Research Memorandum
ETS RM–23-09

Properties of Three Statistics Used to Monitor the Quality of Constructed-Response Scoring Across Occasions

Catherine A. McClellan
John R. Donoghue
Lydia Gladkova

October 2023

ETS Research Memorandum Series

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 Memorandum 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 Memorandum 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.

Properties of Three Statistics Used to Monitor the Quality of Constructed-Response Scoring Across Occasions

Catherine A. McClellan, John R. Donoghue, and Lydia Gladkova
ETS, Princeton, New Jersey, United States

October 2023

Corresponding author: J. R. Donoghue, Email: jdonoghue@ets.org

Suggested citation: 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.

Action Editor: Tim Davey

Reviewers: Adrienne Sgammato and Jodi Casabianca-Marshall

Abstract

When constructed-response items are scored on separate occasions, one issue of interest is whether the scores assigned in the second context/time are equivalent to those assigned in the first. Research has demonstrated that the selection of papers at Occasion A (the rescore design) can have profound effects on the common measures of rescore, the percent exact agreement and the paired *t*-test. To be independent of the rescore design, cross-occasion inference must be based on conditional probabilities (given X_A is the original score for a response and X_B is the second score, $\gamma_{n|m} = P(X_B = n \mid X_A = m)$). The appropriate null hypothesis is that the conditional probabilities $\gamma_{n|m}$ on Occasion A are the same as the conditional probabilities at Occasion B $\eta_{n|m}$ for all values of *n* and *m*.

This manuscript examines three other statistics—Cohen's kappa, quadratically weighted kappa, and the odds-ratio—under the null hypothesis that the scoring is the same on Occasion B as it was on Occasion A $(\boldsymbol{\gamma} = \boldsymbol{\eta})$. Strictly as a function of the rescore design, both kappa and quadratically weighted kappa yield expected values that are higher or lower than computed for the original within-occasion scoring. The odds ratio is shown to not be affected by the rescore design, but it is unclear how best to apply odds ratios for assessment of rater association in tables larger than two by two.

Keywords statistics, scoring, constructed response, occasions, rescore design, Cohen's kappa, quadratically weighted kappa, null hypothesis, odds ratio, monitoring scoring, trend scoring

Acknowledgments

Catherine A. McClellan worked as a psychometrician and research center director at ETS from November 1998 to October 2011. She is currently the deputy chief executive officer (research & assessment) at the Australian Council of Educational Measurement in Camberwell, Victoria, Australia.

John R. Donoghue is a principal research scientist at ETS.

Lydia Gladkova worked as a research scientist, psychometrician at ETS from January 2004 to October 2007. She is currently an artist in Metuchen, New Jersey, United States.

Portions of this work were presented at the annual meeting of the National Council on Measurement in Education in San Diego, California, April 12–16, 2004.

Parts of the work reported herein was supported under the National Assessment of Educational Progress (ED 02-CO-0023) as administered by the Institute of Educational Sciences, U. S. Department of Education.

The authors wish to thank Jodi Casabianca-Marshall, Tim Davey, and Adrienne Sgammato for numerous helpful suggestions.

When constructed-response (CR) items are scored on separate occasions, one issue of interest is whether the scores assigned in the second context/time are sufficiently similar to those assigned in the first. One example of such a comparison is when CR items are used in multiple administrations of a test. Here, concern is whether the scores from raters at Occasion A are comparable to those assigned by raters at Occasion B; do the raters apply the rubric in a consistent manner? Another example is when scores from an automated scoring engine are compared to those assigned by human scorers. For simplicity, we will refer to the first set of scores as Occasion A and the second set as Occasion B, although, as in the automated scoring example, mode rather than time may be the key difference between the two sets of scores.

In each of these situations, the usual approach to monitoring is to select a sample of responses from Occasion A and then have them rescored by Occasion B raters. Usually, the Occasion A scores are chosen according to some guidelines. For example, an equal number of responses from each response category may be sampled. Or the number of scores at each category may be proportional to the overall marginal distribution at Occasion A. We term the choice of distribution of Occasion A responses the *rescore design*, which fixes the distribution of Occasion A scores.

For an *M*-category CR item, the two sets of scores are arranged into an *M* x *M* contingency table. The usual assumption in analyzing such a table is that it follows a multinomial sampling distribution. However, in the context of a rescore study, the Occasion A scores are fixed by the rescore design and the sampling model is no longer multinomial. Instead, the sample for each score category follows a separate multinomial distribution, and sampling for the table is the product of these distributions: a product-multinomial (e.g., Fienberg & Holland, 1970, p. 15). Donoghue et al. (2022) demonstrated that the sampling model can have profound effects on the common measures of rescore agreement, the percent exact agreement and the paired *t*-test (see also Abdalla, 2019). They showed that even when the null hypothesis of equivalence of the scoring process was true, the rescore statistic could be higher than, lower than, or the same as the within-Occasion A value, simply as a function of the rescore design. When the null hypothesis was not true, it was possible to choose the rescore sample with effectively zero power to detect that difference.

The work reported here extends the Donoghue et al. (2022) results to two widely used statistics, Cohen's (1960) kappa and (quadratically) weighted kappa (Cohen, 1968). We also examine the (log) odds ratio and find that it is not sensitive to the rescore design. Although Donoghue et al. relied on numerical examples, the present text will focus more on algebraic relations under the null hypothesis that the scoring is invariant across the occasions.

Within-Occasion Rescore Table

In monitoring scoring within an occasion, it is common to randomly select a sample of the papers and assign them to a second rater to assess agreement. Usually, the second rater is randomly chosen with the stipulation that the paper cannot go to the original rater. In this scenario, individual raters serve the role of Rater 1 for some papers and Rater 2 for others, making Rater 1 a sample of raters from a pool and Rater 2 another sample from the same pool. Because of this interchangeability of raters, the within-occasion rescore table is expected to be symmetric. Table 1 shows a typical rescore table.

Table 1. Example of Within-Occasion Data

Within-year reliability / First score at Occasion A	Second score at Occasion A			Marginal rate
	Score category 0	Score category 1	Score category 2	
Score category 0	f_{00}	f_{01}	f_{02}	f_{2+}
Score category 1	f_{10}	f_{11}	f_{12}	f_{1+}
Score category 2	f_{20}	f_{21}	f_{22}	f_{0+}
Marginal rate	f_{2+}	f_{1+}	f_{0+}	1

Rescore Study

Assume that the distribution of scores for an *M*-category polytomous item given at Occasion A is denoted ***f*** (i.e., $f_0, f_1, \ldots, f_{M-1}, \sum_{k=0}^{M-1} f_k = 1$). A set of papers is chosen from the responses at Occasion A to carry forward to Occasion B. This trend set has a distribution of scores given at Occasion A of ***h*** *(* $h_0, h_1, ..., h_{M-1}, \sum_{k=0}^{M-1} h_k = 1$), where the value of ***h*** is fixed in the rescore design. This differs from monitoring within Occasion A, where the margins of the table are a random variable.

In general, we cannot assume that **h** = **f**. Indeed, there are often reasons to explicitly select a trend set with a distribution of papers (**h**) that differs from the original (**f**). For example, if an item has few responses in the highest category at Occasion A, and item responses are chosen for the trend set in the same proportions as the overall item score distribution, few if any responses in the highest category will be chosen for the trend set. If, on Occasion B, students perform better on the item and there are more responses in the highest score category, it will be difficult to detect errors in scoring this category across occasion, as there is very little data from Occasion A for use in verification. Selecting extra papers in the most extreme score categories, or choosing a uniform distribution of papers across score categories, alleviates this concern, but in that case, it is clear that **f** is not equal to **h**.

Conditional and Marginal Probability

Because **h** may not equal **f**, cross-occasion inference must be based on conditional, rather than marginal probabilities. In other words, given that the paper received a score of 1 from Rater 1, what is the probability that it received a score of 1 from Rater 2? If X_1 is the original score for a response and X_2 is the second score, then the conditional probability we care about is $P(X_{A2} = n \mid X_{A1} = m) = \gamma_{n|m}$. The same quantity at Occasion B (i.e., cross-occasion) in the rescore will be denoted as $P(X_{B2} = n \mid X_{A1} = m) = \eta_{n|m}$. Because they are invariant and do not depend upon the rescore design, conditional probabilities $\gamma_{n|m}$ and $\eta_{n|m}$ can serve as the basis of valid measures of the across-occasion scoring quality. An appropriate null hypothesis is that $\eta_{n|m} = \gamma_{n|m}$ for all values of *n* and *m*.

Cohen's (1960) Kappa

Donoghue et al. (2022) showed that the structure of the cross-occasion table can yield unintuitive results for tests (such as a paired *t*-test) that the rescore table's margins are the same. They also demonstrated that the proportion of exact agreement is affected by the rescore design. An obvious question at this point is does using an alternative measure that adjusts for the marginal frequency, such as Cohen's kappa, encounter the same problems. Cohen (1960) noted that for items with extreme margins (i.e., very hard or very easy items), the percent exact agreement p_A could be high even in the presence of chance-level agreement. He introduced kappa to remove the effect of chance agreement,

$$\kappa = \frac{p_A - p_C}{1 - p_C}, \tag{1}$$

where p_A is the proportion exact agreement as computed from the rescore table, and p_C is the proportion agreement expected due to chance.

Consider a within-occasion rescore two-way table such as Table 2.

Table 2. Example of Within-Occasion Data in Terms of Conditional Probability γ

Within-year reliability/First score at Occasion A	Second score at Occasion A		Marginal rate
	Score category 1	Score category 0	
Score category 1	$\gamma_{1\|1} f$	$\gamma_{0\|1} f$	f
Score category 0	$\gamma_{1\|0}(1-f)$	$\gamma_{0\|0}(1-f)$	$1-f$
Marginal rate	f	$1-f$	1

As noted previously, because of the interchangeability of raters at Occasion A, the within-occasion table is symmetric. When symmetry is assumed,

$$\gamma_{1|0}(1-f) = \gamma_{0|1} f \tag{2}$$

and Table 2 can be rewritten in terms of $\gamma_{1|1}$ and f only, as in Table 3.

Table 3. Example of Within-Occasion Data – Simplified

Within-year reliability/ First score at Occasion A	Second score at Occasion A		Marginal rate
	Score category 1	Score category 0	
Score category 1	$\gamma_{1\|1} f$	$\left(1-\gamma_{1\|1}\right) f$	f
Score category 0	$\left(1-\gamma_{1\|1}\right) f$	$1-f-\left(1-\gamma_{1\|1}\right) f$	$1-f$
Marginal rate	f	$1-f$	1

Based on Table 3, we can determine whether the value of κ is sensitive to the rescore design.

The within-occasion value κ_w is then:

$$\kappa_w = \frac{p_A - p_C}{1 - p_C} \tag{3}$$

$$\kappa_w = \frac{\gamma_{1|1} f + \left[1 - 2f + \gamma_{1|1} f\right] - f^2 - \left(1-f\right)^2}{1 - f^2 - \left(1-f\right)^2} \tag{4}$$

$$\kappa_w = \frac{\gamma_{1|1} - f}{1 - f}. \tag{5}$$

Note that for $\gamma_{1|1} = f$, $\kappa_w > 0$ and the agreement seen in the table are no greater than that expected by chance, while if $\gamma_{1|1} > f, \kappa_w > 0$ demonstrating that kappa is sensitive to the rescore design.

Now consider the cross-occasion calculation of Cohen's kappa under the null hypothesis that $\boldsymbol{\eta} = \boldsymbol{\gamma}$. The formulae to calculate the proportion of scores in each cell for a trend table are shown in Table 4. These proportions have been reformulated so that they are in terms of $\gamma_{1|1}$, ***f*** and ***h*** only, as was done in Table 3.

Table 4. Across-Occasion Rescore Table Under the Null Hypothesis

Occasion A score	Occasion B score		Marginal rate
	1	0	
1	$\gamma_{1\|1} h$	$\left(1-\gamma_{1\|1}\right)h$	*h*
0	$\left(1-\gamma_{1\|1}\right)\left(\frac{f}{1-f}\right)(1-h)$	$\left[1-\left(1-\gamma_{1\|1}\right)\left(\frac{f}{1-f}\right)\right](1-h)$	*1–h*
Marginal rate	$\frac{f(1-h)+\gamma_{1\|1}(h-f)}{1-f}$	$\frac{1-2f+fh+\gamma_{1\|1}(f-h)}{1-f}$	*1*

Under the null hypothesis that $\eta_{1|1}=\gamma_{1|1}$, the population value of p_A is:

$$p_A=\gamma_{1|1}h+\left[1-\left(1-\gamma_{1|1}\right)\left(\frac{f}{1-f}\right)\right](1-h) \tag{6}$$

$$p_A=\frac{\gamma_{1|1}h(1-f)+\left[(1-f)-\left(f-\gamma_{1|1}f\right)\right](1-h)}{1-f} \tag{7}$$

$$p_A=\frac{\gamma_{1|1}h(1-f)+\left(1-2f+\gamma_{1|1}f\right)(1-h)}{1-f} \tag{8}$$

$$p_A=\frac{\gamma_{1|1}h-\gamma_{1|1}fh+1-2f+\gamma_{1|1}f-h+2fh-\gamma_{1|1}fh}{1-f} \tag{9}$$

$$p_A=\frac{\gamma_{1|1}f+\gamma_{1|1}h-2\gamma_{1|1}fh+1-2f-h+2fh}{1-f}. \tag{10}$$

Similarly, the value of p_C is

$$p_C=\frac{\left[f(1-h)+\gamma_{1|1}(h-f)\right]h}{1-f}+\frac{\left[1-2f+fh+\gamma_{1|1}(f-h)\right](1-h)}{1-f} \tag{11}$$

$$p_C=\frac{fh-fh^2+\gamma_{1|1}h^2-\gamma_{1|1}fh+1-h-2f+2fh+fh-fh^2+\gamma_{1|1}f-\gamma_{1|1}h-\gamma_{1|1}fh+\gamma_{1|1}h^2}{1-f} \tag{12}$$

$$p_C=\frac{4fh-2fh^2+2\gamma_{1|1}h^2-2\gamma_{1|1}fh+1-h-2f+\gamma_{1|1}f-\gamma_{1|1}h}{1-f} \tag{13}$$

$$\begin{aligned} p_A - p_C = {} & \frac{\gamma_{1|1} f + \gamma_{1|1} h - 2\gamma_{1|1} fh + 1 - 2f - h + 2fh}{1-f} \\ & - \frac{4fh - 2fh^2 + 2\gamma_{1|1} h^2 - 2\gamma_{1|1} fh + 1 - h - 2f + \gamma_{1|1} f - \gamma_{1|1} h}{1-f} \end{aligned} \tag{14}$$

$$\begin{aligned} p_A - p_C = {} & \frac{\gamma_{1|1} f + \gamma_{1|1} h - 2\gamma_{1|1} fh + 1 - 2f - h + 2fh}{1-f} \\ & + \frac{-4fh + 2fh^2 - 2\gamma_{1|1} h^2 + 2\gamma_{1|1} fh - 1 + h + 2f - \gamma_{1|1} f + \gamma_{1|1} h}{1-f} \end{aligned} \tag{15}$$

$$p_A - p_C = \frac{2\gamma_{1|1} h - 2fh + 2fh^2 - 2\gamma_{1|1} h^2}{1-f} \tag{16}$$

$$p_A - p_C = \frac{2h\left(\gamma_{1|1} - f\right) + 2h^2\left(f - \gamma_{1|1}\right)}{1-f} \tag{17}$$

$$p_A - p_C = \frac{2h\left(1-h\right)\left(\gamma_{1|1} - f\right)}{1-f} \tag{18}$$

$$p_A - p_C = \kappa_w 2h\left(1-h\right). \tag{19}$$

The denominator of the rescore κ_r is

$$denom = 1 - p_C \tag{20}$$

$$denom = 1 - \frac{4fh - 2fh^2 + 2\gamma_{1|1} h^2 - 2\gamma_{1|1} fh + 1 - h - 2f + \gamma_{1|1} f - \gamma_{1|1} h}{1-f} \tag{21}$$

$$denom = \frac{1 - f - 4fh + 2fh^2 - 2\gamma_{1|1} h^2 + 2\gamma_{1|1} fh - 1 + h + 2f - \gamma_{1|1} f + \gamma_{1|1} h}{1-f} \tag{22}$$

$$denom = \frac{-2\gamma_{1|1} h^2 + 2\gamma_{1|1} fh - \gamma_{1|1} f + \gamma_{1|1} h - 4fh + 2fh^2 + h + f}{1-f} \tag{23}$$

$$denom = \frac{-2\gamma_{1|1} h^2 + 2\gamma_{1|1} fh - \gamma_{1|1} f + \gamma_{1|1} h + 2f - 4fh + 2fh^2 + h - f}{1-f} \tag{24}$$

$$denom = \frac{2\gamma_{1|1}h\left(f-h\right)-\gamma_{1|1}\left(f-h\right)+2f\left(1-h\right)^2+\left(h-f\right)}{1-f} \tag{25}$$

$$denom = \frac{2\gamma_{1|1}h\left(f-h\right)-\gamma_{1|1}\left(f-h\right)-\left(f-h\right)+2f\left(1-h\right)^2}{1-f} \tag{26}$$

$$denom = \frac{\left(f-h\right)\left(2\gamma_{1|1}h-\gamma_{1|1}-1\right)+2f\left(1-h\right)^2}{1-f}\,. \tag{27}$$

Therefore, κ_r is

$$\kappa_r = \frac{2h\left(1-h\right)\left(\gamma_{1|1}-f\right)}{\left(f-h\right)\left(2\gamma_{1|1}h-\gamma_{1|1}-1\right)+2f\left(1-h\right)^2}\,. \tag{28}$$

Note that the within-occasion $\kappa_w = \frac{\gamma_{1|1}-f}{1-f}$ yields

$$\kappa_r = \kappa_w \frac{2h\left(1-h\right)\left(1-f\right)}{\left(f-h\right)\left(2\gamma_{1|1}h-\gamma_{1|1}-1\right)+2f\left(1-h\right)^2}\,. \tag{29}$$

It is useful to notice that if *h* = *f*, $\kappa_r = \kappa_w$.

Equation 29 is hard to develop intuition about, so some numerical examples were constructed. To make the examples realistic, the constraints $\gamma_{1|1} \geq 0.5$ and $\gamma_{0|0} \geq 0.5$ were enforced, indicating that for either score, the probability of assigning the correct score is ≥ 0.5. This yielded valid values of *f* in the region $f \in \left[0.2, 0.8\right]$ and likewise $h \in \left[0.2, 0.8\right]$.

Figure 1 shows the results κ_r plotted against κ_w. The null hypothesis is true, so if κ did not depend on the rescore design, all points would fall along the diagonal. In fact, the majority of points are above the line, indicating that in most circumstances, $\kappa_r < \kappa_w$, which might lead the researcher to incorrectly conclude there was a problem with the scoring.

Figure 1. Across-Occasion κ_r (k_r) Plotted Against Within-Occasion κ_w (k_w)

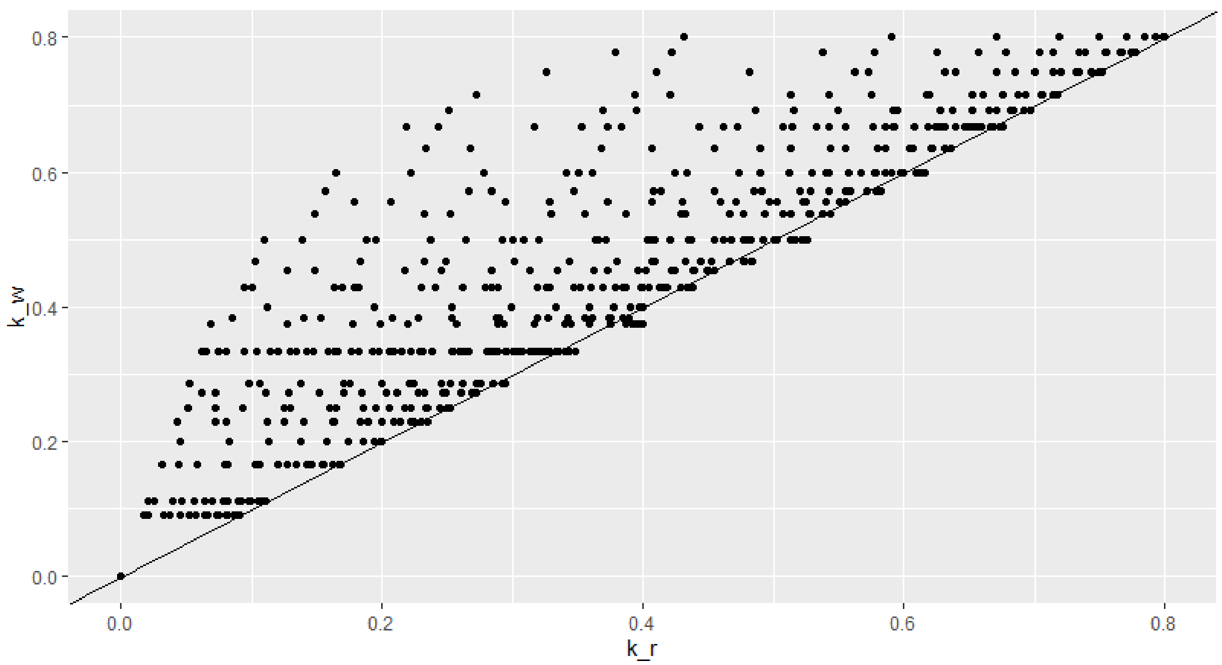
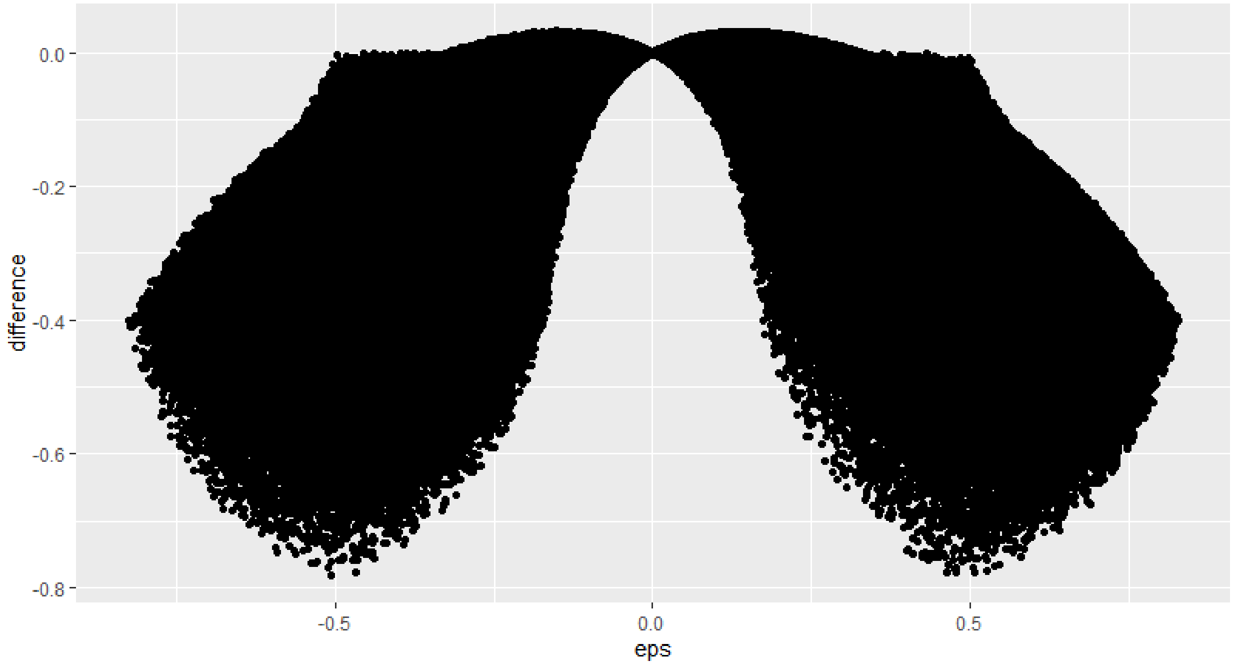


Figure 2 contains a much larger number of points and shows the relationship between the difference in margins for the within-occasion and rescore design $eps = h - f$ and the difference $\kappa_w - \kappa_r$. Relatively few of the plotted points are above 0, and all are less than 0.033. This indicates that when the null hypothesis is true, the researcher is very unlikely to obtain a κ_r value that indicates the scoring is better on Occasion B. On the other hand, the difference can be quite negative, approaching -0.8 in the worst cases, which would likely be seen as strong evidence that the scoring at Occasion B departed from that at Occasion A. It should be re-emphasized that the null is true here $(\eta = \gamma)$, and the differences are solely a function of the rescore design.

Figure 2. Difference $\kappa_w - \kappa_r$ as a Function of $eps = h - f$



Tables 5, 6, and 7 give concrete examples of cases where the null is true, but κ_r is the same as, less than, or greater than κ_w.

Table 5. Illustration of Across-Occasion Table: Cohen's Kappa for Dichotomous Item – κ = 0.50

Trend-year reliability/ Occasion A score	Occasion B score		Marginal rate
	Score category 1	Score category 0	
Score category 1	0.72	0.08	0.80
Score category 0	0.08	0.12	0.20
Marginal rate	0.80	0.20	0.68

Table 6. Illustration of Across-Occasion Table: Cohen's Kappa for Dichotomous Item – κ = 0.41

Trend reliability/ Occasion A score	Occasion B score		Marginal rate
	Score category 1	Score category 0	
Score category 1	0.81	0.09	0.9
Score category 0	0.04	0.06	0.1
Marginal rate	0.85	0.15	0.87

Table 7. Illustration of Across-Occasion Table: Cohen's Kappa for Dichotomous Item – κ = 0.525

Trend reliability/ Occasion A score	Occasion B score		Marginal rate
	Score category 1	Score category 0	
Score category 1	0.63	0.07	0.7
Score category 0	0.12	0.18	0.3
Marginal rate	0.75	0.25	0.81

Quadratically Weighted Kappa

Cohen (1968) introduced weighted kappa to account for the possibility that some disagreements are more important than others. This is certainly the case for the ordered category scores assigned to CRs, and weighted kappa is used much more frequently than unweighted kappa as a measure of rating accuracy. To account for this difference, testing programs that use weighted kappa usually use quadratically weighted kappa, where the weight is a function of the square of the difference between the two scores: $w = \left(x_1 - x_2\right)^2$.

In the case of dichotomous items, quadratically weighted kappa is equal to regular kappa. Therefore, in order to investigate the properties of weighted kappa, we must consider items with at least three categories as in Table 8.

Table 8. Within-Occasion Rescore Table for Three-Category Item

Original score	Rescore 0	Rescore 1	Rescore 2	Margin
0	$\gamma_{0\|0}f_0$	$\gamma_{1\|0}f_0$	$\gamma_{2\|0}f_0$	f_0
1	$\gamma_{0\|1}f_1$	$\gamma_{1\|1}f_1$	$\gamma_{2\|1}f_1$	f_1
2	$\gamma_{0\|2}f_2$	$\gamma_{1\|2}f_2$	$\gamma_{2\|2}f_2$	f_2
Margin	f_0	f_1	f_2	1

Note that $\sum_{k=0}^{m}\gamma_{k|j}=1$.

By symmetry of the rescore table:

$$\begin{aligned}\gamma_{0|1}f_1 &= \gamma_{1|0}f_0\\ \gamma_{0|2}f_2 &= \gamma_{2|0}f_0\\ \gamma_{2|1}f_1 &= \gamma_{1|2}f_2.\end{aligned} \tag{30}$$

Making these substitutions, it is straightforward to show the column totals are equal to the row totals.

Weighted Kappa for Within-Occasion Rescore Table

Weighted kappa is given as follows:

$$\kappa_w = 1-\frac{q_a}{q_c} \tag{31}$$

$$q_a = \sum_{i=0}^{m}\sum_{j=0}^{m} f_{ij}w_{ij} \tag{32}$$

$$q_a = \sum_{i=0}^{m}\sum_{j=0}^{m} f_{ij}\left(i-j\right)^2 \tag{33}$$

$$q_c = \sum_{i=0}^{m}\sum_{j=0}^{m} f_i f_j\left(i-j\right)^2 . \tag{34}$$

Applying this to the within-occasion rescore table reveals

$$\begin{aligned} q_{aw} &= \gamma_{1|0} f_0 + 4\gamma_{2|0} f_0 \\ &+\gamma_{0|1} f_1 + \gamma_{2|1} f_1 \\ &+4\gamma_{0|2} f_2 + \gamma_{1|2} f_2 \end{aligned} \tag{35}$$

and

$$\begin{aligned} q_{cw} &= f_0 f_1 + 4 f_0 f_2 \\ &+f_1 f_0 + f_1 f_2 \\ &+4 f_2 f_0 + f_2 f_1. \end{aligned} \tag{36}$$

Table 9 shows the rescore table for the three-category item.

Table 9. Across-Occasion Rescore Table Under Null Hypothesis

Original score	Rescore			Margin
	0	1	2	
0	$\gamma_{0\|0} h_0$	$\gamma_{1\|0} h_0$	$\gamma_{2\|0} h_0$	h_0
1	$\gamma_{0\|1} h_1$	$\gamma_{1\|1} h_1$	$\gamma_{2\|1} h_1$	h_1
2	$\gamma_{0\|2} h_2$	$\gamma_{1\|2} h_2$	$\gamma_{2\|2} h_2$	h_2
Margin	$\gamma_{0\|0} h_0 + \gamma_{0\|1} h_1 + \gamma_{0\|2} h_2$	$\gamma_{1\|0} h_0 + \gamma_{1\|1} h_1 + \gamma_{1\|2} h_2$	$\gamma_{2\|0} h_0 + \gamma_{2\|1} h_1 + \gamma_{2\|2} h_2$	1

For simplicity, we select the rescore sample by shifting the only the proportions only in Score Categories 0 and 1:

$$\begin{aligned} h_0 &= f_0 + \delta \\ h_1 &= f_1 - \delta \ . \\ h_2 &= f_2 \end{aligned} \tag{37}$$

Using the symmetry assumptions from the original within-occasion table yields Table 10.

Table 10. Across-Occasion Rescore Table With Symmetry Assumptions

Original score	Rescore			Margin
	0	1	2	
0	$\gamma_{0\|0} f_0 + \gamma_{0\|0}\delta$	$\gamma_{0\|1} f_1 + \gamma_{1\|0}\delta$	$\gamma_{0\|2} f_2 + \gamma_{2\|0}\delta$	$f_0 + \delta$
1	$\gamma_{1\|0} f_0 - \gamma_{0\|1}\delta$	$\gamma_{1\|1} f_1 - \gamma_{1\|1}\delta$	$\gamma_{1\|2} f_2 - \gamma_{2\|1}\delta$	$f_1 - \delta$
2	$\gamma_{2\|0} f_0$	$\gamma_{2\|1} f_1$	$\gamma_{2\|2} f_2$	f_2
Margin	$f_0 + m_0$	$f_1 + m_1$	$f_2 + m_2$	1

Where
$$\begin{aligned} m_0 &= \delta\left(\gamma_{0|0} - \gamma_{0|1}\right) \\ m_1 &= \delta\left(\gamma_{1|0} - \gamma_{1|1}\right) \\ m_2 &= \delta\left(\gamma_{2|0} - \gamma_{2|1}\right). \end{aligned} \tag{38}$$

We note in passing that, for quality scoring, we expect the following:

$$\begin{aligned} &m_0\left(exact - adjacent \text{ for category } 0\right) > 0, \\ &m_1\left(adjacent - exact \text{ for category } 1\right) < 0, \\ &m_2\left(discrepant - adjacent \text{ for category } 2\right) < 0. \end{aligned}$$

To compute the weighted kappa for the rescore table, the quadratically weighted proportion of disagreement q_{ar} is

$$\begin{aligned} q_{ar} = {} & \gamma_{0|1} f_0 + \gamma_{1|0}\delta + 4\gamma_{0|2} f_2 + 4\gamma_{2|0}\delta \\ & + \gamma_{1|0} f_0 - \gamma_{0|1}\delta + \gamma_{1|2} f_2 - \gamma_{2|1}\delta \\ & + 4\gamma_{2|0} f_0 + \gamma_{2|1} f_1 \end{aligned} \tag{39}$$

$$q_{ar} = q_{aw} + \gamma_{1|0}\delta + 4\gamma_{2|0}\delta - \gamma_{0|1}\delta - \gamma_{2|1}\delta \tag{40}$$

$$q_{ar} = q_{aw} + b \tag{41}$$

$$b = \delta\left(\gamma_{1|0} - \gamma_{0|1} + \gamma_{2|0} - \gamma_{2|1}\right) + 3\gamma_{2|0}\delta \tag{42}$$

$$\begin{aligned} q_{cr} = {} & f_0 f_1 + f_1 m_0 + 4 f_0 f_2 + 4 f_2 m_0 \\ & + f_1 f_0 + f_0 m_1 + f_1 f_2 + f_2 m_1 \\ & + 4 f_2 f_0 + 4 f_0 m_2 + f_2 f_1 + f_1 m_2 \end{aligned} \tag{43}$$

$$\begin{aligned} q_{cr} = {} & q_{cw} + f_1 m_0 + 4 f_2 m_0 \\ & + f_0 m_1 + f_2 m_1 \\ & + 4 f_0 m_2 + f_1 m_2 \end{aligned} \tag{44}$$

$$q_{cr} = q_{cw} + e \tag{45}$$

$$e = \left(m_1 + m_2\right) f_0 + \left(m_0 + m_2\right) f_1 + \left(m_0 + m_1\right) f_2 + 3 f_2 m_0 + 3 f_0 m_2 \tag{46}$$

$$\begin{aligned} m_1 + m_2 &= \delta\left(\gamma_{1|0} - \gamma_{1|1}\right) + \delta\left(\gamma_{2|0} - \gamma_{2|1}\right) \\ m_0 + m_2 &= \delta\left(\gamma_{0|0} - \gamma_{0|1}\right) + \delta\left(\gamma_{2|0} - \gamma_{2|1}\right) \\ m_0 + m_1 &= \delta\left(\gamma_{0|0} - \gamma_{0|1}\right) + \delta\left(\gamma_{1|0} - \gamma_{1|1}\right) \end{aligned} \tag{47}$$

$$\kappa_r = 1 - \frac{q_{ar}}{q_{cr}} \tag{48}$$

$$\kappa_r = 1 - \frac{q_{aw} + b}{q_{cw} + e}. \tag{49}$$

Both *b* and *e* only contain terms multiplied by δ , and so when $\delta = 0$, $\kappa_w = \kappa_r$.

$$\kappa_w - \kappa_r = 1 - \frac{q_{aw}}{q_{cw}} - \left(1 - \frac{q_{aw} + b}{q_{cw} + e}\right) \tag{50}$$

$$\kappa_w - \kappa_r = \frac{q_{aw} + b}{q_{cw} + e} - \frac{q_{aw}}{q_{cw}} \tag{51}$$

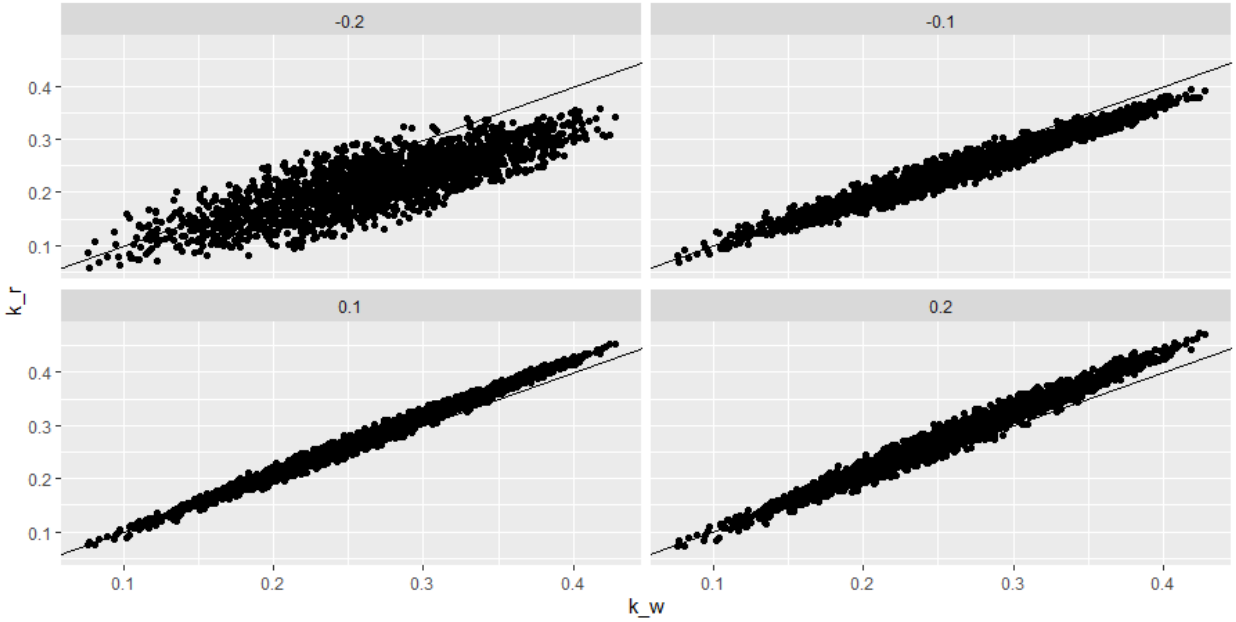
$$\kappa_w - \kappa_r = \frac{q_{aw}q_{cw} + bq_{cw} - q_{aw}q_{cw} - eq_{aw}}{q_{cw}\left(q_{cw} + e\right)} \tag{52}$$

$$\kappa_w - \kappa_r = \frac{bq_{cw} - eq_{aw}}{q_{cw}\left(q_{cw} + e\right)} \tag{53}$$

$$\kappa_w - \kappa_r = \frac{b - \left(1 - \kappa_w\right)e}{q_{cw} + e}. \tag{54}$$

As was the case for unweighted κ , Equation 54 is hard to develop intuition about. Thus, a large number of data sets were sampled. The Occasion A margin was (0.25, 0.5, 0.25). The rescore design shifted values between Categories 0 and 1 by the quantity δ . For each category, the probability of agreement $\gamma_{k|k}$ was sampled from uniform [0.3, 0.6]. The probability $\gamma_{1|0}$, $\gamma_{0|1}$, and $\gamma_{1|2}$ were each sampled from a uniform [0.1, 0.3]. The final conditional probability was computed as one minus the sum of the other two probabilities. Finally, the probabilities were sorted to ensure that the probability of exact agreement ≥ probability of adjacent agreement ≥ probability of discrepant. Then quadratically weighted κ_w and κ_r were computed. The results are presented in Figure 3 for δ = -0.2, -0.1, 0.1, and 0.2 (δ = 0 is not shown because $\kappa_w \equiv \kappa_r$ in that case). Negative values of δ (the margin proportion of rescore papers shifted from category 1 to category 0) generally resulted in lower values of quadratically weighted κ_r than κ_w, while increasing its value to allocate more papers to *h₀* generally resulted in κ_r values that were higher than κ_w.

Figure 3. Across-Occasion κ_r as a Function of Within-Occasion κ_w for Varying Rescore Designs



Note. Panels are defined by the proportion by which the margin proportion is shifted from Category 1 to Category 0.

The conclusion from Figure 3 is that the rescore design can affect the value of quadratically weighted kappa even when the null hypothesis is true.

The Odds Ratio

The odds ratio, or its log, is widely used as a measure of association, especially in fields such as epidemiology. However, with the exception of its use in the Mantel-Haenszel procedure for detecting differential item functioning, the measure has not received much attention in educational research. For a dichotomous variable, such as in Table 1, the odds X = 1 (where P(X_A = 1) = f) are given by

$$odds(X_A = 1) = \frac{f}{1-f} \quad . \tag{55}$$

The odds ratio (α) is the conditional odds (X_B = 1| X_A = 1) divided by the conditional odds (X_B = 1|X_A = 0):

$$\alpha = \frac{odds\left(X_B = 1 \mid X_A = 1\right)}{odds\left(X_B = 1 \mid X_A = 0\right)} = \frac{f_{11} / f_{01}}{f_{10} / f_{00}} = \frac{f_{11} f_{00}}{f_{10} f_{01}} . \tag{56}$$

Note that defining the ratio in terms of conditional odds (X_A = 1|X_B = 1) yields the same quantity:

$$\alpha = \frac{f_{11} / f_{10}}{f_{01} / f_{00}} = \frac{f_{11} f_{00}}{f_{10} f_{01}} . \tag{57}$$

For a 2 x 2 table, once the margins of the table are fixed (i.e., there is one degree of freedom), it is easy to show that the odds ratio uniquely determines the individual cell counts.

The scale of the odds ratio α is $\left(0,\infty\right)$, with an odds ratio of 1 signifying independence. If the rows or columns of the table are switched, the resulting odds ratio is $\alpha' = \frac{1}{\alpha}$. As a result, the scale of the odds ratio is unwieldy in that it is markedly asymmetric. Therefore, the log of the odds ratio is often used. In the log metric, 0 corresponds to independence, negative values indicate negative association, and positive values indicate positive association. Switching the rows or columns of the table yields the same value but with opposite sign.

In the present context, it is useful to consider whether the properties extend to the rescore case: Is the measure influenced by the marginal frequencies of the rescore table, ***h*** ? Referring back to Table 4, the within-occasion odds ratio is

$$\alpha_w = \frac{\gamma_{1|1} f \left[1 - 2f + \gamma_{1|1} f\right]}{\left(1-\gamma_{1|1}\right) f \left(1-\gamma_{1|1}\right) f} \tag{58}$$

$$\alpha_w = \frac{\gamma_{1|1}\left[1-2f+\gamma_{1|1}f\right]}{\left(1-\gamma_{1|1}\right)^2 f}. \tag{59}$$

Using the quantities from Table 5, the across-occasions rescore odds ratio is

$$\alpha_r = \frac{\gamma_{1|1}h\left[1-\left(\frac{f}{1-f}\right)\left(1-\gamma_{1|1}\right)\right](1-h)}{\left(1-\gamma_{1|1}\right)h\left(1-\gamma_{1|1}\right)\left(\frac{f}{1-f}\right)(1-h)} \tag{60}$$

$$\alpha_r = \frac{\gamma_{1|1}\left(1-f-f+\gamma_{1|1}f\right)}{\left(1-\gamma_{1|1}\right)\left(1-\gamma_{1|1}\right)f} \tag{61}$$

$$\alpha_r = \frac{\gamma_{1|1}\left(1-2f+\gamma_{1|1}f\right)}{\left(1-\gamma_{1|1}\right)^2 f} \tag{62}$$

$$\alpha_r = \alpha_w. \tag{63}$$

Thus, we see that the odds-ratio does *not* depend on the marginal proportions ***h*** selected for the trend rescore.

In the context of measuring agreement, use of the odds ratio is somewhat controversial. Shoukri (2004, p. 26) asserted that the odds ratio is a measure of association and therefore not appropriate to measure agreement. Cicchetti and Feinstein (1990) described the odds ratio as a measure of agreement, although they pointed out that it is not corrected for chance level agreement. They also pointed out that observing a single zero cell entry in a 2 x 2 table will give a false picture (falsely good for $f_{10} = 0$ or $f_{01} = 0$; falsely bad for $f_{11} = 0$ or $f_{00} = 0$) of the true level of agreement. However, Agresti (1984, pp. 15–18) noted that adding 0.5 to the frequency count in each cell in the table stabilizes both the odds ratio and its estimated standard error. Fienberg and Holland (1970) provided a more general discussion of handling zero cells in contingency tables. In summary, it appears that the odds ratio may be a useful measure of rater agreement, but the issue requires further study.

Discussion

Donoghue et al. (2022) found that the value of percent exact agreement and *t*-test statistics could be affected by the rescore design. Even when the null hypothesis was true ($\gamma_{n|m} = \eta_{n|m}$ for all *n* and *m*), the statistics could be larger or smaller than in the within Occasion A table in Table 1. This substantially limits the utility of the statistics to accurately detect changes in rating from Occasion A to Occasion B. The goal of this research memorandum was to extend the findings of Donoghue et al. to other measures, Cohen's kappa and weighted kappa, and the odds ratio. It was demonstrated that kappa and quadratically weighted kappa are sensitive to the rescore design. Under the null hypothesis of no change in the scoring process, the value of both statistics can be made higher or lower than the value observed for the original, within-occasion scoring. Seeing a change in the statistic's value from Occasion A to Occasion B, it is unclear whether the change is due to the rescore design or due to changes in the raters' behavior. The findings in Figures 1 and 2, for example, show that even though the rating process is unchanged, the rescore design can cause a large decrease in the value of the κ statistic from Occasion A to Occasion B. This could result in needless costly retraining and/or rescoring. Results for weighted kappa indicate that it, too, is subject to these effects and can lead to erroneous conclusions about the quality of the cross-occasion rating. Given this sensitivity to the rescore design, these measures cannot be recommended for use in evaluating rescore data.

It was also found that a well-known property of the odds ratio, its insensitivity to the marginal distributions, prevents it from being adversely affected by the rescore design. However, a key liability of the odds ratio, like the chi-squared statistic, is that it is a measure of association, not agreement. As such, it measures any form of association rather than agreement, which is the specific form of association of interest (e.g., Cohen, 1968) in rescore studies.

In this paper, the odds ratio has been developed for 2 x 2 tables. In the context of studying rater agreement, this corresponds to dichotomously scored items. However, most CR items are scored in more than two categories. Therefore, an important question is how odds ratio based methods may be extended to items scored in more than two categories.

An approach that has been widely used in the literature is based on computing specialized versions of log-linear models that treat the scores as ordered categories (e.g., Tanner & Young, 1985). Another common approach involves computing multiple odds ratios from the data in the *M* x *M* cross tabulation of ratings. Agresti (1984, pp. 18–23) gives three versions of subsetting the data. Each of these schemes asks a somewhat different question about the data. The ordered nature of the variables restricts the set of comparisons considered.

Future Directions

A promising direction in the analysis of rescore data is to develop measures that explicitly condition on Occasion A (row) score. For example, Donoghue and Eckerly (2019) computed separate statistics within each Occasion A response category. This has the effect of effectively creating *M* test statistics, one per response category. They also examined combining these measures into an omnibus test and obtained very promising results—the statistics had good Type I error control and were not subject to the wild variations seen for the paired *t*-test. Such an approach explicitly accommodates the rescore design and therefore supports valid statistical inferences about the rescore data.

References

Abdalla, W. (2019). *Detecting rater effects in trends scoring* [Unpublished doctoral dissertation]. University of Iowa.

Agresti, A. (1984). *Analysis of ordinal categorical data*. John Wiley & Sons.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

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. (2019, April 4–8). *A new test of rater drift in trend scoring* [Paper presentation]. National Council on Measurement Education Annual Meeting. Toronto, Ontario, Canada.

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

Fienberg, S. E., & Holland, P. W. (1970). Methods for eliminating zero counts in contingency tables. In G. P. Patil (Ed.), *Random counts in scientific work* (pp. 233–260). Pennsylvania State University Press.

Shoukri, M. M. (2003). *Measures of interobserver agreement and reliability.* CRC Press. https://doi.org/10.1201/9780203502594

Tanner, M. A., & Young M. A. (1985). Modeling ordinal scale disagreement. *Psychological Bulletin*, *98*(2), 408–415. https://doi.org/10.1037/0033-2909.98.2.408