

RESEARCH REPORT

Using Ordinal Rescore Measures to Monitor Rater Drift

AUTHORS

John R. Donoghue and Adrienne Sgammato

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

Daniel F. McCaffrey
Lord Chair in Measurement and Statistics

ASSOCIATE EDITORS

Usama Ali
Senior Measurement Scientist

Beata Beigman Klebanov
Principal Research Scientist, Edusoft

Katherine Castellano
Managing Principal Research Scientist

Larry Davis
Director Research

Jamie Mikeska
Managing Senior Research Scientist

Teresa Ober
Research Scientist

Jonathan Schmidgall
Senior Research Scientist

Jesse Sparks
Managing Senior Research Scientist

Zuowei Wang
Senior Measurement Scientist

Klaus Zechner
Senior Research Scientist

Jiyun Zu
Senior Measurement Scientist

PRODUCTION EDITOR

Ayleen Gontz
Mgr. Editorial Services

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Using Ordinal Rescore Measures to Monitor Rater Drift

John R. Donoghue & Adrienne Sgammato

ETS Research Institute, Princeton, New Jersey

Abstract

When constructed response items are used on more than one occasion, a natural concern is whether the scoring is consistent (e.g., not more lenient or strict) across the occasions. It is common to conduct trend scoring, in which a set of Occasion A responses are rescored at Occasion B. The responses are usually selected according to some rescore design, such as being balanced (with an equal number from each score category), proportional to the distribution of Occasion A scores, or a mixed version of these two designs. Recent work has demonstrated that treating the two-way table as if it arose from multinomial sampling is incorrect and can yield seriously biased estimates of whether the scores are lower or higher at Occasion B. The present study builds on these results by incorporating ordinal measures of change. It contrasts the usual trend analysis with an alternative analysis that explicitly conditions on the rescore design and finds only the latter to be effective. Omnibus measures based on combining the individual t -tests or d -statistics are examined. Measures were somewhat conservative in Type I error control and had good power to detect drift. Omnibus measures based on t -tests had marginally higher power, having higher correct detection rates than those based on the d -statistic in 1%–8% of the cases. The difference between the best versions (E_{weighted} , which is based on t -tests, vs. D_{weighted} , which is based on d -statistics) was only 1.8%.

Keywords: constructed response items, score, rescore, bias, Type 1 error, drift, omnibus measures, d -statistic, t -test

Corresponding author: John R. Donoghue, E-mail: jdonoghue@ets.org

Introduction

Use of constructed response (CR) items is widespread. One advantage of CR items is that they require the production of a response, which often taps into different aspects of the domain of interest compared to selected responses (Livingston, 2009). A downside to the use of CR items is

that the responses must be scored. When the same CR items are administered on two occasions, Occasion A and Occasion B, it is important to evaluate whether the scoring is comparable for the administrations. Occasions A and B might be two human scores of responses from two administrations of an assessment. However, the same issues arise when comparing, for example, human scores to those provided by an automated scoring engine or comparing scores of an existing engine to another engine (even if the second engine is a revised/improved version of the first). For simplicity, in this report, the original scoring will be referred to as Occasion A, and the rescoring will be Occasion B.

CR scoring is expensive, and changes in scoring across occasions (i.e., rater drift) can result in biased estimates of the change from Occasion A to Occasion B. In some cases, it may be necessary to treat an item as if it were different items at the two occasions. In the most extreme cases, it may be necessary not to use (to “drop”) the item for Occasion B.

In *trend scoring*, a selection of the Occasion A responses are rescored at Occasion B, and the scores are compared. The two sets of scores are usually cross-tabulated to form a two-way table. In evaluating trend scoring, it is common to treat the table generated as a two-way contingency table, arising from multinomial sampling. If the margins were of interest (i.e., whether Occasion B scores are higher than Occasion A's), one would then compute either a paired *t*-test or an alternative, such as Stuart's (1955) *Q*. If agreement were the chief feature of interest, one would use a measure such as Cohen's (1960) kappa or weighted kappa (Cohen, 1968).

Significance tests of these statistics assume that the table is a sample from some population of responses and that the table follows a multinomial distribution. This is appropriate if the set of scores is sampled and the margins are the observed totals of the observed scores. However, this is usually not true of trend scoring. In most cases, the responses from Occasion A are selected according to some plan, such as (a) ensuring an equal number from each of the response categories; (b) selecting responses proportionately to the Occasion A distribution; or (c) employing a mixture of the two, such as 50% equal distribution and 50% proportional. We refer to this planned distribution of Occasion A responses as the *rescore design*.

When responses are selected according to a rescore design, the Occasion A margins of the rescore table are fixed by the rescore design. In this case, the sampling is no longer multinomial. Instead, each level of Occasion A scores follows a separate multinomial distribution. Because the

table margins are fixed by the rescore design, the proper sampling model is a product-multinomial (Feinberg, 1980, p. 30). Donoghue et al. (2022) and McClellan et al. (2023) showed that treating the table as if it were multinomial can lead to biased t -statistics and kappa coefficients, where the bias can be either positive or negative.

Table 1 provides examples in which the conclusion would be no drift, scorers are more lenient, or scorers are more stringent strictly as a function of the rescore design. Correct analysis of the rescore data requires acknowledging the fact that the Occasion A margins are fixed. Donoghue and Eckerly (2024) suggested computing t -statistics separately within each Occasion A score point and then aggregating the results. They also suggested three omnibus E -statistics (made by combining the individual t -tests). These measures had good Type I error rates and power and were not subject to the bias seen in the paired t -test.

Table 1. Example Rescore Tables With Identical Conditional Row Probabilities but Differing in Trend Design, Illustrating Difference in t -Test and d -Statistic

Occasion B score	Occasion A score		Total
	1	2	
Mean diff = 0, $t = 0$, $d_{wb} = 0$, $z_{wb} = 0$			
1	25	25	50
2	25	25	50
Totals	50	50	
Mean diff = 0.4, $t = 6.83$, $d_{wb} = 0.4$, $z_{wb} = 6.83$			
1	5	5	10
2	45	45	90
Totals	50	50	
Mean diff = -0.3, $t = -4.66$, $d_{wb} = -0.6$, $z_{wb} = -4.61$			
1	40	40	80
2	10	10	20
Totals	50	50	

One weakness of that work is that the t -statistic treats the values as interval scores, whereas the reality is that CR scores are only ordinal indicators of the underlying response quality. More recently, Sgammato and Donoghue (2018) recommended using Stuart's (1955) Q -statistic for marginal homogeneity, which treats the margins as nominal. Bowker (1948), Clayton (1974), and Agresti (1983) demonstrated tests of marginal homogeneity for ordinal measures and others formed of regression analyses (e.g., McCullagh, 1977, 1980; Long, 1997). Other measures, such as the Mann–Whitney U or Cliff's (1993) d , correctly reflect the ordinal level of measurement. Under certain circumstances, these ordinal tests can be more powerful than the t -

test (Feng & Cliff, 2004), and Cliff (1993, 1996a, 1996b) has noted that many times, the ordinal statistics align directly with the research question “Are Occasion B scores higher than Occasion A scores?” Unfortunately, these measures fail to reflect the product-multinomial sampling in rescore tables.

The purpose of the current study was to bring together these lines of work using a measure that reflects the ordinal nature of the data while simultaneously acknowledging the product-multinomial sampling scheme. Its unique contribution is the use of ordinal measures (Cliff’s d) in evaluating a rescoring study. The ordinal d -statistic was chosen because it has been shown to have good power, at times exceeding that of the t -test when applied to the same data (Feng & Cliff, 2004). In addition, the measure has an intuitive interpretation as an effect size: What proportion of the scores for Group 1 are higher than Group 2 scores, versus the opposite?

The rest of the report is organized as follows. First, Cliff’s d is introduced in the general case of comparing two independent groups, then extended to the within-subjects case. Next, the report examines a trend analysis of rescore data and points out the observation from Donoghue et al. (2022) that the usual multinomial sampling assumption does not hold in the presence of a rescore design. Conditional analysis is introduced, and six statistics based on conditional analysis are given. That is followed by a large simulation study. Results are presented for trend analysis, followed by results for conditional analysis, including comparisons of the six conditional analysis measures. Finally, the report finishes with discussion and concluding remarks.

Cliff’s d

This section introduces the d -statistic in general. Its use in trend analysis is discussed in the next section. In the general case of comparing two independent groups, the ordinal¹ d -statistic is defined:

$$d = \frac{\#(X > Y) - \#(Y > X)}{mn}, \quad (1)$$

where function $\#()$ indicates the count of cases in which the argument is true, n is the number of X -scores, and m is the number of Y -scores. Cliff (1993) proposed using “dominance relations” to address the question “Are the scores in Group X higher than those in Group Y?” A dominance relation d_{ij} is defined as

$$d_{ij} = \text{sign}(x_i - y_j), \quad (2)$$

which can be arranged into a matrix as shown in Table 2. Note that the entries indicate whether the row value is larger than the column entry. It is also useful to define the marginal proportion:

$$d_{i\cdot} = \frac{\sum_{j=1}^m d_{ij}}{m}, \quad (3)$$

with an analogous definition of the row proportion $d_{\cdot j}$.

Table 2. Matrix of Dominance Relations

X-score	Y-score					
	1	3	4	7	8	d
6	1	1	1	-1	-1	0.2
7	1	1	1	0	-1	0.4
9	1	1	1	1	1	1
10	1	1	1	1	1	1
d	1	1	1	0.25	0	0.65

Note. Adapted from “Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions,” by N. Cliff, 1993, *Quantitative Methods in Psychology*, 114(3), p. 500, Table 1. Copyright 1993 by the American Psychological Association Inc.

The d -statistic can be readily defined in terms of d_{ij} :

$$d = \frac{\sum_{i=1}^n \sum_{j=1}^m d_{ij}}{mn}. \quad (4)$$

The standard error of d is given as

$$\hat{s}_d^2 = \frac{m^2 \sum_{i=1}^n (d_{i\cdot} - d)^2 + n^2 \sum_{j=1}^m (d_{\cdot j} - d)^2 - \sum_{i=1}^n \sum_{j=1}^m (d_{ij} - d)^2}{mn(m-1)(n-1)}. \quad (5)$$

Cliff (1993) also presented a paired version of the d -statistic, along with associated standard error to facilitate significance testing and the construction of confidence intervals. In this case, rows represent scores in condition X , and columns represent the scores in condition Y . Cliff pointed out that three interrelated questions were of interest:

1. **Within-subject, measured by d_w .** Are the responses of a subject higher in the Y -condition than they are in the X -condition? This is the diagonal of the dominance matrix:

$$d_w = \frac{\sum_{i=1}^n d_{ii}}{n}, \quad s_{d_w}^2 = \frac{\sum_{i=1}^n (d_{ii} - d_w)^2}{n-1}. \quad (6)$$

2. **Between subject, measured by d_b .** Do different members of the group score higher in the Y -condition than they do in the X -condition? This compares the off-diagonal elements of the dominance matrix:

$$d_b = \frac{\sum \sum_{j \neq i} d_{ij}}{n(n-1)}, \quad s_{d_b}^2 = \frac{s_{ij}^2 + \text{cov}(d_{ij}, d_{ji}) + (n-2) \left[s_{d_{i.}}^2 + s_{d_{.i}}^2 + 2 \text{cov}(d_{i.}, d_{.i}) \right]}{n(n-2)}. \quad (7)$$

3. **Combined, measured by d_{bw} .** Overall, are scores in the Y -condition higher than those in the X -condition? This combines information from the whole matrix, that is, from d_w and d_b :

$$d_{bw} = d_b + d_w, \quad s_{d_{bw}}^2 = s_{d_b}^2 + s_{d_w}^2 + 2 \text{cov}(d_w, d_b), \quad (8)$$

where

$$\text{cov}(d_w, d_b) = \frac{\sum_{i=1}^n \left[\left(\sum_{j \neq i} d_{ij} + \sum_{j \neq i} d_{ji} \right) d_{ii} \right] - 2n(n-1)d_b d_w}{n(n-1)(n-2)}. \quad (9)$$

Note that d_{bw} can be larger than 1 and so is no longer interpretable as a probability.

Approaches to Analyzing Trend Scoring Data

We differentiate two forms of analysis of the cross-occasion data. *Trend analysis* refers to analyzing the rescore table as if it were a two-way table derived from multinomial sampling. *Conditional analysis* refers to explicitly accounting for the product-multinomial sampling of the rescore table.

In trend analysis, the scores are paired. A common test to determine if scores at Occasion B are lower or higher than they are at Occasion A is a paired t -test. For the d -statistic, trend analysis uses the paired d -statistics (d_w , d_b , and d_{wb}) computed from the rescore table. The dominance matrix is constructed, and the statistics are computed according to the preceding equations. As noted, the three d -statistics ask slightly different questions. For simplicity, this report focuses on d_{wb} . Results for d_w and d_b show the same patterns and so are not presented in the interest of space.

Conditional Analysis

Because the Occasion A margins of the rescore table are fixed by the rescore design, comparisons like the trend analysis based on the margins are at best misleading. Statistics that

are invariant to the margins are the conditional probabilities $P(Y|X)$, the probability of score Y on Occasion B given that a score of X was observed at Occasion A. However, to evaluate the conditional probabilities, a comparison is needed. Here it is assumed that there was within-occasion monitoring at Occasion A, which involved having a second score assigned at Occasion A, and so a within-occasion rescore table is available. The conditional probabilities from the within-occasion table are then compared to those from the rescore table (for more detail, see Donoghue & Eckerly, 2024). The key idea is to consider only papers that received a specific score k from the Occasion A first rater. From the within-occasion score table, extract that row of counts. Extract the same row from the rescore table. Finally, compute an independent-groups test (t -test or d -statistic) comparing these two sets of scores.

One challenge of this approach is that it yields one test statistic for each level of the Occasion A score. Frequently, an omnibus statistic is required to answer the question “Overall, are Occasion B scores higher or lower than Occasion A scores?” To address this, Donoghue and Eckerly (2024) proposed three E -statistics, based on different ways of combining the individual t -tests. E_{pooled} sums the numerators and denominators, then divides the two to come up with a test statistic:

$$E_{\text{pooled}} = \frac{\sum_{k=0}^K (\bar{x}_k - \bar{y}_k)}{\sqrt{\sum_{k=0}^K s_{(\bar{x}_k - \bar{y}_k)}^2}}. \quad (10)$$

Relying on the t -test’s approach to the normal distribution as the degrees of freedom increase, tests for this statistic are conducted comparing it to a standard normal distribution.

The second statistic, E_{weighted} , weights the individual statistics by their frequency in the Occasion A scoring, forming a weighted sum of the numerators and a weighted sum of the denominators:

$$E_{\text{weighted}} = \frac{\sum_{k=0}^K w_k (\bar{x}_k - \bar{y}_k)}{\sqrt{\sum_{k=0}^K w_k^2 s_{(\bar{x}_k - \bar{y}_k)}^2}}. \quad (11)$$

E_{weighted} is compared to a standard normal distribution.

The third statistic, E_{χ^2} , is formed by squaring the individual t -tests:

$$E_{\chi^2} = \sum_{k=0}^K t_k^2. \quad (12)$$

Under the assumption that the individual tests are approximately standard normal, E_{χ^2} is compared to a χ^2 variate with degrees of freedom equal to the number of terms summed (the number of categories of the Occasion A score).²

Cliff notes that d divided by its standard error is asymptotically distributed as a standard normal variable. On this basis, omnibus measures of d were computed, hereinafter referred to as D -statistics, that were exact analogs of the E -statistics:

$$D_{\text{pooled}} = \frac{\sum_{k=0}^K d_k}{\sqrt{\sum_{k=0}^K s_{d_k}^2}}, \quad (13)$$

$$D_{\text{weighted}} = \frac{\sum_{k=0}^K w_k d_k}{\sqrt{\sum_{k=0}^K w_k^2 s_{d_k}^2}}, \quad (14)$$

$$D_{\chi^2} = \sum_{k=0}^K \left(\frac{d_k}{s_{d_k}} \right)^2. \quad (15)$$

The difference is that the terms in the summations are individual d -tests instead of t -tests.

Because the two sets of scores are separate responses in conditional analysis, only the between-subjects independent d -test was used for the conditional analyses.

One detail in computing the omnibus E - and D -statistics is that, for extreme distributions of Occasion A scores (typically a high IRT a -parameter coupled with an extreme b -parameter), it is possible that one level might yield a set of scores for which the t -test and d -test could not be computed. In these cases, the computation was modified to ignore the level in question. In this case, the degrees of freedom for E_{χ^2} and D_{χ^2} were modified accordingly.

Method

To explore the design space, an extensive simulation study was conducted to examine Type I error rate and power. The factors are summarized in Table 3. To model use of the same test taker responses, the same θ (representing the quality of the CR) was used with the Occasion A and Occasion B item response theory (IRT) parameters to generate responses. Data were generated using Python 3.9. Most data manipulation and computation of the target statistics was conducted in R (Version 4.2.2; R Core Team, 2022). The exception was that computation of the dependent Cliff d -statistics was done using a Java 11 program for better performance. Finally, statistical analysis of the outcome data used SAS and R.

Table 3. Factors Varied in Simulation

Factor	Number of levels	Levels
Number of cases	6	50, 100, 200, 400, 600, 1,000
Rescore design	3	Proportional, balanced, mixed (50% proportional, 50% balanced)
Occasion A b -parameter	5	$b_0 = -1.0, -0.5, 0, 0.5, 1.0$
Change in b -parameter from Occasion A to Occasion B	5	$b_{\text{shift}} = -1.0, -0.5, 0, 0.5, 1.0$
Occasion A a -parameter	5	$a_0 = 0.7, 1.0, 1.3, 1.5, 2.0$
Occasion B a -parameter	5	$a_{\text{alt}} = 0.7, 1.0, 1.3, 1.5, 2.0$
No. score categories and IRT model	7	2 (2PL), 3, 4, 5 (GPCM or GRM)

Note. 2PL = two-parameter logistic; GPCM = generalized partial credit model; GRM = graded response model; IRT = item response theory.

Item Response Theory Models Used

For dichotomous items, the two-parameter logistic (2PL) model was used:

$$P(\theta) = \frac{\exp[1.7a(\theta - b)]}{1 + \exp[1.7a(\theta - b)]}. \quad (16)$$

For polytomous items, the graded response model (GRM),

$$P_k(\theta) = P_k^+ - P_{k-1}^+, \quad (17)$$

with

$$P_k^+ = P(x \geq k | \theta) = \frac{\exp[1.7a(\theta - b + d_k)]}{1 + \exp[1.7a(\theta - b + d_k)]}, \quad (18)$$

was used for half of the items, and the generalized partial credit model (GPCM),

$$P_k(\theta) = \frac{\exp\left[1.7a \sum_{n=0}^k (\theta - b + d_n)\right]}{\sum_{v=0}^K \exp\left[1.7a \sum_{n=0}^v (\theta - b + d_v)\right]}, \quad (19)$$

was used for the other half. For the polytomous items, b was determined by the value of b_0 for items without scoring drift, and for items exhibiting drift, by $b_0 + b_{\text{shift}}$. The category thresholds d_k were chosen: $(-0.75, 0.75)$ for three-category items, $(-0.75, 0.0, 0.75)$ for four-category items, and $(-0.75, -0.25, 0.25, 0.75)$ for five-category items. There is no assertion that these parameters are equivalent across the two polytomous IRT models; rather, the parameters were chosen to yield data that look like scoring data.

Data Generation Factors

As shown in Table 3, the design contained seven factors, each with several levels. The factors were fully crossed:

- **Number of response categories and IRT model (seven levels).** The number of response categories was two, three, four, or five. The 2PL model was used for two-category data. The remaining six levels come from cross-generating model GRM or GPCM with the three levels of numbers of categories. Note that the same IRT model was used for all scores of an item, although (as described later) the item parameters could be different if the item exhibited drift. Holding θ fixed for the two scoring occasions corresponds to the fact that the underlying quality of the CRs has not changed. Changing the IRT parameters for Occasion B represents a shift in the overall scoring process (e.g., due to training differences) at Occasion B.
- **Number of cases (six values).** The six values are 50, 100, 200, 400, 600, and 1,000.
- **Occasion A b -parameter, b_0 (five values).** The five values are -1 , -0.5 , 0 , 0.5 , and 1 .
- **Change in b -parameter from Occasion A to Occasion B, b_{shift} (five values).** The five values are -1 , -0.5 , 0 , 0.5 , and 1 .
- **Occasion A a -parameter, a_0 (five values).** The five values are 0.7 , 1.0 , 1.3 , 1.5 , and 2.0 .
- **Occasion B a -parameter, a_{alt} (five values).** The five values are 0.7 , 1.0 , 1.3 , 1.5 , and 2.0 .
- **Rescore design (three levels), balanced, proportional, or mixed.** In the balanced design, an equal number of papers was generated for each Occasion A response category. In the proportional design, the number of Occasion A papers mirrored the expected distribution of Occasion A responses. Using the IRT parameters and assuming an $N(0, 1)$ distribution of ability, the item response function was evaluated at 41 points $[-4, 4]$. This was multiplied by the height of the normal density at that point and summed to compute the expected proportion in that category. This was then multiplied by the number of cases to come up with the number of responses for each category. Fractional responses were arbitrarily assigned to the lowest response category. For the mixed design, one-half of the papers were selected according to the balanced design, and one-half were selected according to the proportional design.

The seven design factors were crossed to yield $7 \times 6 \times 5 \times 5 \times 5 \times 5 \times 3$ (78,750) data generation conditions; 1,000 replications were generated for each cell. In some replications, all responses fell into one of the categories, making the paired t -test and Q impossible to compute. Another situation was if agreement happened to be perfect (all off-diagonal cells = 0.0), the denominator of the t -test was undefined. In a small number of additional conditions, the covariance matrix used in computing Stuart's Q (see the appendix for details) was singular, preventing its inversion. This was associated with extreme combinations of b_0 and b_{shift} and with high a -parameter values. These replications were replaced until the full 1,000 were obtained for each cell. The proportion of the 1,000 values for which the statistic was significant was recorded, and these rates are the outcome measures for the study.

Analysis Factors

For trend analysis, the rescore data were treated as a two-way table. Paired t -test and Cliff's paired d were computed. For conditional analysis, independent-groups t -test and Cliff's d were computed separately for each level of Occasion A scores. One set of values comprised the within–Occasion A second scores. The other set of values comprised the Occasion B scores. Thus there are four analyses of each data set:³ (a) trend, paired t -test; (b) trend, Cliff's paired d ; (c) conditional, t -test, with three omnibus measures considered (E_{pooled} , E_{weighted} , and E_{χ^2}); and (d) conditional, Cliff's d , again with three omnibus versions considered (D_{pooled} , D_{weighted} , and D_{χ^2}).

Data Generation

For each response, a θ -value (representing quality of the response) was drawn from an $N(0, 1)$ distribution. Next, using the IRT model, the probability of each response category (conditional on θ) was computed, then summed to form a cumulative distribution. A uniform random number was drawn, and the response category was assigned based on which of the category probability values contained the uniform value. This was the Occasion A first score. According to the rescore design, if the number of responses for that category had already been reached, the θ and response were discarded. Another θ was drawn, and an associated response was generated. This process continued until the number of responses required by the rescore design was obtained. For within–Occasion A rescoring, the same θ was used with the same item parameters. A second uniform random number was drawn and used to assign a within-occasion rescore response.

Generation of the cross-occasion rescore table proceeded similarly. To reflect the independence of the rescore table from the within-occasion scoring, a new θ value was drawn, and an Occasion A response was generated, subject to the constraint on the limits imposed by the rescore design. For the Occasion B response, the same θ was used, but the item parameters for the second score were chosen according to the condition. These new item parameters were used to compute Occasion B probabilities, and a uniform number was then drawn to determine the Occasion B score.

The final results of the data generation were two tables with the same row margins (first score Occasion A, determined by the rescore design). The cell values and column (rescore) totals were free to vary.

Results

The results are presented in two phases. The first phase reports the analyses of the trend analysis. The second phase reports the results for the conditional analyses.

Trend Analyses

Type I Error

We first examine the Type I error behavior. The data were a subset of the 3,150 conditions in which the null hypothesis was true: $b_{\text{shift}} = 0$ and $a_0 = a_{\text{alt}}$. A descriptive analysis of variance (ANOVA) was computed to identify which factors were associated with large proportions of variance in the Type I error rate. A practical effect size of

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \geq 0.01 \quad (20)$$

was adopted. Selected results are given in Table 4. Note that b_{shift} is not in the ANOVA because it is constant. Similarly, a_{alt} is not included because it must equal a_0 in the null condition.

The D [Rescore Design $\times B$ (b_0) $\times N$ (number of cases)] was identified as salient, as were each of the constituent main effects and two-way interactions. None of the other two-way or three-way interactions were identified as salient. Figures 1 and 2 show the three-way interaction for t -test and Cliff's paired d , respectively. As the figures make clear, the Type I error rate is grossly inflated for the balanced design, while it is well controlled for the proportional design. As would be expected, the mixed design falls between these two extremes but generally displays

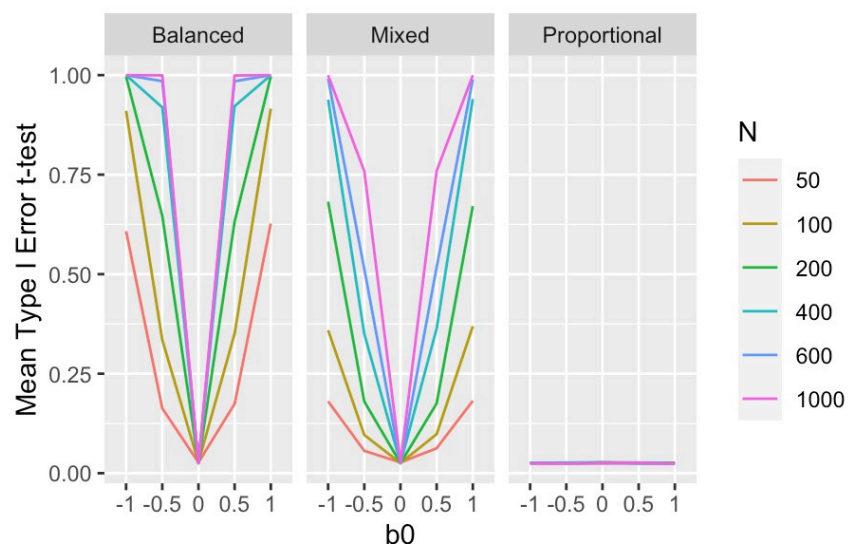
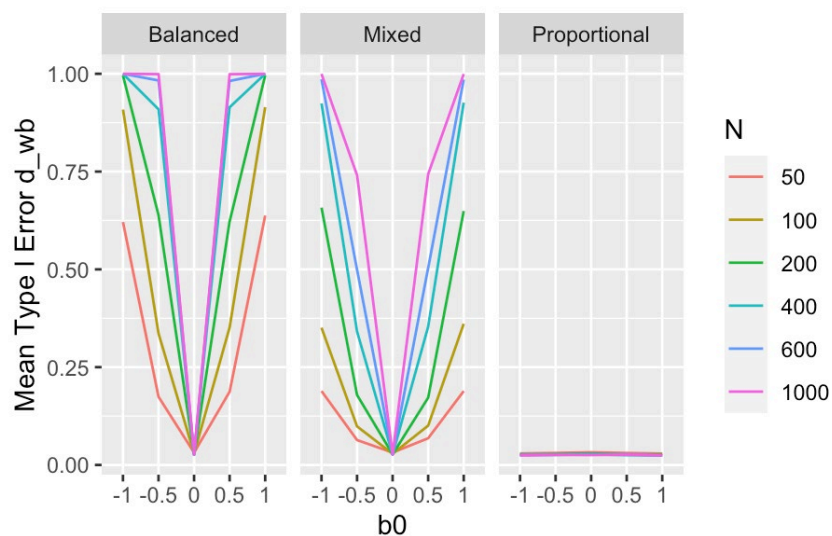
inflated Type I error. Cliff's d is not immune to these effects. The pattern largely parallels that of the other statistic.

Table 4. Selected Analysis of Variance Results for Paired t -Test and d_{wb}

Variable	df	t -test		d_{wb}	
		ANOVA SS	η^2	ANOVA SS	η^2
M	2	0.03	0.000	0.11	0.000
K	3	0.04	0.000	0.01	0.000
N	5	53.76	0.106	94.37	0.220
D	2	205.25	0.403	124.17	0.289
B	4	114.26	0.224	70.73	0.165
A	4	0.16	0.000	0.11	0.000
M \times N	10	0.01	0.000	0.04	0.000
M \times D	4	0.04	0.000	0.11	0.000
M \times B	8	0.02	0.000	0.04	0.000
M \times A	8	0.21	0.000	0.13	0.000
K \times N	15	0.01	0.000	0.03	0.000
K \times D	6	0.05	0.000	0.04	0.000
K \times B	12	0.05	0.000	0.03	0.000
K \times A	12	0.16	0.000	0.17	0.000
N \times D	10	30.09	0.059	46.22	0.108
N \times B	20	16.66	0.033	25.69	0.060
N \times A	20	0.02	0.000	0.04	0.000
D \times B	8	65.82	0.129	41.58	0.097
D \times A	8	0.15	0.000	0.12	0.000
B \times A	16	0.07	0.000	0.05	0.000
N \times D \times B	40	20.54	0.040	24.20	0.056
Total	3,149	509.20		429.74	

Note. Effect sizes $\eta^2 \geq .01$ are in boldface. ANOVA SS = analysis of variance sum of squares; design; M = item response theory model; K = number of response categories; N = number of cases; D = rescore; B = b_0 .

The findings in Figures 1 and 2 parallel the findings in Donoghue et al. (2022) and the simulation results of Donoghue and Eckerly (2024). The result is clear: Ignoring the sampling model and treating rescore data as if the data arise from a multinomial, two-way table can yield very misleading results. The exception is when the rescore design specifies numbers of responses that are proportional to the Occasion A marginal distribution. Note that Type I error is noticeably lower for the condition $b_0 = 0$ than it is for the other values. In this condition, the proportions in each category are equal. Thus the balanced design, proportional design, and mixed design correspond in this condition.

Figure 1. Type I Error Rate for t -Test as a Function of Design, b_0 , and N **Figure 2. Type I Error Rate for d_{wb} -Statistic as a Function of Rescore Design, b_0 , and N** ***Detection/Power***

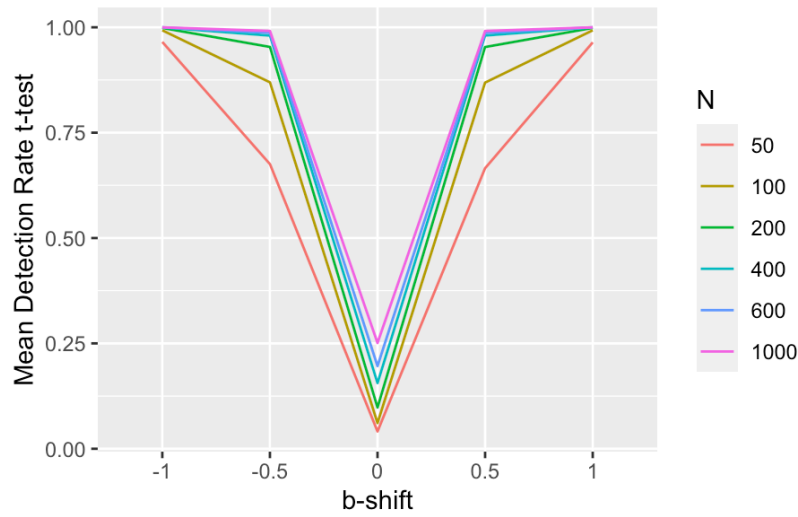
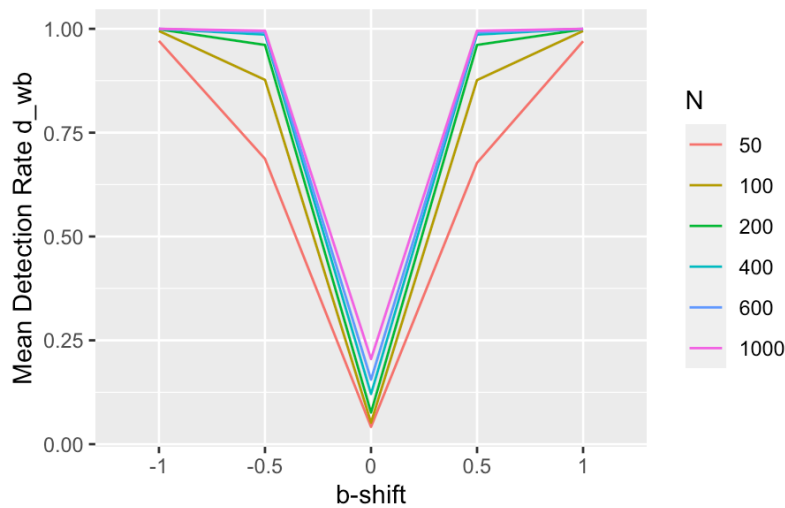
Because of the grossly inflated Type I error rates observed for the mixed and balanced conditions, this analysis is restricted to the proportional condition where the Type I error rate was well controlled. Therefore these results can accurately be termed “power.” Table 5 gives selected ANOVA results for detection rates for each of the measures.

Table 5. Selected Analysis of Variance Results for Detection (Power) Proportional Design Only

Variable	df	<i>t</i> -test		z- <i>d</i> _{wb}	
		ANOVA SS	η^2	ANOVA SS	η^2
M	2	5.33	0.002	3.94	0.001
K	3	6.39	0.002	4.42	0.001
N	5	99.47	0.033	98.30	0.032
B	4	0.17	0.000	0.06	0.000
Db	4	2,372.02	0.781	2,485.28	0.818
A	4	1.09	0.000	2.17	0.001
Aalt	4	8.33	0.003	7.43	0.002
M × N	10	27.74	0.009	22.72	0.007
M × B	8	0.74	0.000	0.67	0.000
M × Db	8	58.26	0.019	43.28	0.014
M × A	8	1.76	0.001	1.63	0.001
M × Aalt	8	6.69	0.002	6.15	0.002
K × N	15	14.19	0.005	12.33	0.004
K × B	12	0.14	0.000	0.14	0.000
K × Db	12	16.68	0.005	19.00	0.006
K × A	12	0.68	0.000	0.88	0.000
K × Aalt	12	1.98	0.001	2.12	0.001
N × Db	20	62.61	0.021	67.30	0.022
N × A	20	9.68	0.003	10.84	0.004
N × Aalt	20	9.38	0.003	9.43	0.003
B × Db	16	32.41	0.011	23.25	0.008
B × A	16	0.29	0.000	0.20	0.000
B × Aalt	16	0.44	0.000	0.20	0.000
Db × A	16	11.30	0.004	10.42	0.003
Db × Aalt	16	20.78	0.007	16.28	0.005
A × Aalt	16	106.11	0.035	107.76	0.035
Total	25,199	3,038.85		3,057.28	

Note. Effect sizes $\eta^2 \geq .01$ are in boldface. A = a_0 ; Aalt = a_{alt} ; ANOVA = analysis of variance; B = b_0 ; D = rescore design; Db = b_{shift} ; K = number of response categories; M = item response theory model; N = number of cases.

Compared to Table 4, design is not present, as it is held constant for this analysis. On the other hand, both b_{shift} and a_{alt} are now included. Again, an effect size criterion of $\eta^2 \geq 0.01$ is adopted. As would be expected, N , b_{shift} , and their interaction are identified as salient. Figures 3 and 4 show this interaction for each of the measures.

Figure 3. Interaction of Sample Size N and b_{shift} for t -Test**Figure 4. Detection for d_{wb} by b_{shift} and Sample Size**

There is a U-shaped relationship between b_{shift} and detection that is minimal at $b_{\text{shift}} = 0$ and increases as the value diverges from 0. The steepness of the curve is affected by N . When $b_{\text{shift}} = 0$, the only way that the null can be false is if the a -parameters differ. For small samples, the power is quite poor, but for large samples, the power is nonnegligible. Referring to Table 5, the interaction of $a_0 \times a_{\text{alt}}$ was identified as salient. Tables 6 and 7 show this interaction. When the two a -parameters differ, there is moderate power. The elevation of the diagonal of equality is an artifact of the design. When the two slopes are the same, the b_{shift} must be nonzero. Because

the measures are more sensitive to differences in item difficulty, power is relatively good in this condition. Figures 5 and 6 portray this interaction.

Table 6. Mean Detection Rate t -Test as a Function of a_0 and a_{alt}

Value a_{alt}	a_0				
	0.7	1	1.3	1.5	2
0.7	0.90	0.75	0.76	0.76	0.77
1	0.76	0.95	0.78	0.79	0.80
1.3	0.79	0.78	0.97	0.79	0.80
1.5	0.80	0.79	0.79	0.98	0.80
2	0.81	0.81	0.80	0.80	0.99

Table 7. Mean Detection Rate of d_{wb} -Statistic as a Function of a_0 and a_{alt}

Value a_{alt}	a_0				
	0.7	1	1.3	1.5	2
0.7	0.90	0.75	0.76	0.77	0.78
1	0.76	0.95	0.78	0.79	0.80
1.3	0.78	0.78	0.97	0.79	0.80
1.5	0.79	0.79	0.79	0.98	0.80
2	0.80	0.81	0.80	0.80	0.99

Figure 5. Interaction of Item Response Theory Model and b_{shift} for Detection by t -Test

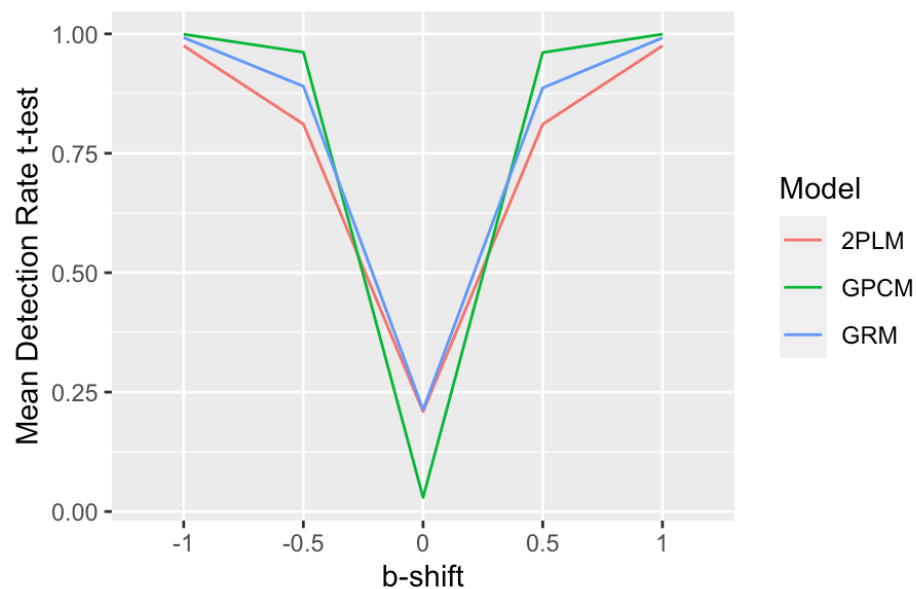
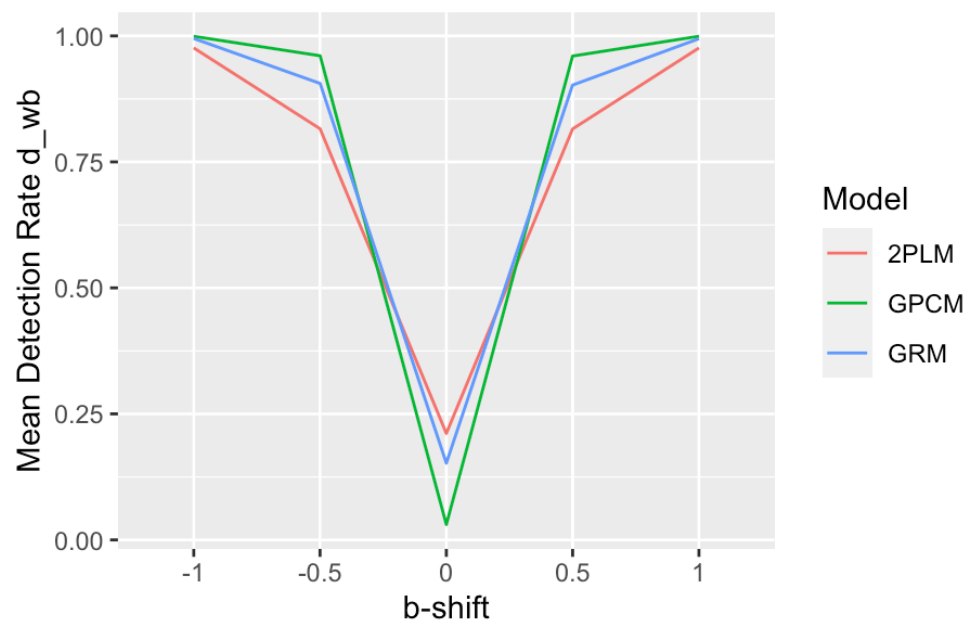
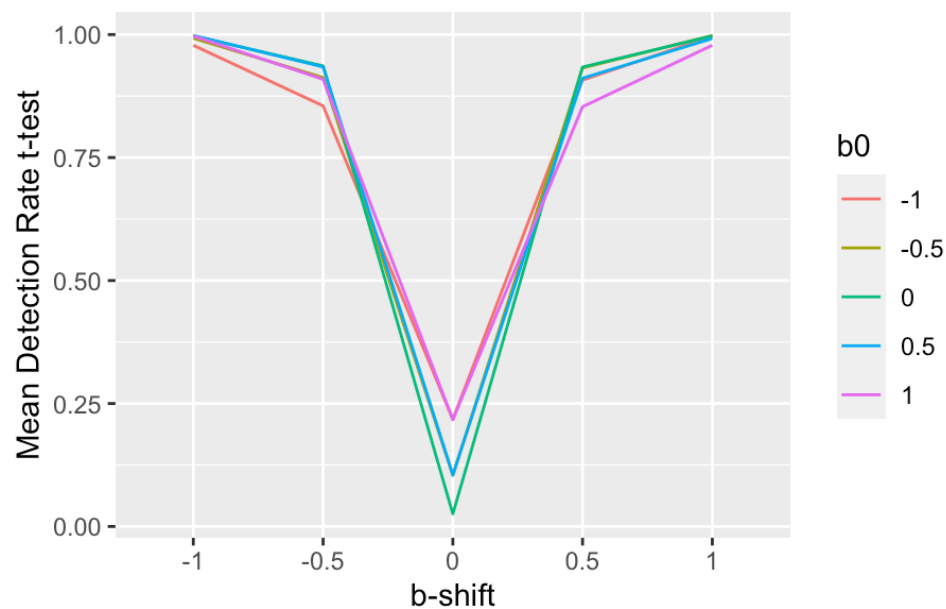


Figure 6. Interaction of Item Response Theory Model and b_{shift} for Detection by d_{wb} 

One other interaction was flagged for the paired t -test: The $b_0 \times b_{\text{shift}}$ interaction was identified as salient. Figure 7 shows this interaction.

Figure 7. Mean Detection Rate for t -Test for Interaction of b_0 and b_{shift} 

Conditional Analyses

Table 8 gives the average Type I error rate for each of the statistics by rescore design. Compared to the trend analyses, the overall Type I error rates of all of the omnibus statistics are somewhat conservative, and none shows a strong effect for rescore design. Despite their overall conservative Type I error rates, the omnibus E_{χ^2} and D_{χ^2} measures were the only ones to demonstrate mild Type I error inflation (0.075–0.10).

Table 8. Type I Error Rate for Conditional Measures by Design

Design	E_{pooled}	E_{weighted}	E_{χ^2}	D_{pooled}	D_{weighted}	D_{χ^2}
Balanced	0.021	0.021	0.016	0.017	0.017	0.012
Proportional	0.022	0.022	0.018	0.020	0.015	0.016
Mixed	0.021	0.021	0.016	0.018	0.016	0.013

Figure 8 shows density plots of the Type I error rates. Both measures show conservative Type I error rates, with densities peaking around 0.01. However, both also show marked positive skew. The maximum value for E_{χ^2} is 0.083, and for D_{χ^2} , it is 0.10. Overall, 14 of 7,300 conditions (5/3,150 E_{χ^2} and 9/3,150 D_{χ^2}) were found to have elevated Type I error rates (Table 9). All but two instances were associated with the smallest sample size of 50. Type I inflation for E_{χ^2} with a sample size of 50 was also noted by Donoghue and Eckerly (2024).

Figure 8. Density Plot of D_{χ^2} and E_{χ^2} for Null Case

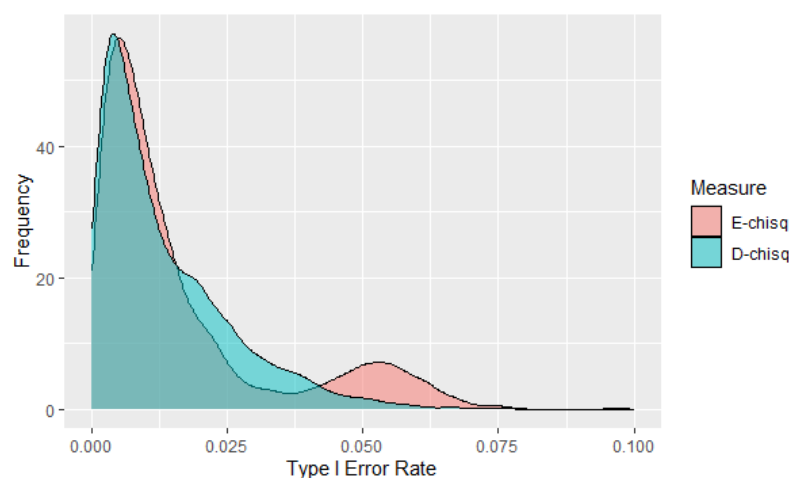


Table 9. Conditions in Which E_{χ^2} and D_{χ^2} Measures Showed Elevated Type I Error Rates

Model	N categories	N cases	Design	b_0	a_0	Measure	Type I
2PL	2	100	Mixed	0.5	1	E	0.080
2PL	2	1,000	Balanced	0.5	1.5	E	0.077
2PL	2	50	Balanced	1	1.5	E	0.076
2PL	2	50	Proportional	1	0.7	E	0.083
2PL	2	50	Mixed	0.5	2	E	0.076
GPCM	5	50	Proportional	1	0.7	D	0.085
GRM	5	50	Mixed	1	0.7	D	0.076
GRM	5	50	Proportional	-1	0.7	D	0.096
GRM	5	50	Proportional	-1	1	D	0.100
GRM	5	50	Proportional	-1	1.3	D	0.076
GRM	5	50	Proportional	-0.5	0.7	D	0.090
GRM	5	50	Proportional	0.5	0.7	D	0.097
GRM	5	50	Proportional	1	0.7	D	0.084
GRM	5	50	Proportional	1	1	D	0.096

Note. 2PL = two-parameter logistic. D = D -statistic; E = E -statistic GPCM = generalized partial credit model; GRM = graded response model.

Because of the overall good Type I behavior for all of the statistics, ANOVA analyses were deemed of limited interest and so are not presented here.

Detection/Power

Table 10 provides summary statistics for the nonnull case. All measures demonstrate good detection rates, with the median falling at 1.0 and the first quartile at 0.5 or higher. E_{χ^2} and D_{χ^2} show lower values than the pooled or weighted statistics. Also, the D -based omnibus statistics have somewhat lower means than the E -based statistics.

Table 10. Overall Detection Rates for Omnibus Measures

Statistic	E_{pooled}	E_{weighted}	E_{χ^2}	D_{pooled}	D_{weighted}	D_{χ^2}
Min	0.00	0.00	0.00	0.00	0.00	0.00
1st quartile	0.70	0.76	0.58	0.66	0.70	0.52
Median	1.00	1.00	1.00	1.00	1.00	1.00
Mean	0.78	0.79	0.77	0.77	0.78	0.75
3rd quartile	1.00	1.00	1.00	1.00	1.00	1.00
Max	1.00	1.00	1.00	1.00	1.00	1.00

To determine which factors had the largest effect on detection rates, a series of descriptive ANOVAs was conducted. As earlier, an effect size criterion of $\eta^2 \geq 0.01$ was adopted. Table 11 shows results for the E -statistics, and Table 12 shows results for the D -statistics.

Table 11. Selected Analysis of Variance Results Detection for E -Statistics

Source	df	E_{pooled}		$E_{weighted}$		E_{χ^2}	
		ANOVA SS	η^2	ANOVA SS	η^2	ANOVA SS	η^2
M	2	28.85	0.003	23.44	0.002	2.9	0.000
K	3	28.12	0.003	25.09	0.003	3.95	0.000
N	5	527.69	0.053	446.32	0.046	1,100.05	0.112
D	2	11.77	0.001	0.04	0.000	0.30	0.000
B	4	5.65	0.001	0.05	0.000	1.12	0.000
Db	4	8,318.19	0.828	7,872.32	0.814	6,789.87	0.69
A	4	9.79	0.001	9.6	0.001	59.49	0.006
Aalt	4	26.24	0.003	21.69	0.002	8.61	0.001
M \times Db	8	65.88	0.007	111.28	0.012	186.92	0.019
N \times Db	20	403.84	0.04	320.08	0.033	668.48	0.068
A \times Aalt	16	333.82	0.033	328.89	0.034	271.69	0.028
Total	75,599	10,044.33		9,675.77		9,838.81	

Note. Effect sizes $\eta^2 \geq .01$ are in boldface. A = a_0 ; Aalt = a_{alt} ; ANOVA = analysis of variance; B = b_0 ; D = rescore design; Db = b_{shift} ; K = number of response categories; M = item response theory model; N = number of cases.

Table 12. Selected Analysis of Variance Results Detection for D -Statistics

Source	df	D_{pooled}		$D_{weighted}$		D_{χ^2}	
		ANOVA SS	η^2	ANOVA SS	η^2	ANOVA SS	η^2
M	2	53.57	0.005	64.07	0.006	30.00	0.003
K	3	54.74	0.005	64.89	0.006	26.44	0.003
N	5	596.53	0.058	534.49	0.053	1,112.71	0.107
D	2	11.05	0.001	1.05	0.000	1.00	0.000
B	4	8.71	0.001	2.94	0.000	5.78	0.001
Db	4	8,490.22	0.824	8,243.82	0.813	7,535.07	0.725
A	4	10.81	0.001	8.56	0.001	48.35	0.005
Aalt	4	22.5	0.002	19.49	0.002	8.58	0.001
M \times Db	8	63.8	0.006	95.44	0.009	116.49	0.011
N \times Db	20	457.03	0.044	388.24	0.038	733.77	0.071
A \times Aalt	16	338.88	0.033	337.98	0.033	288.42	0.028
Total	75,599	10,044.33		9,675.77		9,838.81	

Note. Effect sizes $\eta^2 \geq .01$ are in boldface. A = a_0 ; Aalt = a_{alt} ; ANOVA = analysis of variance; B = b_0 ; D = rescore design; Db = b_{shift} ; K = number of response categories; M = item response theory model; N = number of cases.

For all statistics, the interaction of number of cases N and b_{shift} is flagged as salient, as are the main effects. The $a_0 \times a_{alt}$ interaction is also flagged for all statistics. Finally, the two-way interaction of model by number of cases is significant for several measures. Figure 9 portrays the means for the $N \times b_{shift}$ interaction. In all cases, the b_{shift} forms a “V,” with little detection for $b_{shift} = 0$ increasing to (nearly) perfect detection for $b_{shift} = 1.0$. The slope is fairly gentle for $N = 50$. The “V” becomes steeper for values of $b_{shift} = \pm 0.5$.

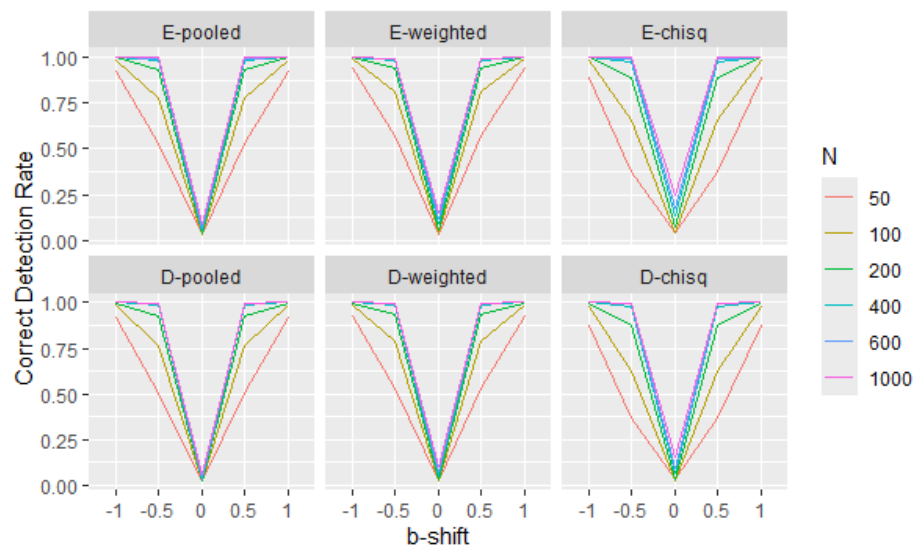
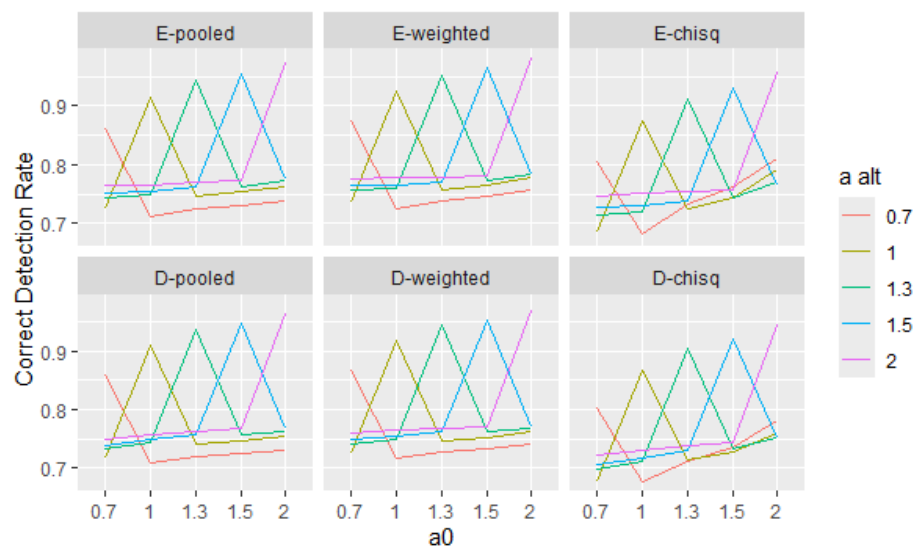
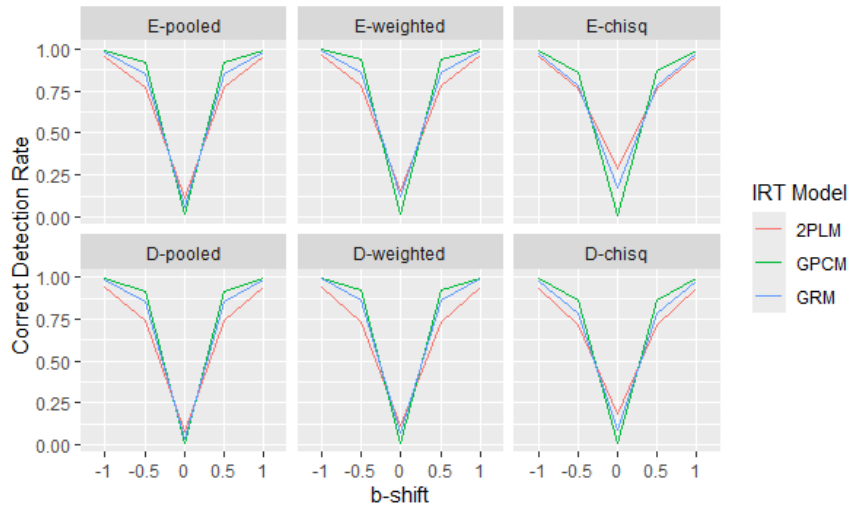
Figure 9. Mean Detection Rate for Interaction of b_{shift} With Sample Size N 

Figure 10 shows the $a_0 \times a_{\text{alt}}$ interaction. As was noted in the trend results, in all cases, there is a spike when $a_0 = a_{\text{alt}}$. In this case, $b_{\text{shift}} \neq 0$; otherwise, it would be a null case. Because the t -test and d -statistic are sensitive to changes in location, cases when $b_{\text{shift}} \neq 0$ are detected well. Outside of the spike, there is a tendency for detection to increase as the difference between a_0 and a_{alt} increases. Also, the curves have a slightly upward tilt moving from left to right, indicating that larger values of the IRT a -parameter are associated with better detection.

Figure 10. Mean Detection Rates for Interaction of a_0 and a_{alt} 

Finally, Figure 11 shows the interaction of model and b_{shift} . The curves tend to be shallowest for 2PL items and steepest for GPCM items.

Figure 11. Mean Detection Rates by b_{shift} and IRT Model Used to Generate the Data



Comparison of Methods

The final comparison concerns comparing the conditional methods. Table 13 gives the correlations between the omnibus measures. All correlations are >0.95 . Given the high correlations, results are likely to be similar across methods. However, the correlation loses information about the level of the variables, a critical feature of a significance test. It is therefore of interest to see what proportion of the time each method gives a better detection rate than the other. Because this is inherently a question of order, the d -statistic was used for these comparisons. Although the comparisons are paired, the original omnibus d_{wb} loses interpretability as a probability. Thus the independent d is used as an effect size measure for these comparisons. The mixed rescore design was used due to computer memory limitations.⁴ The comparisons of methods are presented in Table 14. Positive numbers indicate that the method listed in the column is higher than the method listed in the row. First, the E -statistics outperformed the similarly defined D -statistics: E_{pooled} is $d = 0.01$ higher than D_{pooled} , E_{weighted} is $d = 0.018$ higher, and E_{χ^2} is $d = 0.021$ higher than D_{χ^2} . Second, the weighted version of the statistic produced the best results: E_{weighted} is 0.012 higher than E_{pooled} and 0.058 higher than E_{χ^2} . For D -statistics, D_{weighted} was not significantly lower than D_{pooled} ($d = -0.006$) and was significantly

higher than D_{χ^2} , $d = 0.061$. Overall, E_{weighted} had the best performance, being higher than the other methods 1.2%–7.9% of the time.

Table 13. Correlations Among Omnibus Measures for Detection Rates

Statistic	E_{pooled}	E_{weighted}	E_{χ^2}	D_{pooled}	D_{weighted}	D_{χ^2}
E_{pooled}	—					
E_{weighted}	0.987	—				
E_{χ^2}	0.957	0.962	—			
D_{pooled}	0.997	0.981	0.952	—		
D_{weighted}	0.991	0.994	0.958	0.991	—	
D_{χ^2}	0.971	0.965	0.989	0.973	0.974	—

Table 14. d -Statistics for Comparison of Conditional Measures

Statistic	E_{pooled}	E_{weighted}	E_{χ^2}	D_{pooled}	D_{weighted}	D_{χ^2}
E_{pooled}	—					
E_{weighted}	−0.012	—				
E_{χ^2}	0.047	0.058	—			
D_{pooled}	0.010	0.022	−0.037	—		
D_{weighted}	0.006	0.018	−0.041	−0.003	—	
D_{χ^2}	0.067	0.079	0.021	0.057	0.061	—

Note. Positive entries indicate that the method listed in the column outperformed the method in the row. Boldface entries differ significantly from 0.0.

Discussion

This work compared methods of analyzing rescore data. Results for the trend analyses support the findings in Donoghue et al. (2022) that treating the rescore table as a two-way contingency table can yield very misleading results. Type I error was adequately controlled only when the rescore design was proportional to the Occasion A margins. As noted earlier, there may be good reasons to deviate from a strictly proportional design. This is especially true if some categories have a low proportion of responses. It may be critical to have sufficient numbers of responses in these categories to diagnose errors for retraining if the scoring is amiss. As an example, the National Assessment of Educational Progress uses a mixed design similar to that used in this study. Part of the reason for the design is to ensure sufficient instances to diagnose and remediate problems in applying the rubric even in the presence of unpopular categories.

The results for the E -statistics largely replicate the results in Donoghue and Eckerly (2024). The statistics have well-controlled Type I error behavior and good power. The novel contribution of this report is the use of Cliff's (1993) d in the context of monitoring across occasion trend scoring. When analyzing the rescore table as a two-way contingency table, the

paired d -statistics showed the same poor control of Type I error seen for the t -test. When appropriately analyzing the data using conditional analysis (E -statistics and D -statistics), the Type I error rate was well controlled regardless of the trend study design.

In comparing the E -statistics to the D -statistics, the results are very similar. The ordinal D -statistics exhibited slightly less power than the E -statistics. On the basis of the comparison of the methods, the best overall method was E_{weighted} . The advantage was not large; better results were obtained in 1.2%–7.9% of the data sets. The behaviors of E_{χ^2} and D_{χ^2} showed similar patterns. Overall, the Type I error rates of the indexes were quite conservative. However, they were also the only indexes to show inflated (>0.075) Type I error rates. Modified versions of these statistics may yield better (less conservative) Type I error rates and a corresponding increase in power.

The simulation contained in this report was large. However, there remain some important limitations. The most important limitation is the IRT models used to generate the data. The IRT d -parameters were equally spaced, which tended to create symmetric marginal distributions of scores, especially for the balanced design. This may have advantaged the E -statistics indexes based on the t -test, as Feng and Cliff (2004) found that the d -statistic showed more advantage over the t -test when the distributions of the two groups differed in shape as well as in location. It would be useful to extend this work to asymmetric IRT d -parameters. Another interesting option would be to use probabilities based on empirical rescore tables. The challenge in such an approach is how to manipulate the shift in difficulty.

Conclusion

This report has two main takeaways. The most important is that treating an across-occasions rescore table as a two-way contingency table derived from multinomial sampling can lead to very misleading results and so should not be done. Instead, score monitoring needs to acknowledge the product-multinomial sampling of the rescore table and monitor based on the conditional probabilities that are invariant to the rescore design–specified marginal distribution. To provide a meaningful comparison, the within–Occasion A rescore data need to be utilized. Using this information allows for defensible tests for each score category and for the computation of omnibus statistics with accurate Type I error control and good power to detect scoring drift when it occurs.

This study demonstrates that appropriate ordinal measures can function well in rater monitoring. The omnibus measures displayed good Type I error rate controls across the rescore designs. The omnibus measures were also powerful in detecting rater drift, especially in changes in rater severity.

From one view, the similarity of the results for the D -statistics with those of the E -statistics can be seen as giving little reason to shift from t -test-based measures. The other view is that there is little reason to use inappropriate t -tests. The d -statistics match the ordinal nature of CR scores. Second, the results show little to no loss of power to detect misfit. Finally, the d -statistic has a natural use as an effect size: what proportion of Occasion B scores were higher than Occasion A scores, as opposed to the opposite.

Results of this study inform practice for monitoring trend scoring. This study gives concrete guidance for the best way to design and analyze trend rescore studies. CR scoring is expensive, and changes in scoring can result in biased estimates of Occasion A to Occasion B change. In extreme cases, it may necessitate treating the item as separate in the two assessments or even not using (“dropping”) it at Occasion B. Assuming that rescored responses are representative, dependent sampling has the potential to improve monitoring. The E -statistics and D -statistics maintained good Type I error control and showed good power regardless of the rescore design, making them useful in this setting.

References

- Agresti, A. (1983). Testing marginal homogeneity for ordinal categorical variables. *Biometrics*, 39(2), 505–150. <https://doi.org/10.2307/2531022>
- Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43(244), 572–574. <https://doi.org/10.1080/01621459.1948.10483284>
- Clayton, D. G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika*, 61(3), 525–531. <https://doi.org/10.1093/biomet/61.3.525>
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114(3), 494–509. <https://doi.org/10.1037/0033-2909.114.3.494>
- Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31(3), 331–350. https://doi.org/10.1207/s15327906mbr3103_4

- Cliff, N. (1996b). *Ordinal methods for behavioral data analysis*. Erlbaum.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>
- Donoghue, J. R., & Eckerly, C. (2024). New tests of rater drift in trend scoring. *Applied Measurement in Education*, 37(3), 225–239. <https://doi.org/10.1080/08957347.2024.2386935>
- Donoghue, J. R., McClellan, C. A., & Hess, M. R. (2022). *Investigating constructed response scoring over time: The effects of study design on trend rescore statistics* (Research Report No. RR-22-15). ETS. <https://doi.org/10.1002/ets2.12360>
- Feinberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). MIT Press.
- Feng, D., & Cliff, N. (2004). Monte Carlo evaluation of ordinal d with improved confidence interval. *Journal of Modern Applied Statistical Methods*, 3(2), 322–332. <https://doi.org/10.22237/jmasm/1099267560>
- Livingston, S. A. (2009, September). *Constructed-response test questions: Why we use them; how we score them* (ETS R&D Connections No. 11). ETS. https://www.ets.org/Media/Research/pdf/RD_Connections11.pdf
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. SAGE.
- McClellan, C. A., Donoghue, J. R., & Gladkova, L. (2023). *Properties of three statistics used to monitor the quality of constructed-response scoring across occasions* (Research Memorandum No. RM-23-09). ETS.
- McCullagh, P. (1977). A logistic model for paired comparisons with ordered categorical data. *Biometrika*, 64(3), 449–453. <https://doi.org/10.2307/2345320>
- McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42(2), 109–142. <https://doi.org/10.1111/j.2517-6161.1980.tb01109.x>
- R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Sgammato, A., & Donoghue, J. R. (2018). On the performance of the marginal homogeneity test to detect rater drift. *Applied Psychological Measurement*, 42(4), 307–320.

<https://doi.org/10.1177/0146621617730390>

Stuart, A. (1955). A test for homogeneity of marginal distributions in a two-way classification.

Biometrika, 42(3/4), 412–416. <https://doi.org/10.1093/biomet/42.3-4.412>

Appendix

Results for Trend Analysis of Stuart's Q

This appendix presents results for the measure of marginal homogeneity, Stuart's (1955) Q -statistic. Because the measure requires paired data, it could be computed only for trend analysis. In trend analysis, the scores are paired. A common test to determine if scores at Occasion B are lower or higher than at Occasion A is a paired t -test. More recently, Sgammato and Donoghue (2018) suggested using Stuart's (1955) Q -statistic in place of the paired t -test:

$$Q = \mathbf{d}'\mathbf{V}^{-1}\mathbf{d}, \quad (\text{A1})$$

where \mathbf{d} is the vector of differences in marginal proportions and \mathbf{V} is the covariance matrix obtained under the assumption that the two sets of margins are identical (marginal homogeneity). Sgammato and Donoghue found that Q was more powerful than the paired t -test in some conditions, whereas there were very few cases in which the observed t -test was significant and Q was not. They therefore recommended use of Q instead of the paired t -test.

Because the Occasion A data are distinct from the Occasion B data, there is no information to estimate the covariance matrix \mathbf{V} . Therefore Q cannot be applied to the conditional analysis data.

Trend Analysis Using Q

Table A1 presents the ANOVA results for Q .

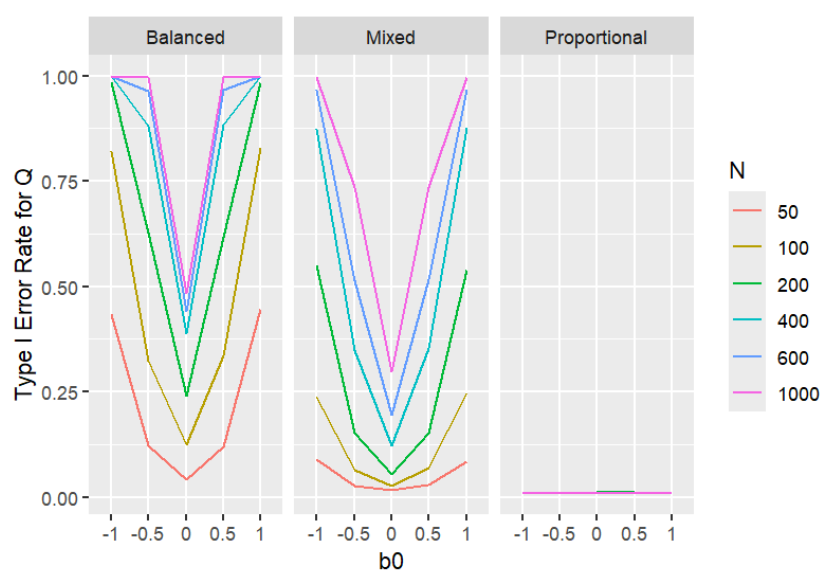
The D (Rescore Design $\times B$ [b_0] $\times N$ [number of cases]) was identified as salient, as were each of the constituent main effects and two-way interactions. None of the other two-way or three-way interactions were identified as salient. Figure A1 shows the three-way interaction.

As Figure A1 makes clear, the Type I error rate is grossly inflated for the balanced design, while it is well controlled for the proportional design. As would be expected, the mixed design falls between these two extremes but generally displays inflated Type I error. One unexpected feature of Figure A1 is that, especially for larger samples, the Q -statistic remains sensitive, incorrectly flagging results at a rate higher than the nominal Type I error rate. This was unexpected, as the statistic shows excellent control for the proportional design. Thus the finding warrants further study in the future.

Table A1. Selected Analysis of Variance Results of Type I Error Rate for Q

Variable	ANOVA SS	η^2
M	2.08	0.004
K	5.26	0.010
N	83.74	0.157
D	230.01	0.432
B	57.57	0.108
A	0.35	0.001
M \times N	0.64	0.001
M \times D	1.39	0.003
M \times B	2.40	0.005
M \times A	0.59	0.001
K \times N	0.79	0.001
K \times D	2.57	0.005
K \times B	3.06	0.006
K \times A	3.85	0.007
N \times D	46.32	0.087
N \times B	7.14	0.013
N \times A	0.04	0.000
D \times B	31.20	0.059
D \times A	0.30	0.001
B \times A	0.05	0.000
N \times D \times B	14.46	0.027
Total	532.76	

Note. Effect sizes $\eta^2 \geq .01$ are in boldface. ANOVA SS = analysis of variance sum of squares; B = b_0 ; D = rescore design; K = number of response categories; M = item response theory model; N = number of cases.

Figure A1. Type I Error Rates for Stuart's Q as a Function of Design, b_0 , and N 

Power Under the Proportional Design

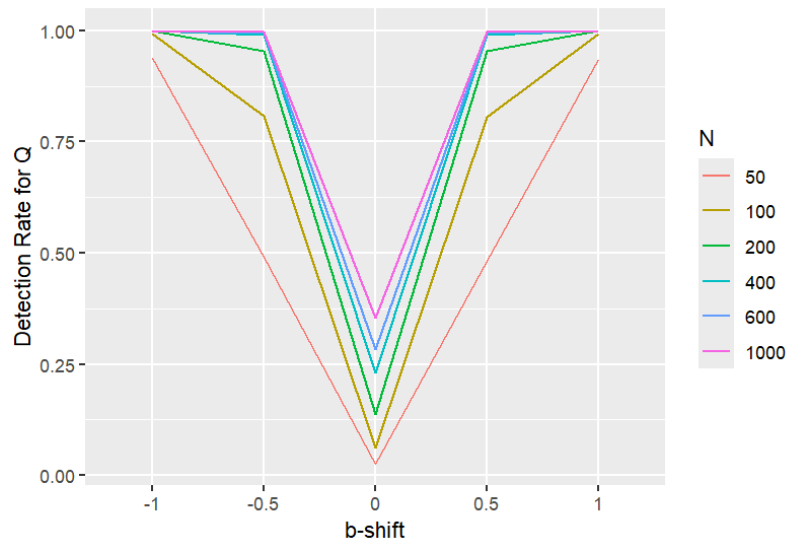
As was found for the other measures when using trend analysis, Q exhibited inflated Type I error for the balanced design and the mixed design. Analysis of the detection of a true difference was restricted to the proportional design. Table A2 presents an ANOVA for the detection rates.

Table A2. Selected Analysis of Variance Results for Q for Detection (Power), Proportional Design Only

Variable	df	ANOVA SS	η^2
M	2	5.97	0.002
K	3	3.17	0.001
N	5	262.96	0.086
B	4	1.29	0.000
Db	4	2028.49	0.661
A	4	4.66	0.002
Aalt	4	5.62	0.002
M \times N	10	33.13	0.011
M \times B	8	0.08	0.000
M \times Db	8	112.44	0.037
M \times A	8	6.33	0.002
M \times Aalt	8	6.30	0.002
K \times N	15	1.51	0.000
K \times B	12	0.08	0.000
K \times Db	12	5.06	0.002
K \times A	12	0.97	0.000
K \times Aalt	12	3.27	0.001
N \times Db	20	156.69	0.051
N \times A	20	20.33	0.007
N \times Aalt	20	14.38	0.005
B \times Db	16	9.88	0.003
B \times A	16	0.02	0.000
B \times Aalt	16	0.14	0.000
Db \times A	16	28.39	0.009
Db \times Aalt	16	27.16	0.009
A \times Aalt	16	86.92	0.028
Total	25,199	3067.5	

Note. Effect sizes $\eta^2 \geq .01$ are in boldface. A = a_0 ; Aalt = a_{alt} ; ANOVA = analysis of variance; B = b_0 ; D = rescore design; Db = b_{shift} ; K = number of response categories; M = item response theory model; N = number of cases.

Compared to Table A1, design is not present, as it is held constant for this analysis. On the other hand, both b_{shift} and a_{alt} are now included in the analysis. Again, an effect size criterion of $\eta^2 \geq .01$ is adopted. For this analysis, N , b_{shift} , and their interaction are identified as salient. Figure A2 gives the means for this interaction.

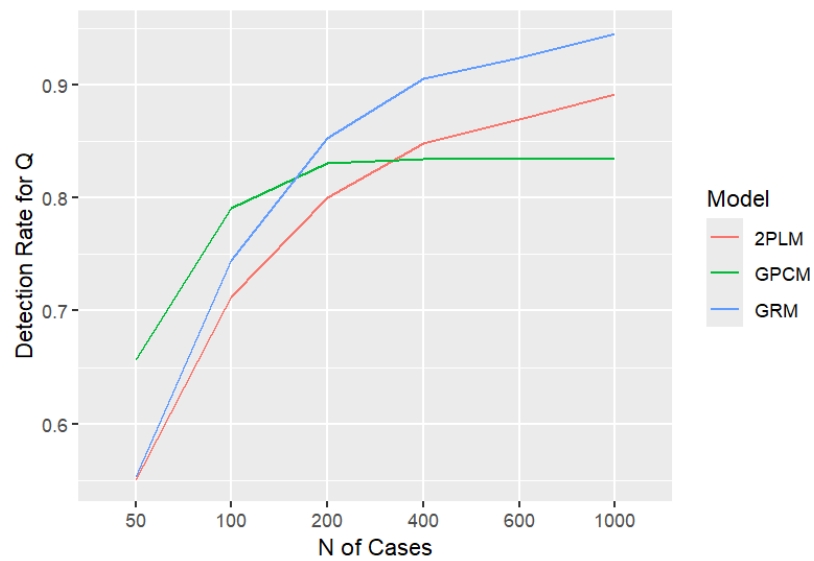
Figure A2. Detection Rates for Stuart's Q by b_{shift} and Sample Size N 

The interaction of a_0 and a_{alt} also meets the effect size criterion. Table A3 gives the mean detection rates for this interaction. As for t and d_{bw} , there is a higher mean detection for Q for cases in which $a_0 = a_{\text{alt}}$. In this case, $b_{\text{shift}} \neq 0$, and the statistic is sensitive to this difference.

Table A3. Mean Detection Rate Q -Statistic as a Function of a_0 and a_{alt}

Value a_{alt}	a_0				
	0.7	1	1.3	1.5	2
0.7	0.84	0.73	0.78	0.79	0.81
1	0.74	0.91	0.75	0.78	0.81
1.3	0.78	0.76	0.95	0.77	0.79
1.5	0.80	0.78	0.77	0.96	0.78
2	0.81	0.81	0.79	0.78	0.98

One additional effect met the effect size criterion: the interaction of number of cases N with the IRT model. The means for this are shown in Figure A3. Detection for data generated by the GPCM model levels off at 200, while the other models show improvement up through a sample size of 1,000.

Figure A3. Mean Q -Statistic for Model by Number of Cases N 

¹ The d -statistic is described as ordinal rather than nonparametric. There is a population parameter δ that is being estimated.

² The actual distribution of E_{χ^2} is likely to be much more complicated. However, in this report, we are interested in how well the approximation works.

³ Results for trend analysis using Stuart's (1955) Q -statistic are included in the appendix.

⁴ The analysis was repeated for each of the proportional and balanced designs (only a single rescore design's data could be analyzed at a time).

Suggested Citation:

Donoghue, J. R., & Sgammato, A. (2025). *Using ordinal rescore measures to monitor rater drift* (Research Report No. RR-25-15). ETS. <https://doi.org/10.64634/fgv13811>

Action Editor: Usama Ali

Reviewers: Jodi Casabianca-Marshall and Carol Eckerly

ETS and the ETS logo are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the [ETS ReSEARCHER](#) database.

