ETS Research Report Series ISSN 2330-8516
MANUSCRIPT IN PRESS 6/18/2025

Using Ordinal Rescore Measures to Monitor Rater Drift

John R. Donoghue & Adrienne Sgammato

ETS Research Institute, ETS, Princeton, New Jersey, United States

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 re-scored 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/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 is effective. Omnibus measures based on combining the individual *t*-tests/*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, v. $D_{weighted}$, 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, Email: 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 paper 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 to not use ("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 are of interest (i.e., are occasion B scores 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 was the chief feature of interest, one would use a measure such as Cohen's (1960) kappa or weighted kappa (1973).

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) an equal number from each of the response categories, b) select responses proportionate to the occasion A distribution, or c) a mixture of the two, such as 50% equal distribution and 50% proportional. We will refer to this planned distribution of occasion A responses as the "rescore design."

When papers 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) show

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 suggest three omnibus *E*-statistic (made by combining the individual *t*-tests). These measures had good Type I error rate and power and were not subject to the bias seen in the paired *t*-test.

One weakness of that work is that the *t*-statistic treats the values as interval score, whereas the reality is that CR scores are only ordinal indicators of the underlying response quality. More recently, Sgammato and Donoghue (2018) recommended using Stuarts's (1955) *Q* statistic for marginal homogeneity, which treats the margins as nominal. Bowker (1948), Clayton (1974) and Agresti (1983) demonstrate tests of marginal homogeneity for ordinal measures, and other formed of regression analyses (e.g., Long, 1997). Other measures such as the Mann-Whitney *U* or Cliff's (1993) *d*-statistic, 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, 1996ab) notes that many times the ordinal statistics align directly with the research question: "Are occasion B scores higher than occasion A?" Unfortunately, these measures fail to reflect the product multinomial sampling in rescore tables.

The purpose of the current study is 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*-statistic) in the evaluation of 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, versus the opposite?

The rest of the paper is organized as follows: First Cliff's *d*-statistic is introduced in the general case of comparing two independent groups and then extended to the within-subjects case. Next, the paper examines 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 paper finishes with discussion and concluding remarks.

Cliff's *d*-Statistic

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 ordinal[1] *d*-statistic is defined:

$$d = \frac{\#\left(X > Y\right) - \#\left(Y > X\right)}{mn} \tag{1}$$

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

$$d_{ij} = sign\left(x_i - y_j\right). \tag{2}$$

which can be arranged into a matrix as shown in Figure 1. Note that the entries indicate whether the row value is larger than the column entry. It is also useful to define the marginal proportion:

$$d_{i\bullet} = \frac{\sum_{j=1}^{m} d_{ij}}{m} \tag{3}$$

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

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

$$d = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}}{mn} \tag{4}$$

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

The standard error of *d* can is given as:

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

Figure 1. Matrix of Dominance Relations

		y-scores					
		1	3	4	7	8	*d.*
x-scores	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
		1	1	1	0.25	0	0.65

Adapted from Table 1: Illustration of Independent Groups Dominance Analysis, "Dominance Statistics: Ordinal Analyses to Answer Ordinal Questions," by N. Cliff, 1993, *Quantitative Methods in Psychology, 114*(3), p. 500. Copyright 1993 by the American Psychological Association, Inc.

Cliff (1993) also presented a paired version of the *d*-statistic, along with associated standard error to facilitate significance testing and 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.

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

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

$$d_b = \frac{\sum\sum_{j\neq i} d_{ij}}{n\left(n-1\right)}\ ,\qquad s_{d_w}^2 = \frac{s_{ij}^2 + \mathrm{cov}\left(d_{ij},d_{ji}\right) + \left(n-2\right)\left(s_{d_{i\cdot}}^2 + s_{d_{\cdot i}}^2 + 2\,\mathrm{cov}\left(d_{i\cdot},d_{\cdot i}\right)\right)}{n\left(n-2\right)} \tag{7}$$

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

$$d_{bw} = d_b + d_w \ , \qquad s^2_{d_{bw}} = s^2_{d_b} + s^2_{d_w} + 2\,\mathrm{cov}\left(d_w, d_b\right) \tag{8}$$

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

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

Approaches to Analyze Trend Scoring Data

We differentiate two forms of analysis of the cross-occasion data. "Trend analysis" will refer to analyzing the rescore table as if it was a two-way table derived from multinomial sampling. "Conditional analysis" will refer 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 occasion B is a paired *t*-test. For the *d*-statistic, trend analysis uses the paired *d*-statistics listed above (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 equations above. As noted above, the three *d*-statistics ask slightly different questions. For simplicity this paper will focus on d_{wb}. Results for d_w and d_b showed the same patterns and so are not presented in the interest of space.

Table 1.

Example rescore tables with identical conditional row probabilities but differing in trend design illustrating difference in *t*-test and *d*-statistic

Table 1a mean diff = 0, t = 0, d_{wb} = 0, z_{wb} = 0

		Occasion A Score 1	2	
Occasion B Score	1	25	25	50
	2	25	25	50
		50	50	

Table 1b mean diff = 0.4, t = 6.83, d_{wb} = 0.4, z_{wb} = 6.83

		Occasion A Score 1	2	
Occasion B	1	5	5	10

Score	2	45	45	90
		50	50	

Table 1c mean diff = -0,3, t = -4.66, d_{wb} = -0.6, z_{wb} = -4.61

		Occasion A Score		
		1	2	
Occasion B	1	40	40	80
Score	2	10	10	20
		50	50	

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 (see Donoghue & Eckerly 2024 for more detail). 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?" 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 and then divides the two to come up with a test statistic:

$$E_{pooled} = \frac{\sum_{k=0}^{K}\left(\bar{x}_k - \bar{y}_k\right)}{\sqrt{\sum_{k=0}^{K} s^2_{\left(\bar{x}_k - \bar{y}_k\right)}}} \,. \tag{10}$$

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:

$$E_{weighted} = \frac{\sum_{k=0}^{K} w_k \left(\overline{x}_k - \overline{y}_k \right)}{\sqrt{\sum_{k=0}^{K} w_k^2 s^2_{(\overline{x}_k - \overline{y}_k)}}} \tag{11}$$

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

The third statistic, $E_{\chi 2}$, is formed by squaring the individual t-tests:

$$E_{\chi^2} = \sum_{k=0}^{K} t_k^2 \tag{12}$$

Under the assumption that the individual tests are approximately standard normal, $E_{\chi 2}$ 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. Based on this, omnibus measures of *d* were computed, here after referred to a *D*-statistics, that were exact analogs of the *E*-statistics.

$$D_{pooled} = \frac{\sum_{k=0}^{K} d_k}{\sqrt{\sum_{k=0}^{K} s^2_{d_k}}} \tag{13}$$

$$D_{weighted} = \frac{\sum_{k=0}^{K} w_k d_k}{\sqrt{\sum_{k=0}^{K} w_k^2 s^2_{d_k}}} \tag{14}$$

[2] The actual distribution of $E_{\chi 2}$ is likely to be much more complicated. However, in this paper we are interested in how well the approximation works.

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

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 high *a*-parameter coupled with 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 $E_{\chi 2}$ and $D_{\chi 2}$ 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 2. 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 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 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 2. Factors Varied in Simulation

Factor	# Levels	Levels
Number of cases	6	50, 100, 200, 400, 600, 1000
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_{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_{alt} = 0.7, 1.0, 1.3, 1.5, 2.0
Number of score categories and IRT model	7	2 (2PL), 3, 4, 5 (GPCM or GRM)

IRT Models Used

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

$$P(\theta) = \frac{\exp\left(1.7a\left(\theta - b\right)\right)}{1 + \exp\left(1.7a\left(\theta - b\right)\right)} \tag{16}$$

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

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

with

$$P_k^+ = P\left(x \geq k \mid \theta\right) = \frac{\exp\left(1.7a\left(\theta - b + d_k\right)\right)}{1 + \exp\left(1.7a\left(\theta - b + d_k\right)\right)} \tag{18}$$

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

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

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, $b_0 + b_{shift}$. 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 looks like scoring data.

Data Generation Factors

As shown in Table 2, the design contained 7 factors, each with several levels. The factors were fully crossed. The factors were:

- **Number of response categories and IRT Model** (7 levels): The number of response categories was 2, 3, 4, or 5. The 2PLM model was used for two-category data, The remaining 6 levels come from crossing generating model GRM or GCPM with the 3 levels of numbers of categories. Note that the same IRT model was used for all scores of

an item, although (as described below) the item parameters could be different if the item exhibited drift. Holding θ fixed for the two scoring occasions corresponds to the fact that underlying quality of the CRs have 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** (6 values): 50, 100, 200, 400, 600, 1000
- **Occasion A b-parameter** b_0 (5 values): -1, -0.5, 0, 0.5, 1
- **Change in b-parameter from occasion A to occasion B** b_{shift} (5 values): -1, -0.5, 0, 0.5, 1
- **Occasion A a-parameter** a_0 (5 values): 0.7, 1.0, 1.3, 1.5, 2.0
- **Occasion B a-parameter** a_{alt} (5 values): 0.7, 1.0, 1.3, 1.5, 2.0
- **Rescore Design** (3 levels): Balanced, Proportional or Mixed. In the balanced design, an equal number of papers were 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 a $N(0, 1)$ 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, ½ of the papers were selected according to the balanced design and ½ were selected according to the proportional design.

The 7 design factors were crossed to yield 7 x 6 x 5 x 5 x 5 x 5 x 3 (78,750) data generation conditions. 1000 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 is undefined. In a small number of additional conditions, the covariance matrix used in Q was singular preventing its inversion. This was associated with extreme combinations of b_0 and b_{shift}, and high a-parameter values. These replications were replaced until

the full 1000 were obtained for each cell. The proportion of the 1000 values that the statistic was significant was recorded, and these rates are the outcome measures for the study.

Analysis Factors

For trend analysis, the rescore data was treated as a two-way table. Paired *t*-test and Cliff's paired *d*-statistics were computed. For conditional analysis, independent groups *t*-test and Cliff's *d*-statistic were computed were computed separately for each level of occasion A score. One set of values was the within occasion A second score. The other set of values was the occasion B score. Thus, there are 4 analyses of each data set:[3]

- Trend, paired *t*-test
- Trend, Cliff's paired *d*-statistic
- Conditional, *t*-test. Three omnibus measures were considered
  - E_{pooled}
  - $E_{weighted}$
  - $E_{\chi 2}$
- Conditional, Cliff's *d*-statistic. Again, three omnibus version were considered:
  - D_{pooled}
  - $D_{weighted}$
  - $D_{\chi 2}$

Data Generation

For each response, a θ value (representing quality of the response) was drawn from a $N(0,1)$ distribution. Next, using the IRT model, the probability of each response category (conditional on θ) was computed and then summed to form a cumulative distribution. A uniform random number was drawn, and the response category was assigned based upon 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

[3] Results for Trend analysis using Stuart's (1955) *Q* statistic are included in the Appendix.

reached, the θ and response were discarded. Another θ was drawn and associated response generated. This process continued until the number of responses required by the rescore design was obtained. For within occasion A rescores, the same θ was used with the same item parameters. A second uniform random number was drawn and used to assign a within occasion re-score 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 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 uniform number 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 will be presented in two phases. The first will report the analyses of the trend analysis. The second phase will report the results for the conditional analyses.

Trend Analyses

Type I Error

We first examine the Type I error behavior. The data were subset to the 3150 conditions in which the null hypothesis was true: $b_{shift} = 0$ and $a_0 = a_{alt}$. A descriptive 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

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

was adopted. Selected results are given in Table 1. 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. Salient effects are highlighted in bold.

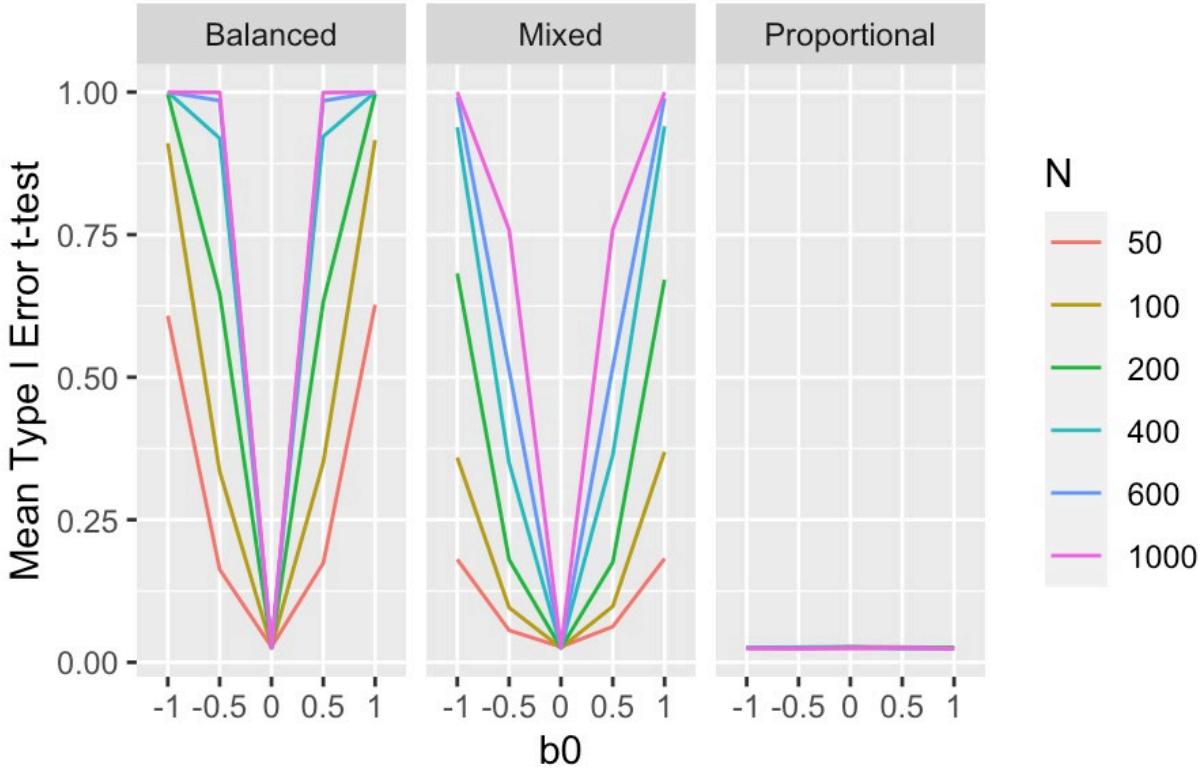
Table 2. Selected ANOVA Results for Paired *t*-Test and d_{wb}

		t-test		d_{wb}	
Variable	df	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*N	10	0.01	0.000	0.04	0.000
M*D	4	0.04	0.000	0.11	0.000
M*B	8	0.02	0.000	0.04	0.000
M*A	8	0.21	0.000	0.13	0.000
K*N	15	0.01	0.000	0.03	0.000
K*D	6	0.05	0.000	0.04	0.000
K*B	12	0.05	0.000	0.03	0.000
K*A	12	0.16	0.000	0.17	0.000
N*D	10	30.09	**0.059**	46.22	**0.108**
N*B	20	16.66	**0.033**	25.69	**0.060**
N*A	20	0.02	0.000	0.04	0.000
D*B	8	65.82	**0.129**	41.58	**0.097**
D*A	8	0.15	0.000	0.12	0.000
B*A	16	0.07	0.000	0.05	0.000
N*D*B	40	20.54	**0.040**	24.20	**0.056**
Total	3149	509.20		429.74	

Note. Effect sizes $\eta^2 \geq 0.01$ are indicated with bold font. D = rescore design, M = IRT model, N = number of cases, K = number of response categories, B = b_0.

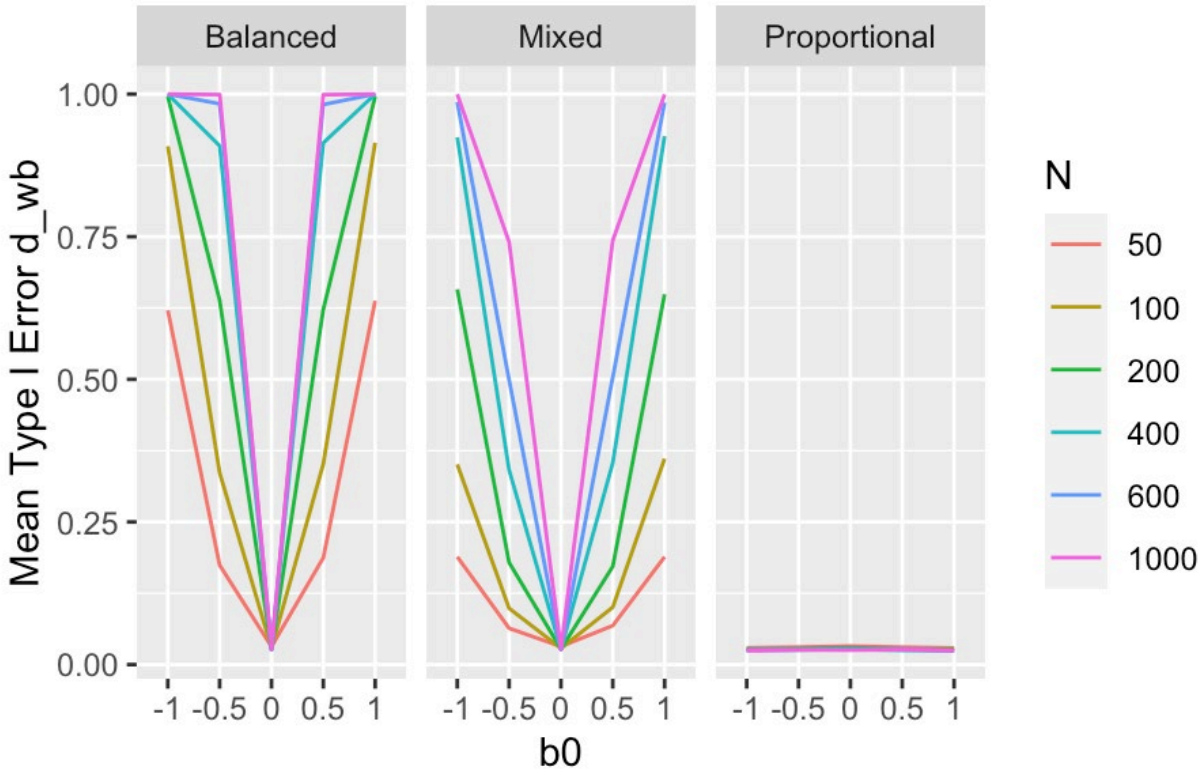
The *D* (rescore design x *B* (b_0) x *N* (number of cases) was identified as salient, as were each of the constituent main effects and two-way interactions. None of the other two-way or three-way interactions were identified as salient. Figures 2 and 3 show the three-way interaction for *t*-test, and the Cliff's paired *d*-statistic respectively. As the figures make clear, 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*-statistic is not immune to these effects. The pattern largely parallels that of the other statistic.

Figure 2. Type I Error Rate for *t*-test as a Function of design, b_0 and *N*



The findings in Figures 2 and 3 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 3. 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 3 gives selected ANOVA results for detection rates for each of the measures.

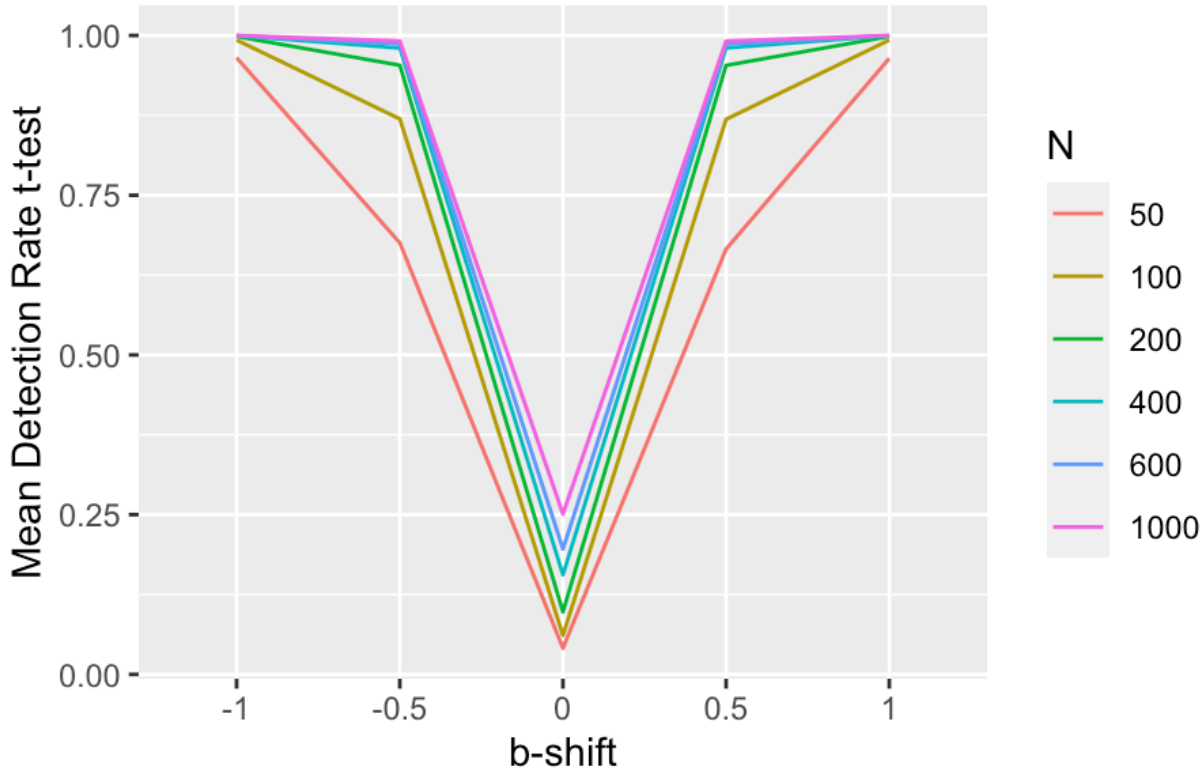
Table 3. Selected ANOVA Results for Detection (Power) Proportional Design Only

		t-test		*z*-d_{wb}	
Variable	df	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	2372.02	**0.781**	2485.28	**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	25199	3038.85		3057.28	

Note. Effect sizes $\eta^2 \geq 0.01$ are indicated with bold font. D = rescore design, M = IRT model, N = number of cases, K = number of response categories, B = b_0, Db = b_{shift}, A = a_0, Aalt=a_{alt}

Compared to Table 2, 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 4 and 5 give this interaction for each of the measures.

Figure 4. Interaction of Sample Size *N* and *b*-shift for *t*-test



There is a U-shaped relationship between b_{shift} and detection which is minimal at $b_{shift} = 0$ and increases as the value diverges from 0. The steepness of the curve is affected by *N*. When $b_{shift} = 0$, 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 non-negligible. Referring to table 3, the interaction of a_0 x a_{alt} was identified as salient. Tables 4 and 5 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 non-zero. Because the measures are more sensitive to the difference in item difficulty, power is relatively good in this condition.

Figure 5. Detection for d_{wb} by b_{shift} and Sample Size

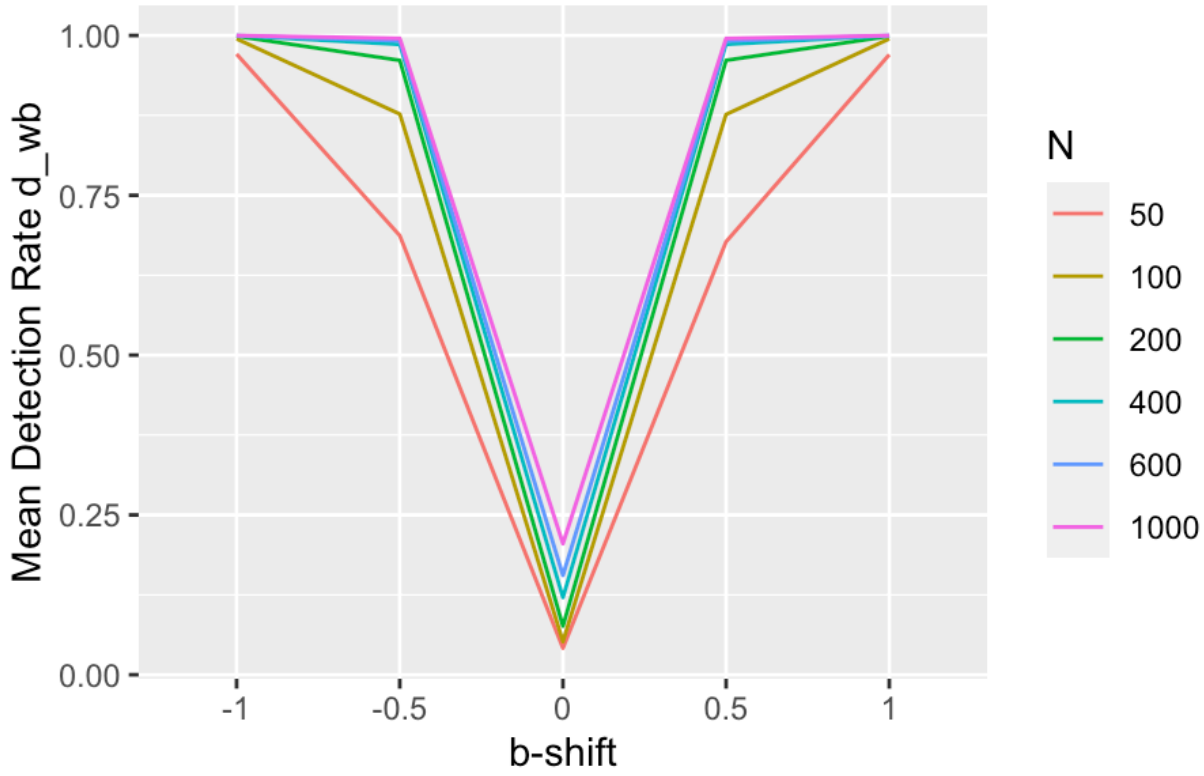


Table 4. Mean Detection Rate *t*-test as function of a_0 and a_{alt}

	a_0				
a_{alt}	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 5. Mean Detection Rate d_{wb} Statistic as a Function of a_0 and a_{alt}

	a_0				
a_{alt}	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

Figures 6 and 7 portray this interaction.

Figure 6. Interaction of IRT Model and b_{shift} for Detection by *t*-test

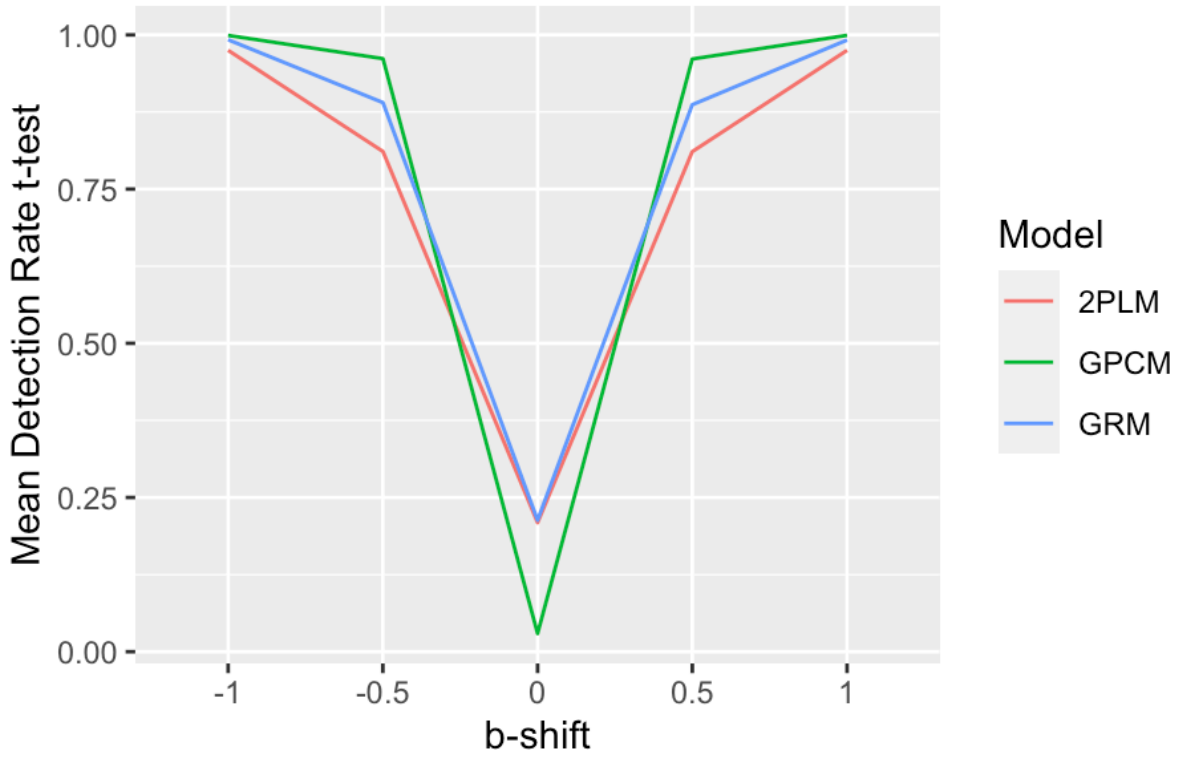
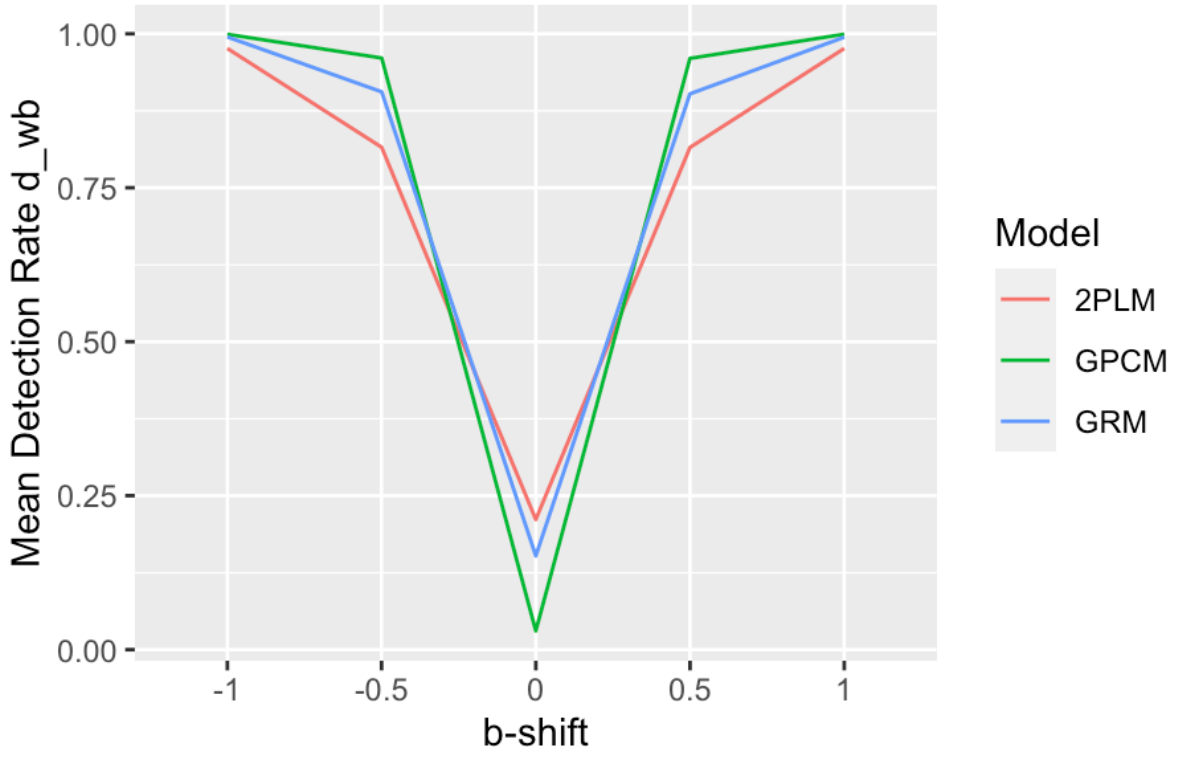
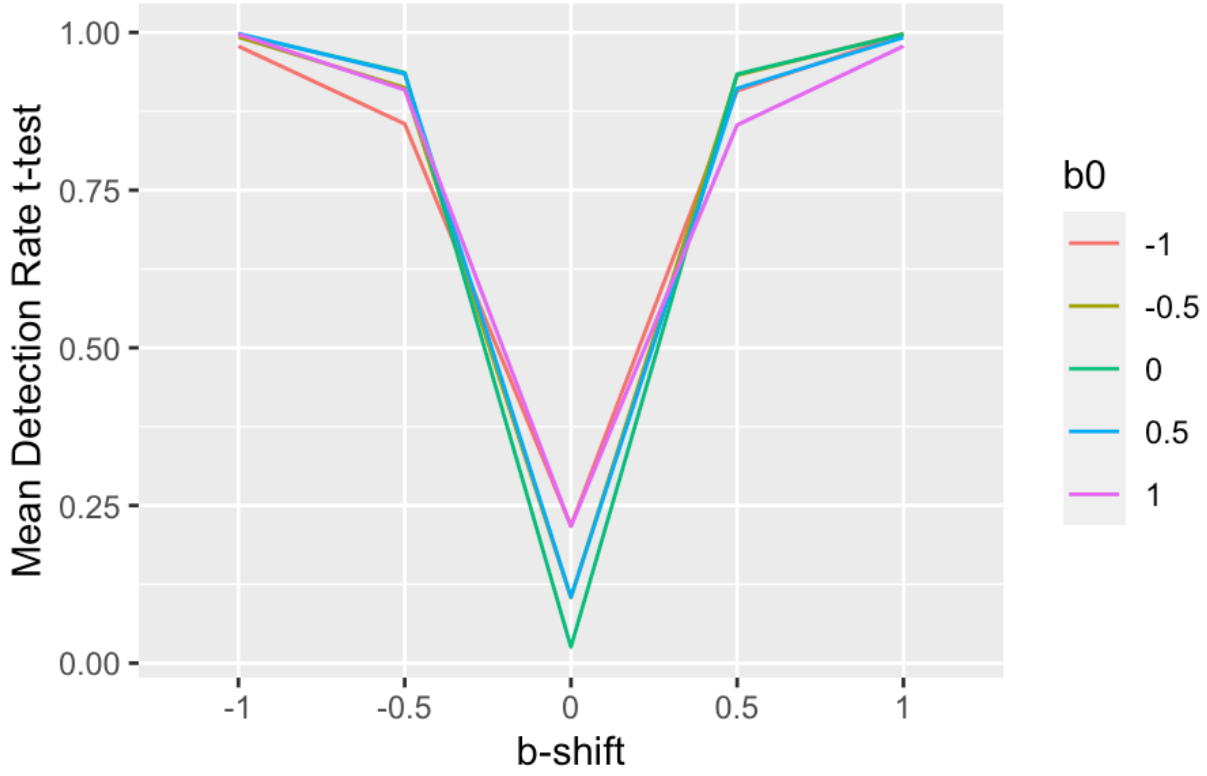


Figure 7. Interaction of IRT Model and b_{shift} for Detection by d_{wb}



One other interaction was flagged for the paired *t*-test. The b_0 by b_{shift} was identified as salient. Figure 8 gives this interaction.

Figure 8. Mean Detection Rate for *t*-test for Interaction of b_0 and b_{shift}



Conditional Analyses

Table 6 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 rate 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 rate, the omnibus $E_{\chi 2}$ and $D_{\chi 2}$ measures were the only ones to demonstrate mild Type I error inflation (0.075-0.10).

Table 6.Type I Error Rate for Conditional Measures by Design

Design	E_{pooled}	$E_{weighted}$	$E_{\chi 2}$	D_{pooled}	$D_{weighted}$	$D_{\chi 2}$
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 9 shows density plots of the Type I error rate. Both measures show conservative Type I error rate, with densities peaking around 0.01. However, both also show marked positive skew. The maximum value for $E_{\chi 2}$ is 0.083 and for $D_{\chi 2}$ it is 0.10. Overall, 14 of 7300 conditions (5/3150 $E_{\chi 2}$ and 9/3150 $D_{\chi 2}$) were found to have elevated Type I error rate. All but two instances were associated with the smallest sample size of 50. Type I inflation for $E_{\chi 2}$ with sample of 50 was also noted by Donoghue and Eckerly (2024).

Figure 9. Density Plot of $D_{\chi 2}$ and $E_{\chi 2}$ for Null Case

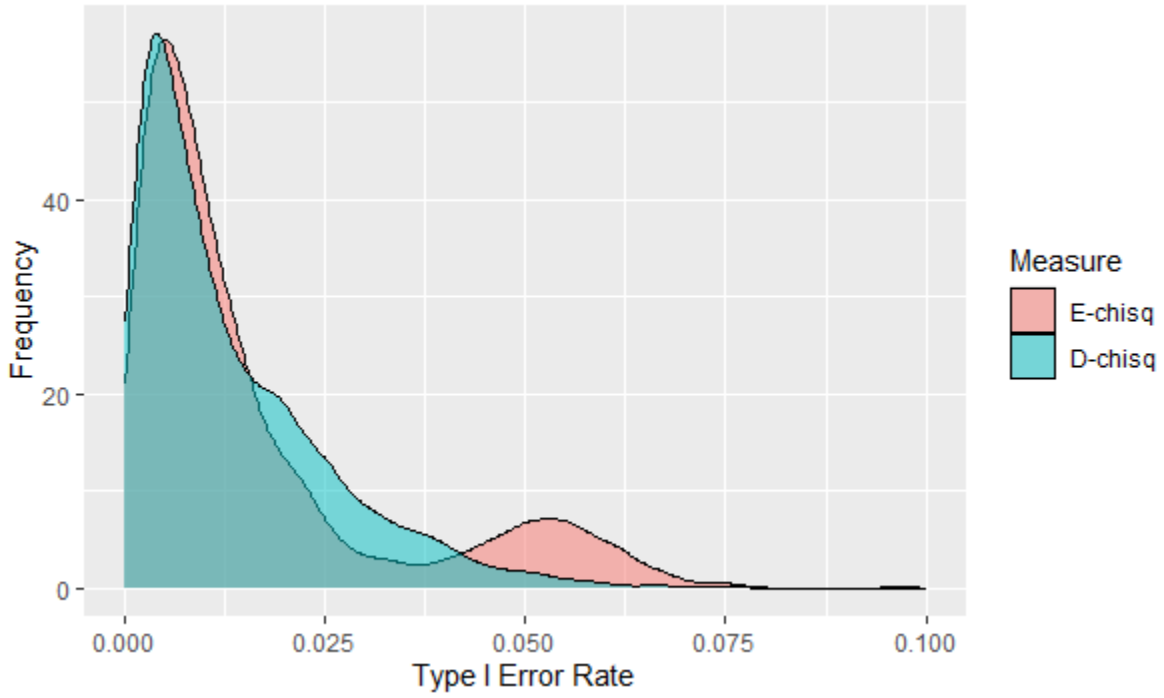


Table 7. Conditions in Which $E_{\chi 2}$ and $D_{\chi 2}$ Measures Showed Elevated Type I Error Rate

Model	N cats	N cases	Design	b_0	a_0	measure	Type_I
2PL	2	100	Mixed	0.5	1	E	0.080
2PL	2	1000	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

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 8 provides summary statistics for the non-null case. All measures demonstrate good detection rates, with the median falling at 1.0 and the first quartile at 0.5 of higher. $E_{\chi 2}$ and $D_{\chi 2}$ 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 8. Overall Detection Rates for Omnibus Measures

	E_{pooled}	$E_{weighted}$	$E_{\chi 2}$	D_{pooled}	$D_{weighted}$	$D_{\chi 2}$
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 above, an effect size criterion of $\eta^2 \geq 0.01$ was adopted. Table 9 shows results for the *E*-statistics, and Table 10 shows results for the *D*-statistics.

Table 9. Selected ANOVA Results Detection for *E*-statistics

		E_{pooled}		$E_{weighted}$		$E_{\chi 2}$	
Source	DF	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**	1100.05	**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	8318.19	**0.828**	7872.32	**0.814**	6789.87	**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*Db	8	65.88	0.007	111.28	**0.012**	186.92	**0.019**
N*Db	20	403.84	**0.04**	320.08	**0.033**	668.48	**0.068**
A*Aalt	16	333.82	**0.033**	328.89	**0.034**	271.69	**0.028**
Total	75599	10044.33		9675.77		9838.81	

Note. Effect sizes $\eta^2 \geq 0.01$ are indicated with bold font. D = rescore design, M = IRT model, N = number of cases, K = number of response categories, B = b_0, Db = b_{shift}, A = a_0, Aalt=a_{alt}

Table 10. Selected ANOVA Results Detection for *D*-statistics

		D_{pooled}		$D_{weighted}$		$D_{\chi 2}$	
Source	DF	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**	1112.71	**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	8490.22	**0.824**	8243.82	**0.813**	7535.07	**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*Db	8	63.8	0.006	95.44	0.009	116.49	**0.011**
N*Db	20	457.03	**0.044**	388.24	**0.038**	733.77	**0.071**
A*Aalt	16	338.88	**0.033**	337.98	**0.033**	288.42	**0.028**
Total	75599	10308.5		10137.8		10392.7	

Note. Effect sizes $\eta^2 \geq 0.01$ are indicated with bold font. D = rescore design, M = IRT model, N = number of cases, K = number of response categories, B = b_0, Db = b_{shift}, A = a_0, Aalt=a_{alt}

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 by 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 10 portrays the means for the number of cases by 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 10. Mean Detection Rate for Interaction of b_{shift} With Sample Size N

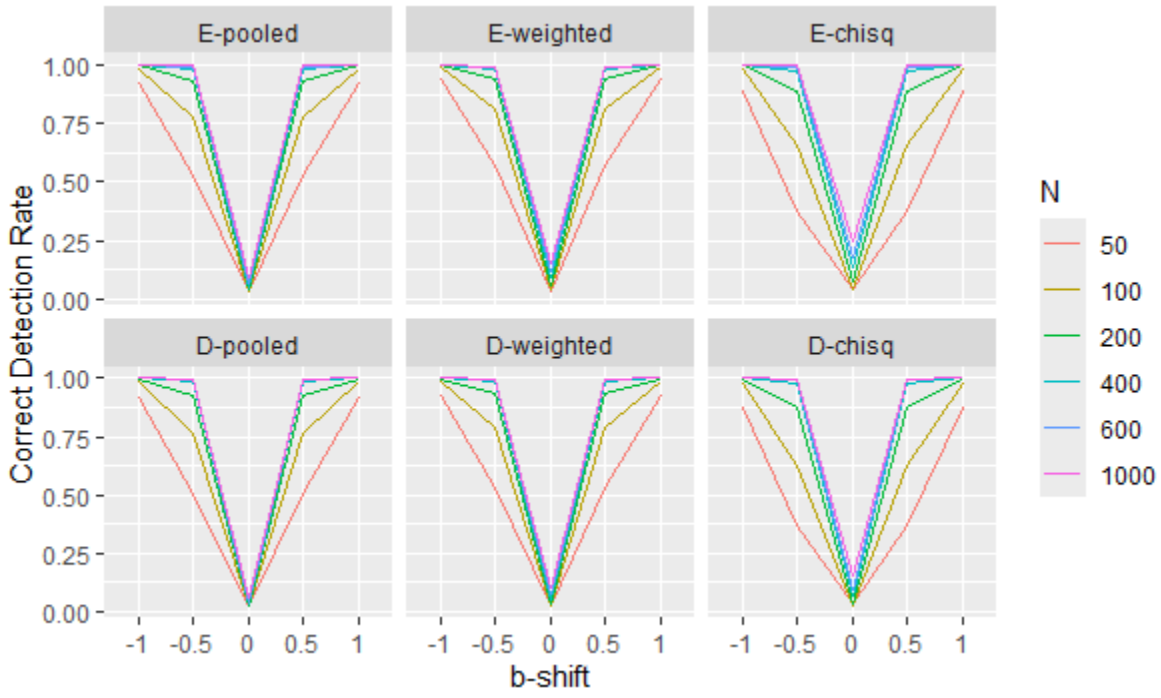
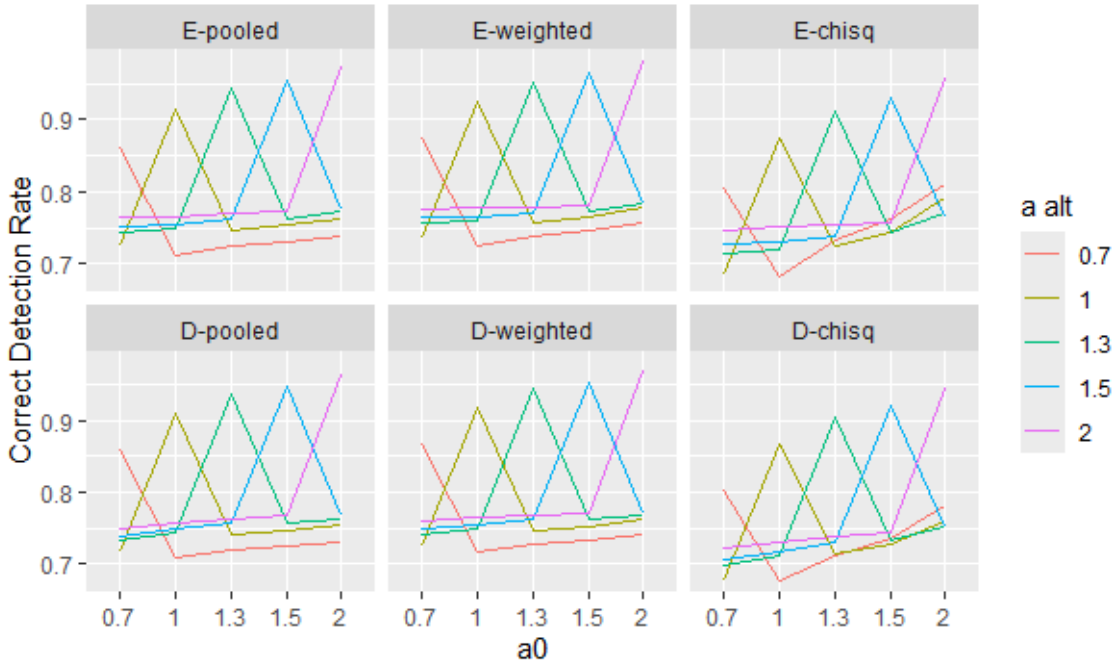


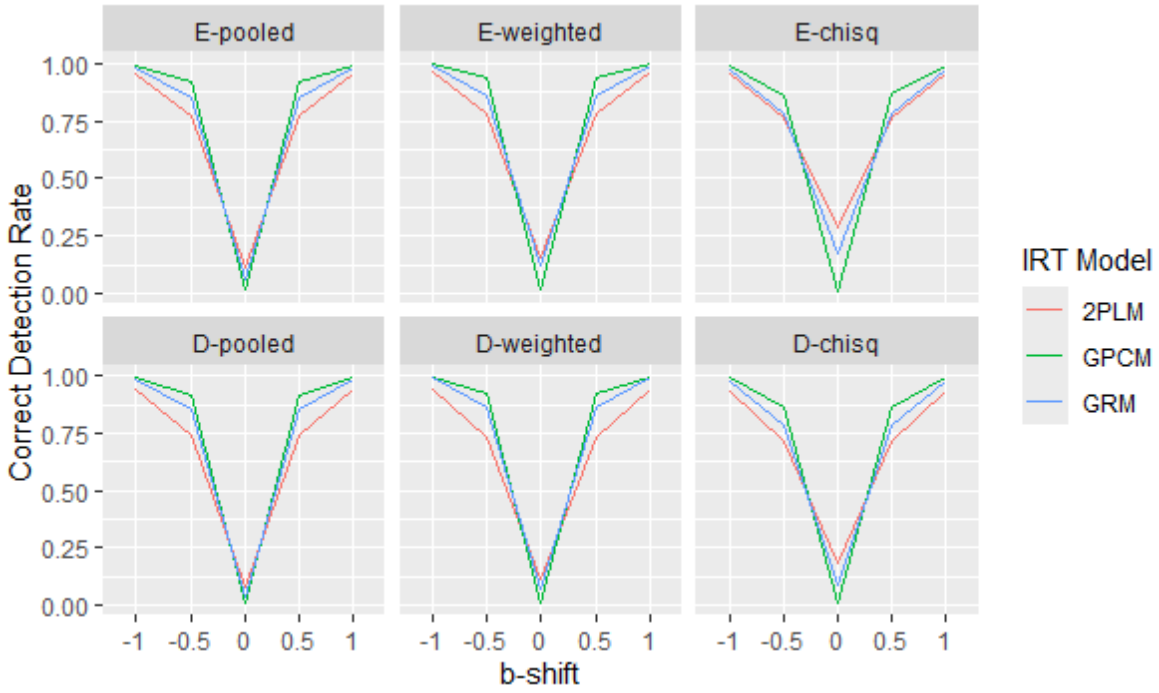
Figure 11 shows the a_0 by a_{alt} interaction. As was noted in the trend results, in all cases there is a spike when $a_0 = a_{alt}$. In this case $b_{shift} \neq 0$; otherwise, it would be a null case. Because the *t*-test and *d*-statistic are sensitive to changes in location, cases when $b_{shift} \neq 0$ 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 11. Mean Detection Rates for Interaction of a_0 and a_{alt}



Finally, Figure 12 shows the interaction of model and b_{shift}. The curves tend to be shallowest for 2PL items, and steepest for GPCM items.

Figure 12. 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 11 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 12. 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 $E_{\chi 2}$ is *d* = 0.021 higher than $D_{\chi 2}$. 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 $E_{\chi 2}$. For *D*-statistics, $D_{weighted}$ was not significantly lower than D_{pooled} (*d* = -0.006) and was significantly higher than $D_{\chi 2}$, *d* = 0.061. Overall, $E_{weighted}$ had the best performance, with being higher than the other methods from 1.2% to 7.9% of the time.

Table 11. Correlations Among Omnibus Measures for Detection Rates

	E_{pooled}	$E_{weighted}$	$E_{\chi 2}$	D_{pooled}	$D_{weighted}$	$D_{\chi 2}$
E_{pooled}	---					
$E_{weighted}$	0.987	---				
$E_{\chi 2}$	0.957	0.962	---			
D_{pooled}	0.997	0.981	0.952	---		
$D_{weighted}$	0.991	0.994	0.958	0.991	---	
$D_{\chi 2}$	0.971	0.965	0.989	0.973	0.974	---

[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).

Table 12. *d*-statistics for Comparison of Conditional Measures

	E_{pooled}	$E_{weighted}$	$E_{\chi 2}$	D_{pooled}	$D_{weighted}$	$D_{\chi 2}$
E_{pooled}	---					
$E_{weighted}$	**-0.012**	---				
$E_{\chi 2}$	**0.047**	**0.058**	---			
D_{pooled}	**0.010**	**0.022**	**-0.037**	---		
$D_{weighted}$	0.006	**0.018**	**-0.041**	-0.003	---	
$D_{\chi 2}$	**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. Bold 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 (NAEP) 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 paper is the use of Cliff's (1993) *d*-statistics 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 (the *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. Based on the comparison of the methods, the best overall method was $E_{weighted}$. The advantage was not large; better results were obtained in 1.2% to 7.9% of the data sets. The behavior of the $E_{\chi 2}$ and $D_{\chi 2}$ showed a similar

pattern to one another. Overall, the Type I error rate of the indexes were quite conservative. However, they were also the only indexes that showed inflated (> 0.075) Type I error rates. Modified versions of these statistics may yield better (less conservative) Type I error rates and corresponding increase in power.

The simulation contained in this paper 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 indices based on the *t*-test, as Feng & Cliff (2004) found that the *d*-statistic showed more advantage over the *t*-test when the distributions of the two groups differed shape as well as 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

There are two main take-aways from the present work. 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 used. Instead, monitoring of scoring 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 defensible tests for each score category, and 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 results for *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*-test. The *d*-statistics match the ordinal nature of CR scores. Second, the results show little/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, 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 – 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*-statistic 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.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the American Statistical Association, 43(*244), 572-574.

Clayton, D. G. (1974). Some odds ratio statistics for the analysis of ordered categorical data. *Biometrika, 61*(3), 525-531.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3) 494-509.

Cliff, N. (1996a). *Ordinal methods for behavioral data analysis*. Erlbaum.

Cliff, N. (1996b). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31*(3) 331-350.

Donoghue, J. R. & Eckerly, C. (2024). New tests of rater drift in trend scoring. *Applied Measurement in Education, 37(3),* 225-239.

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*. (ETS RR–22-15). ETS.

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.

Livingston, S. A. (2009). Constructed-response test questions: Why we use them; How we score them. *ETS R&D Connections, 11*.

Long, J. S. (1997). Regression *models for categorical and limited dependent variables*. age.

McClellan, C. A., Donoghue, J. R., & Gladkova, L. (2023). Properties of three statistics used to monitor the quality of constructed-response scoring across occasions. (ETS Research Memorandum RM-23-09). ETS.

McCullagh, P. (1977). A logistic model for paired comparisons with ordered categorical data. *Biometrika, 64(*3) 449-453.

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.

Stuart, A. (1955). A test for homogeneity of marginal distributions in a two-way classification. *Biometrika, 42*(3/4), 412-416.

Appendix

Results for Trend Analysis of Stuart's *Q*

This appendix presents results for the measure of marginal homogeneity, Stuart's (1959) *Q* statistic. Because the measure requires paired data, it could only be computed 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 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 **d** is the vector of differences in marginal proportions and **V** is the covariance matrix of obtained under the assumption that the two sets of margins are identical (marginal homogeneity). Sgammato and Donoghue found that *Q* was more powerful that the paired *t*-test in some conditions, while there were very few cases where 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 **V**. Therefore, *Q* cannot be applied to the conditional analysis data.

Trend Analysis Using Q

Table A1 presents the ANOVA results for *Q*.

Table A1. Selected ANOVA Results of Type I Error Rate for *Q*

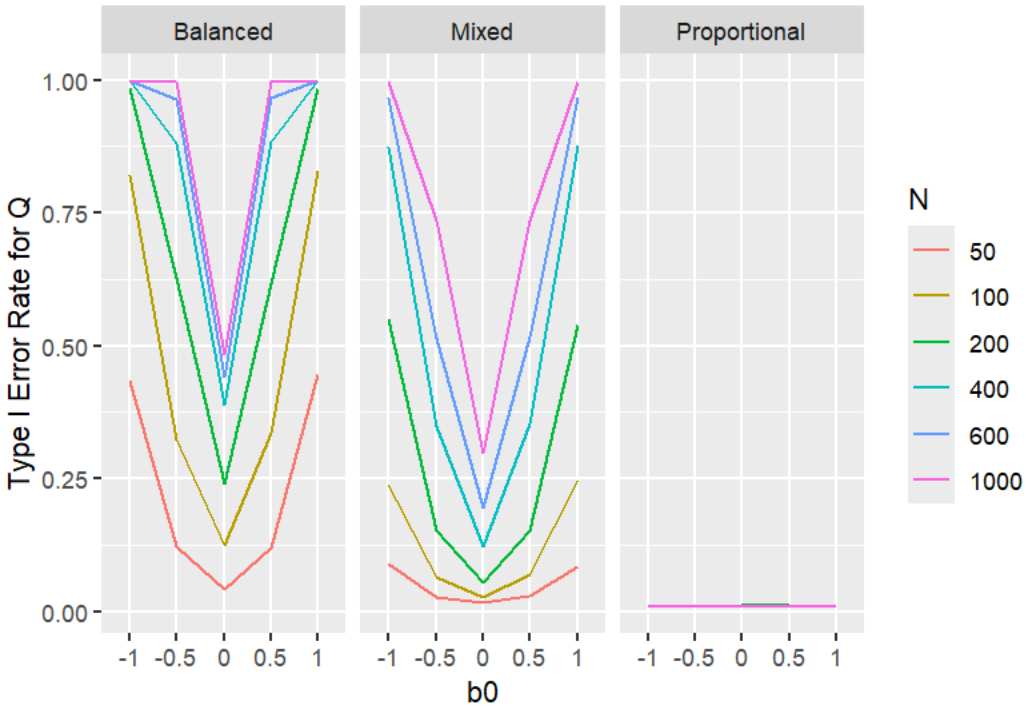
Variable	ANOVA SS	η^2
M	2.08	0.004
K	5.26	0.010
N	83.74	**0.157**
D	230.01	**0.432**
B	57.57	**0.108**
A	0.35	0.001
M*N	0.64	0.001
M*D	1.39	0.003
M*B	2.40	0.005
M*A	0.59	0.001
K*N	0.79	0.001
K*D	2.57	0.005
K*B	3.06	0.006

K*A	3.85	0.007
N*D	46.32	**0.087**
N*B	7.14	**0.013**
N*A	0.04	0.000
D*B	31.20	**0.059**
D*A	0.30	0.001
B*A	0.05	0.000
N*D*B	14.46	**0.027**
Total	532.76	

Note. Effect sizes $\eta^2 \geq 0.01$ are indicated with bold font. D = rescore design, M = IRT model, N = number of cases, K = number of response categories, B = b_0.

The *D* (rescore design x *B* (b_0) x *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.

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



As Figure A1 makes clear, 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 future.

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.

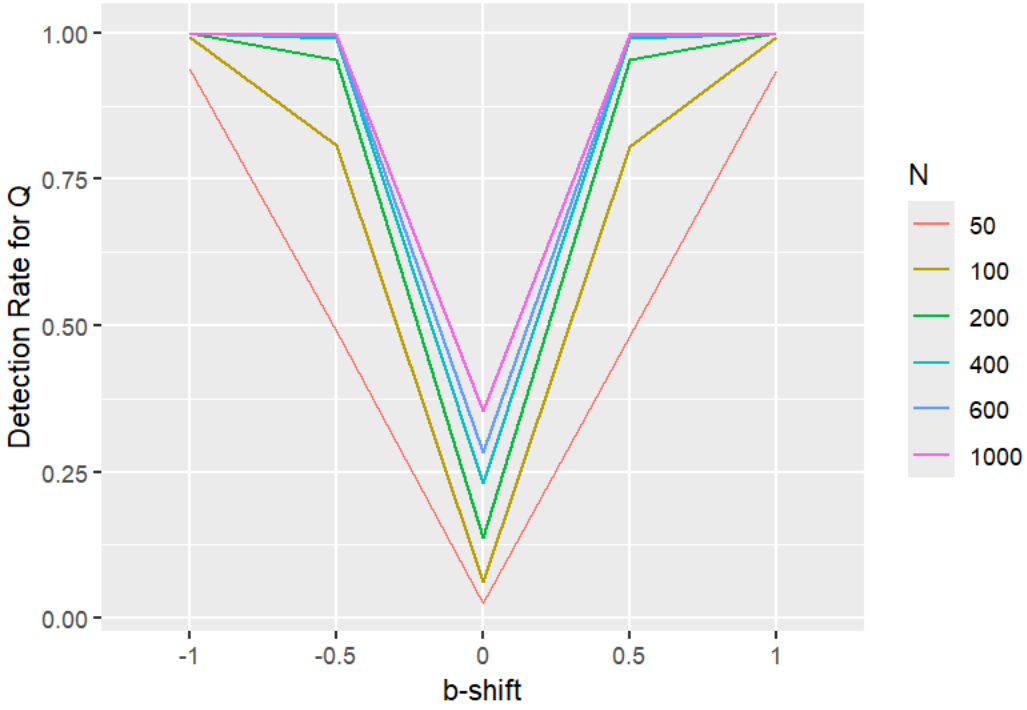
Table A2. Selected ANOVA Results for *Q* for Detection (Power), Proportional Design Only

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*N	10	33.13	**0.011**
M*B	8	0.08	0.000
M*Db	8	112.44	**0.037**
M*A	8	6.33	0.002
M*Aalt	8	6.30	0.002
K*N	15	1.51	0.000
K*B	12	0.08	0.000
K*Db	12	5.06	0.002
K*A	12	0.97	0.000
K*Aalt	12	3.27	0.001
N*Db	20	156.69	**0.051**
N*A	20	20.33	0.007
N*Aalt	20	14.38	0.005
B*Db	16	9.88	0.003
B*A	16	0.02	0.000
B*Aalt	16	0.14	0.000
Db*A	16	28.39	0.009
Db*Aalt	16	27.16	0.009
A*Aalt	16	86.92	**0.028**
Total	25199	3067.5	

Note. Effect sizes $\eta^2 \geq 0.01$ are indicated with bold font. D = rescore design, M = IRT model, N = number of cases, K = number of response categories, B = b_0, Db = b_{shift}, A = a_0, Aalt=a_{alt}

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 $\eta^2 \geq 0.01$ 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*-statistic 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.

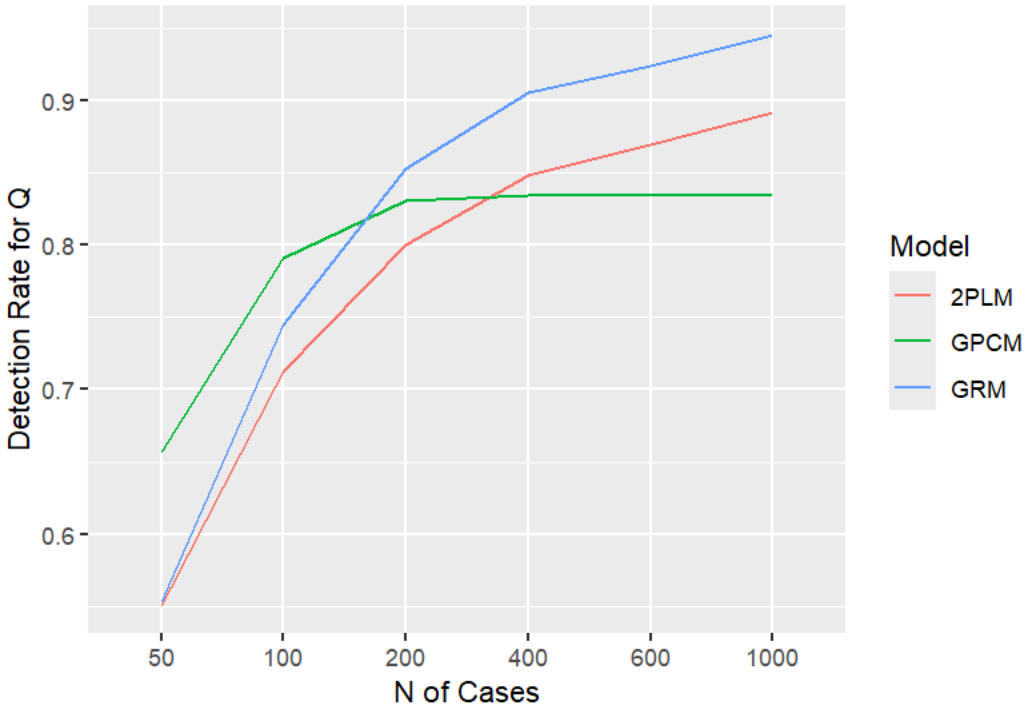
Table A3. Mean Detection Rate *Q* statistic as Function of a_0 and a_{alt}

	a_0				
a_{alt}	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

As for *t*-test and d_{bw}, there is a higher mean detection for Q for cases in which $a_0 = a_{alt}$. In this case, $b_{shift} \neq 0$, and the statistic is sensitive to this difference.

There was one additional effect which met the effect size criterion, the interaction of number of cases *N* with 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 1000.

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



Suggested Citation:
Donoghue, J. R., & Sgammato, A. (in press). *Using ordinal rescore measures to monitor rater drift* (ETS Research Reports). ETS.

Action Editor:
Usama Ali

Reviewers:
Jodi Casabianca-Marshall and Carol Eckerly