}$	ρ_j^w	95% C.I.	π_j^w	95% C.I.
1	20.9	−1.8	−0.0	1.0	0.989	(0.974, 0.999)	0.919	(0.832, 0.996)
2	6.6	−1.8	−0.5	0.7	0.978	(0.952, 1.00)	0.830	(0.743, 0.997)
3	21.4	−1.8	−0.1	0.6	0.993	(0.984, 1.00)	0.957	(0.856, 0.998)
					ρ_0:	0.625	π_0:	0.250
Reproducibility ρ^b:					0.980	(0.951, 0.992)	π^b:	0.795

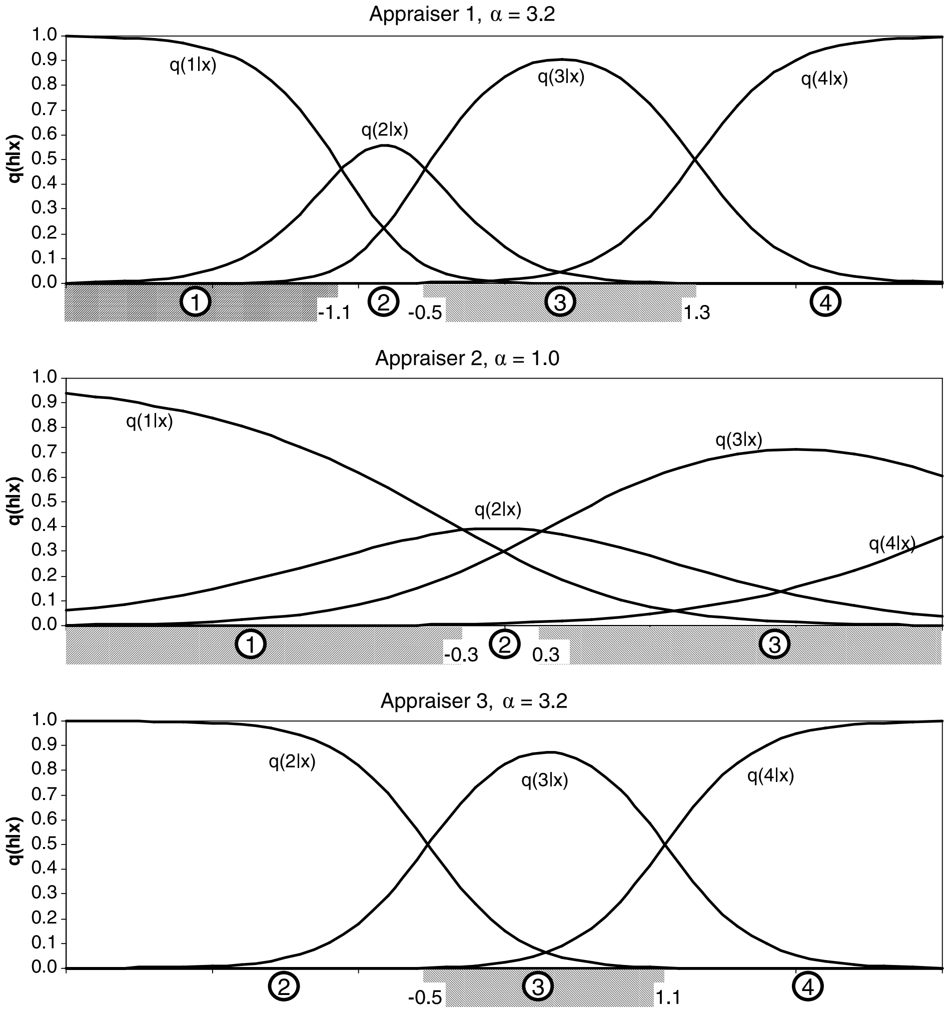

Figure 3. Results of the analysis of the Initial experiment. Characteristic curves are shown for all three appraisers. The δ_{jh} are indicated on the x-axis. By means of shading the areas $\delta_{j,h-1} < x < \delta_{jh}$ (for $h = 1, \ldots, 4$) are demarcated.

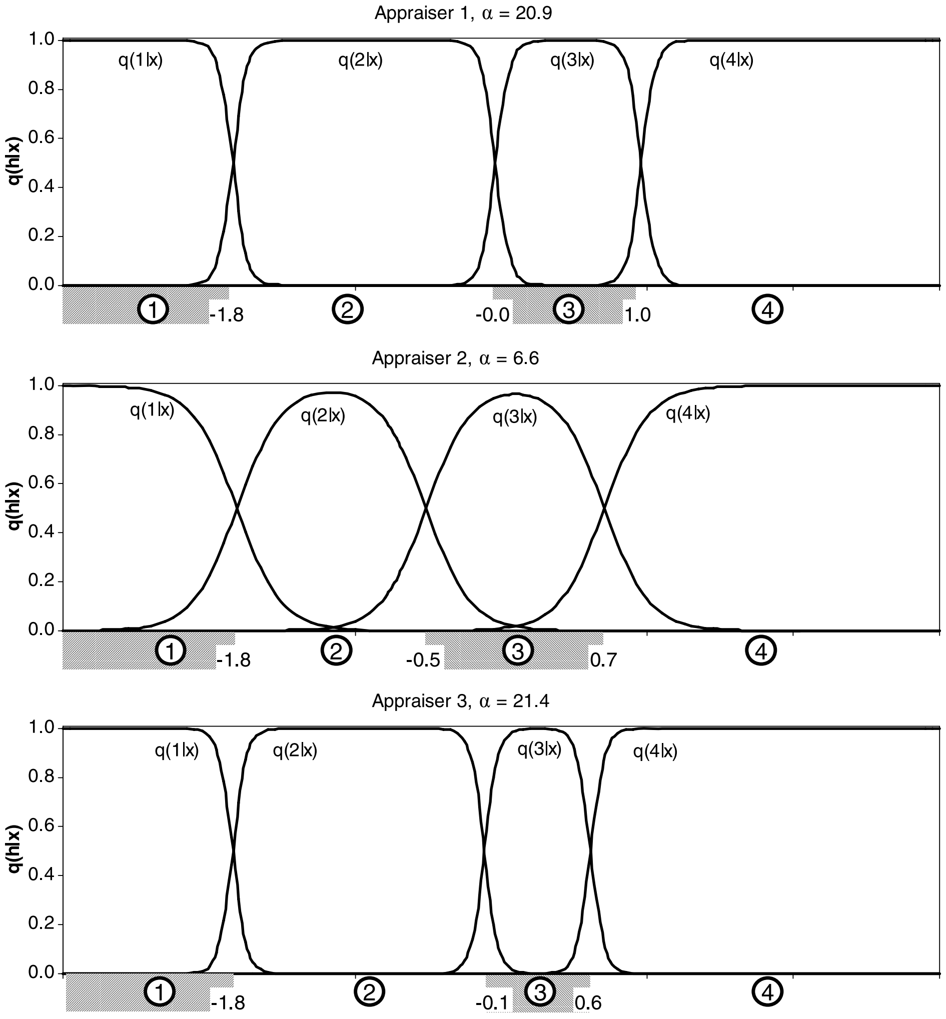

Figure 4. Results of the analysis of the Follow-up experiment. Characteristic curves are shown for all three appraisers. The δ_{jh} are indicated on the x-axis. By means of shading the areas $\delta_{j,h-1} < x < \delta_{jh}$ (for $h = 1, \ldots, 4$) are demarcated.

Figure 5 shows the normal probability plots of the fitted true values $\hat{x}_i$ [which were calculated from (4)]. The stacks of datapoints (e.g., at $x = 1.47$ in the Follow-up experiment) need not worry us, as they are an artifact of the fact that the x_i are reconstructed from categorical data: if $\mathbf{r}_{i_1} = \mathbf{r}_{i_2}$ then $\hat{x}_{i_1} = \hat{x}_{i_2}$. The probability plots do not give us strong evidence that the normality assumption for the X_i is unwarranted.

Three response patterns in the Initial experiment are flagged as "unusual" at the 95% level, namely, the results for PCBs 41, 42, and 44. Looking in the raw data, we see that PCBs 41 and 42 were rated "3" or "4" by appraisers 1 and 3. As can be seen in Figure 3, the typical response from the second appraiser in such cases would be "3," perhaps a "2," but the double "1" is anomalous. Similarly for PCB 44, where the double "4" from the second appraiser is at odds with his typical rating behavior. The regular analysis [based on model (2)] allows for fixed differences among the appraisers in their implicit category bounds (the δ_{jh}); it does not allow for object × appraiser interaction effects (namely, that an appraiser's δ_{jh} are different from object to object). One way in which the model's assumptions can be violated, is that such interaction effects are present. A possible explanation says that the deviations are caused by specific properties of the three PCBs in question; appraiser 2 responds differently to these specific properties than appraisers 1 and 3.

Figure 6 is an interaction plot. It compares whether the analysis results are in line with the results of J individual per-appraiser analyses. In particular, fitting the model to the data of appraiser j only, instead of to the complete dataset, yields fitted values $\hat{x}_i^{(j)}$, where the (j) superscript indicates that it concerns true values predicted from the data from appraiser j only. Figure 6 graphs these $\hat{x}_i^{(j)}$ (with the objects sorted by $\hat{x}_i$). The graphs visualize the differences between the appraisers, with parallel but shifted graphs indicating fixed differences. The discrepancies in Figure 6 between appraiser 2 on the one hand, and the other two appraisers on the other, are not only fixed discrepancies, but differ from object to object, thus visualizing the object × appraiser interaction effects. Note that the largest discrepancies are for PCBs 41, 42, and 44, which are the same ones that we had identified as "unusual observations" above. Thus, the graph reveals a serious problem with this rating process. The analysis provides the user with specific leads as to how to pinpoint the core problem; a sensible next step would be to examine and discuss PCBs 41, 42, and 44 with the three appraisers.

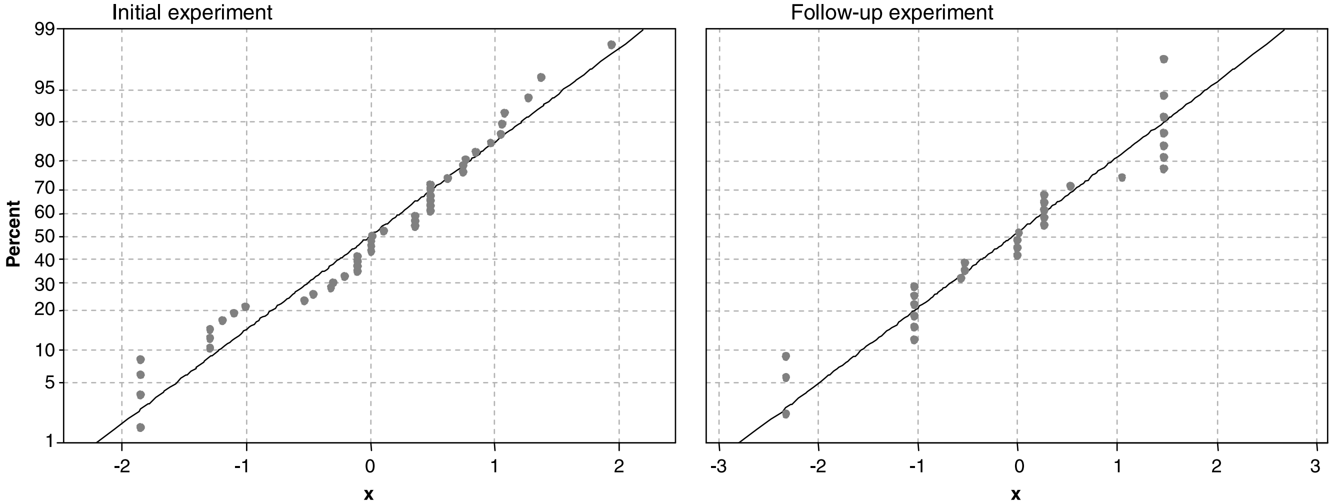

Figure 5. Normal probability plots of the fitted true values $\hat{x}_i$ for the Initial (left) and Follow-up experiment (right).

Note that, in the original analysis, α_2 was estimated substantially lower than α_1 and α_3, suggesting that the repeatability of the second appraiser is substantially worse. At first sight, this seems at odds with the observation that the results of the second appraiser disagree less than those of the other appraisers (15 times out of 45, versus 21 and 16 for the other appraisers). When fitting the model to the data of the second appraiser only, α_2 is estimated as 3.79, which seems better in line with appraiser 2's repeatability. The relatively low estimate in the original analysis is explained by the observation that the abovementioned object × appraiser interaction effect is absorbed in the intraappraiser results of the second appraiser, resulting in a poorer value for α_2.

Figure 7 shows that the Follow-up experiment does not seem to suffer from problems like the Initial experiment. One response pattern is flagged as "unusual," namely, the one corresponding to object number 11. We do not have an explanation for this unusual observation, but the reader should bear in mind that even in datasets that conform perfectly to the model assumptions, some 5% of the observations will be flagged as "unusual."

We summarize our conclusions. There is a general discrepancy between the second appraiser and the other two. The out-of-control behavior visualized in Figure 6, most prominent in the three flagged PCBs, is indicative of ineffective instructions or a general lack of understanding of the property that is being measured. The wild behavior frustrates modeling attempts, and the fitted model and estimated probabilities of correct ordering and consistent classification are unreliable in the Initial experiment, but the analysis pinpoints the source of the problem and

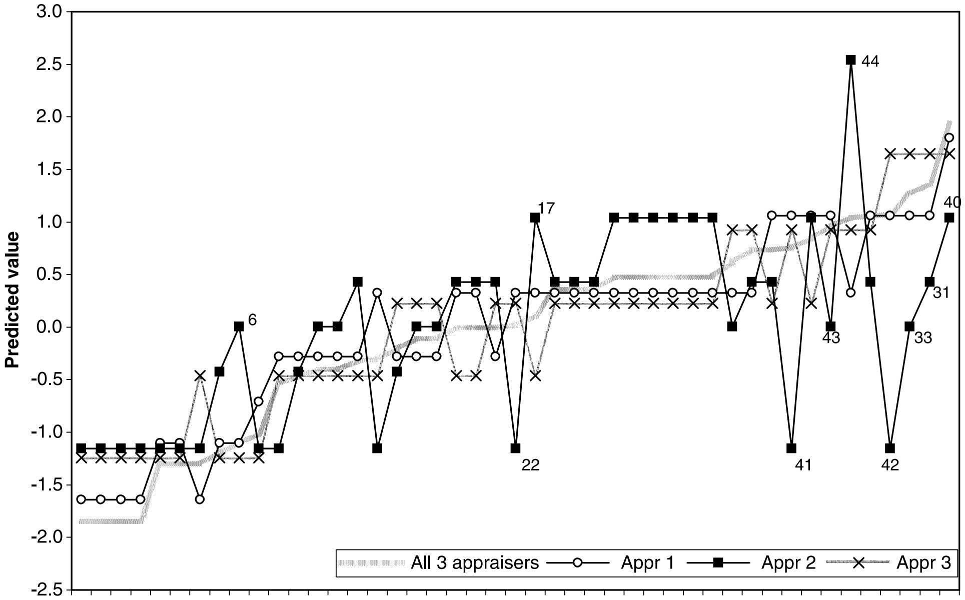

Figure 6. Comparison for the Initial experiment of the true values fitted from each appraiser's data individually (the $\hat{x}_i^{(j)}$), and the true values fitted from the total dataset (the $\hat{x}_i$). The points are sorted by $\hat{x}_i$ (objects i with smallest $\hat{x}_i$ to the left). Labels give the numbers of the 10 objects having the largest difference $\hat{x}_i^{(j)} - \hat{x}_i$.

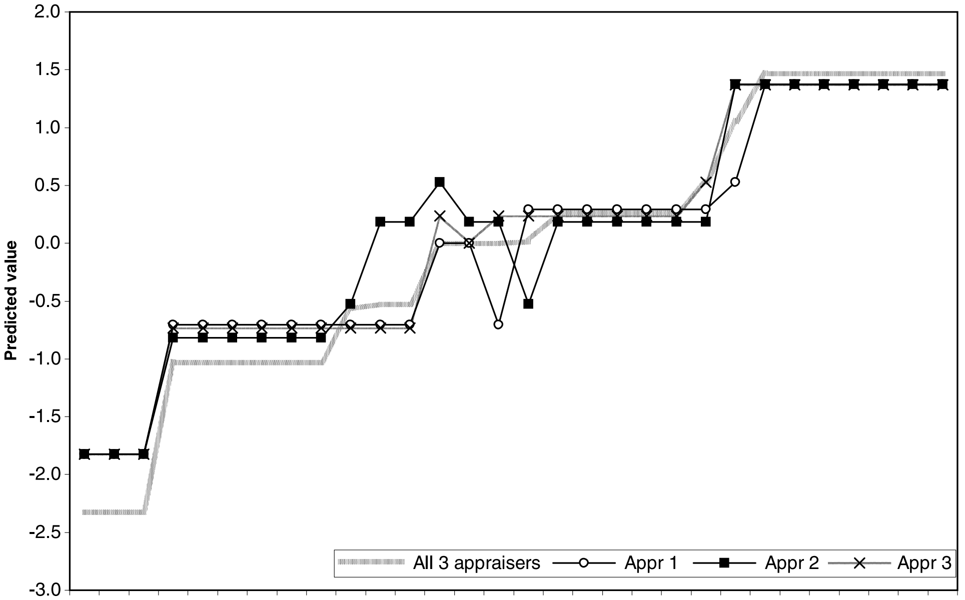

Figure 7. Comparison for the Follow-up experiment of the true values fitted from each appraiser's data individually, and the true values fitted from the total dataset. The points are sorted by $\hat{x}_i$ (objects i with smallest $\hat{x}_i$ to the left).

gives tangible and useful leads for improvement. In the analysis of the Follow-up experiment we do not see serious validity problems, and we conclude that this analysis gives a reliable assessment of the R&R of the final inspection procedure.

5.2 Nonparametric Analysis

Our analysis approach was designed with the pursuit of tangible results in mind. It is not the purpose of this paper to compare our methods with all alternatives found in literature (the interested reader is referred to de Mast and van Wieringen 2004, who compare a fair number of indices). We draw two of the most popular indices in the comparison, Kendall's coefficient of concordance W and Goodman and Kruskal's γ, thus illustrating our general concerns with many of the measures around.

The structure assumed in this paper (replications nested in appraisers) is rarely encountered in the contexts in which W and γ are usually defined. Instead, the data Y_{ij} are usually assumed to be repeated measurements $j = 1, \ldots, J$ of objects $i = 1, \ldots, I$ (in the datasets in Table 1, $I = 45$ and $I = 30$ respectively, while $J = 6$).

Kendall's W (briefly introduced in the Introduction) was originally proposed for rankings, but can be modified for use with (ordinal) ratings (de Mast and van Wieringen 2004). In this form, it is defined as

$$W = \frac{\sum_{i=1}^{I}(\sum_{j=1}^{J} R_{ij} - J(I+1)/2)^2}{IJ^2(I^2-1)/12 - J\sum_{j=1}^{J}\sum_{h=1}^{H} N_{jh}(N_{jh}^2-1)/12},$$

where $N_{jh} = \{\#i : Y_{ij} = h\}$ (for $j = 1, \ldots, J$ and $h = 1, \ldots, H$), and $R_{ij} = \sum_{h=1}^{Y_{ij}-1} N_{jh} + (1 + N_{jY_{ij}})/2$. For the Initial experiment ($I = 45$) we find $W = 0.639$ (computed for all six columns). By applying the formula to pairs of columns, we can calculate a W-value for each appraiser. For appraisers 1, 2, and 3 we find 0.817, 0.866, and 0.846 respectively (thus representing an intraappraiser analysis). The results for the Follow-up experiment ($I = 30$) are $W = 0.935$ (computed for all six columns), and 0.973, 0.971, and 0.982 for the individual appraisers.

Defining the probabilities of concordance, discordance, and ties as

$$\begin{aligned} P_c &:= P(Y_{i_1j_1} < Y_{i_2j_1}, Y_{i_1j_2} < Y_{i_2j_2}) \\ &\quad + P(Y_{i_1j_1} > Y_{i_2j_1}, Y_{i_1j_2} > Y_{i_2j_2}), \\ P_d &:= P(Y_{i_1j_1} < Y_{i_2j_1}, Y_{i_1j_2} > Y_{i_2j_2}) \\ &\quad + P(Y_{i_1j_1} > Y_{i_2j_1}, Y_{i_1j_2} < Y_{i_2j_2}), \\ P_{tie} &:= P(Y_{i_1j_1} = Y_{i_2j_1}) + P(Y_{i_1j_2} = Y_{i_2j_2}), \end{aligned}$$

Goodman and Kruskal's gamma is defined as

$$\begin{aligned} \gamma &= \frac{P_c - P_d}{1 - P_{tie}} \\ &= P(\text{concordance} \mid \text{no ties}) - P(\text{discordance} \mid \text{no ties}). \end{aligned}$$

A value of $\gamma = 1$ implies perfect consistency in order, whereas a value of $\gamma = 0$ means that ratings are done at random (and hence are uninformative). Kendall's tau (Kendall and Gibbons 1990) follows quite a similar line of reasoning. Gamma is defined for pairs of ratings (i.e., for data of the form Y_{ij}, where $i = 1, \ldots, I$ and $j = 1, 2$). Following the formulas in Goodman and Kruskal (1954) (or Haberman 1988), gamma is estimated from data Y_{ij} as follows

$$\hat{C} = 2\frac{1}{I^2}\sum_{h=1}^{H-1}\sum_{m=1}^{H-1}\sum_{h'=h+1}^{H}\sum_{m'=m+1}^{H} N_{hm}N_{h'm'},$$

$$\hat{D} = 2\frac{1}{I^2}\sum_{h=1}^{H-1}\sum_{m=2}^{H}\sum_{h'=h+1}^{H}\sum_{m'=1}^{m-1} N_{hm}N_{h'm'},$$

$$\hat{\gamma} = (\hat{C} - \hat{D})/(\hat{C} + \hat{D}).$$

Here, $N_{hm} = \{\#i : Y_{i1} = h, Y_{i2} = m\}$. Using this formula, we can calculate a $\hat{\gamma}(j_1, j_2)$ for each pair of columns $(j_1, j_2 = 1, \ldots, 6)$. The results for the column pairs corresponding to the same appraiser are $\hat{\gamma}(1, 2) = 0.830$, $\hat{\gamma}(3, 4) = 0.843$, and $\hat{\gamma}(5, 6) = 0.975$ in the Initial experiment, and 1.00, 1.00, and 1.00 in the Follow-up experiment. These values could be taken to represent an intraappraiser analysis. Taking the average of the $\hat{\gamma}$-values of the remaining columns, we find 0.707 (Initial) and 0.987 (Follow-up), which we take as the interappraiser analysis.

6. EVALUATION AND CONCLUSIONS

The purpose of this paper is to evaluate whether IRT modelling could be the basis for a methodology for evaluating the R&R of ordinal classifications in business and industry. We have demonstrated that IRT methodology can be modified for use with the sort of R&R studies that are common in business and industry.

The main obstacle that we are aware of, is that it is not clear whether IRT models can accommodate objects × appraiser interaction effects. The analysis may become intractable, and the inclusion of such an interaction effect runs somewhat against the philosophy of IRT (where the presence of such interaction effects would disqualify the rating process as a measurement process, as the criteria are unstable). The issue is coped with in our current approach by alerting the user to the problem (this is done by reporting unusual observations and by interaction plots such as Figures 6 and 7).

Comparing the analysis based on our IRT model to the analyses based on the W and γ statistics, we note in the first place that abstract values such as $W = 0.639$ and $\gamma = 0.707$ are difficult to interpret in tangible terms; they only say that the ratings are somewhere in between perfectly repeatable and purely random. We claim that the metrics that we propose (probabilities of correct ordering and consistent classification) have a more tangible interpretation than both W and γ.

The second point we wish to make concerns the case that a classification system has a poor performance, such as the Initial experiment's measurements in the example. Our model-based approach allows graphics and diagnostics that provide insight into the workings of a classification process, which is vital information for fixing an unreliable classification system. Besides the R&R metrics ρ and π themselves, the user is presented with:

- Predicted $\hat{x}_i$ and a normal probability plot of them; this helps identifying objects with anomalous values.
- Unusual observations. The power of this diagnostic has been demonstrated in the example, where it helped to pinpoint the problems in the Initial experiment's results (in combination with the interaction plot in Figure 6).
- The order of the estimated $\hat{\delta}_{jh}$ helps to check whether the categories are used in their intended order (see Figure 2, where this is not the case).
- Reproducibility modeling. The estimated δ_{jh} allow a detailed analysis of the nature of the differences among the appraisers (and this can also be done graphically, as in Figures 3 and 4).

Although both W and γ quantify repeatability and reproducibility, they do not provide insight into the structure and nature of intraappraiser and interappraiser differences, and we are not sure how to do diagnostics checking. Unusual observations are not brought to our attention, and as a consequence, the discrepancies between appraiser 2 and the other appraisers go unnoticed, and the salient results for PCBs 41, 42, and 44 are not revealed.

Future research should develop the methodology further by applying it to real cases such as the one above. The critical examination of how well our methods work in practice helps to identify points for improvement. Additional research should also result in recommendations for suitable sample sizes for this type of studies. The most interesting challenge, however, is to go beyond IRT modeling and try to develop models which incorporate objects × appraiser interaction effects and yet are tractable. In conclusion, then, we would say that IRT modeling did, for now, not result in a perfect approach, but has substantial merits above existing nonparametric approaches.

SUPPLEMENTAL MATERIALS

Dataset "Initial Experiment": First dataset of the soldered joints quality example (the initial experiment). (Initial-experiment.txt, tab delimited text file)

Dataset "Follow-up Experiment": Second dataset of the soldered joints quality example (the follow-up experiment). (Followup-experiment.txt, tab delimited text file)

Appendix: Appendix containing the mathematical derivations of the limit behavior of $q_j(h|x)$ and ρ_j^w for $\alpha \downarrow 0$ and $\alpha \to \infty$. (appendix.pdf, Acrobat reader document)

[Received March 2008. Revised March 2009.]

REFERENCES

Agresti, A. (1988), "A Model for Agreement Between Ratings on an Ordinal Scale," *Biometrics*, 44, 539–548. [95]

Burdick, R. K., Borror, C. M., and Montgomery, D. C. (2003), "A Review of Methods for Measurement Systems Capability Analysis," *Journal of Quality Technology*, 35, 342–354. [94]

Cohen, J. (1968), "Weighted Kappa: Nominal Scale Agreement With Provision for Scaled Disagreement or Partial Credit," *Psychological Bulletin*, 70, 213–220. [94]

de Mast, J. (2007), "Agreement and Kappa Type Indices," *The American Statistician*, 61, 148–153. [94]

de Mast, J., and van Wieringen, W. N. (2004), "Measurement System Analysis for Bounded Ordinal Data," *Quality and Reliability Engineering International*, 20, 383–395. [104]

——— (2007), "Measurement System Analysis for Categorical Data: Agreement and Kappa Type Indices," *Journal of Quality Technology*, 39, 191–202. [94,98]

De Menezes, L. M. (1999), "On Fitting Latent Class Models for Binary Data: The Estimation of Standard Errors," *British Journal of Mathematical and Statistical Psychology*, 52, 149–168. [97]

Embretson, S. E., and Reise, S. P. (2000), *Item Response Theory for Psychologists*, London: Law Erlbaum Associates. [95]

Formann, A. K. (2003), "Latent Class Model Diagnostics—A Review and Some Proposals," *Computational Statistics & Data Analysis*, 41, 549–559. [97]

Goodman, L. A., and Kruskal, W. H. (1954), "Measures of Association for Cross Classifications," *Journal of the American Statistical Association*, 49, 732–764. [95,104]

Haberman, S. J. (1988), "Association, Measures of," in *Encyclopedia of Statistical Sciences*, Vol. 1, eds. S. Kotz and N. L. Johnson, Chichester: Wiley. [94,104]

ISO (1995), *Guide to the Expression of Uncertainty in Measurements*, Geneva: ISO. [95]

Kendall, M., and Gibbons, J. D. (1990), *Rank Correlation Methods* (5th ed.), London: Arnold. [95,104]

Lord, F. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale: Lawrence Erlbaum. [95,98]

Masters, G. N. (1982), "A Rasch Model for Partial Credit Scoring," *Psychometrika*, 47, 149–174. [95,96]

Muraki, E. (1992), "A Generalized Partial Credit Model: Application of an EM Algorithm," *Applied Psychological Measurement*, 16, 159–176. [95,96]

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing. [100]

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Chicago: University of Chicago Press. [96]

Samejiima, F. (1969), "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychometrika Monograph Supplement*, 17. [96]

Stroud, A. H., and Secrest, D. (1966), *Gaussian Quadrature Formulas*, Englewood Cliffs, NJ: Prentice Hall. [97]

Wright, B. D. (1997), "A History of Social Science Measurement," *Educational Measurement: Issues and Practice*, 16, 33–45. [96]