Research Report

Investigating Constructed-Response Scoring Over Time: The Effects of Study Design on Trend Rescore Statistics

ETS RR–22-15

John R. Donoghue
Catherine A. McClellan
Melinda R. Hess

December 2022

ETS Research Report 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 Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Investigating Constructed-Response Scoring Over Time: The Effects of Study Design on Trend Rescore Statistics

John R. Donoghue[1], Catherine A. McClellan[2], & Melinda R. Hess[3]

1 ETS, Princeton, NJ

2 Australian Council for Educational Research, Camberwell, VIC, Australia

3 Research, Evaluation, and Assessment of Living and Learning, Inc., Lutz, FL

When constructed-response items are administered for a second time, it is necessary to evaluate whether the current Time B administration's raters have drifted from the scoring of the original administration at Time A. To study this, Time A papers are sampled and rescored by Time B scorers. Commonly the scores are compared using the proportion of exact agreement across times and/or *t*-statistics comparing Time A means to Time B means. It is common to treat these rescores with procedures that assume a multinomial sampling model, which is incorrect. The correct, product-multinomial model reflects the stratification of Time A scores. Using direct computation, the research report demonstrates that both proportion of exact agreement and the *t*-statistic can deviate substantially from expected behavior, providing misleading results. Reweighting the rescore table gives each statistic the correct expected value but does not guarantee that the usual sampling distributions hold. It is also noted that the results apply to a wider class of situations in which a set of papers is scored by one group of raters or scoring engine and then a sample is selected to be evaluated by a different group of raters or scoring engine.

Keywords Constructed-response scoring; trend scoring; rater drift; rater monitoring statistics

doi:10.1002/ets2.12360

Introduction

Many large-scale assessment programs use constructed-response (CR) items. Such items require that the examinee create a response, rather than selecting one from a provided set of choices. CR items are believed to assess somewhat different aspects of a construct and to tap behaviors that cannot be accessed easily by other item formats. However, these benefits are not realized without costs. Although automated scoring has made large advances in the past years, most CR items still require at least partial scoring by human raters, adding expense, and time to large-scale assessments. In addition, lack of consistency between raters can add variability to the data and so may reduce the reliability of the assessment results.

Scoring reliability is generally assessed using measures like percent of exact agreement between raters. In assessment programs that maintain trends in data across time, there are two aspects to measuring the reliability of scoring CR items: within assessment and across assessment.

The issues in the current research report use the terminology of trend scoring to ground the discussion. The issues identified apply in any situation where a set of responses is selected from one scoring mode (e.g., humans or a given scoring engine) and applied to another (e.g., a different set of humans or a different scoring engine).

Interrater Reliability Estimation

Within-year reliability is checked by scoring a sample of papers a second time using a different rater. In practice, many testing programs have adopted effect size criteria to assess whether an item shows less than acceptable interrater agreement. Usually, the criterion varies based on the number of score categories, with items having more score categories having lower criteria. For illustrative purposes, in this report, we examine the criteria adopted by the National Assessment of Educational Progress (NAEP) in 2002.[1] For dichotomous items, the minimum percent agreement for within-year reliability is required to be 85%. Items that have reliability values below 85% are flagged and discussed, and if appropriate, the

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

raters are retrained on the scoring rubric. In extreme cases, when acceptable reliability cannot be obtained, all scores for an item are discarded, and a different set of raters is trained to score the item. Note that other measures of reliability, such as Cohen's (1968) quadratically weighted kappa and the proportional reduction in mean square error (Haberman, 2020), are used by other testing programs. We chose to focus on percent agreement and the *t*-statistic because the procedures are simple and therefore widely used, and they form an important part of NAEP's scoring monitoring.

Trend Reliability

To measure across-time (trend) reliability, a set of papers from the earlier assessment is rescored during the current administration's scoring session. This allows a direct comparison of scores from the current assessment (Time B) with those from the earlier assessment (Time A). The NAEP criteria for trend reliability are as follows:

- Between Time A and Time B, there can be no more than a 7% decrease from the within-year percent exact agreement from Time A.
- The *t*-statistics for mean differences comparing the item mean score at Time A to that at Time B may not exceed ± 1.5.

Items that fail to meet these criteria are subject to review. If the lack of agreement is severe, the scores may need to be discarded and the raters retrained. These consequences are expensive, both financially and in time lost.

The selection of trend papers from the previous assessments may affect the accuracy of the measure of trend reliability. However, some aspects of the rescore process are not specified, and the literature provides little to no guidance:

- How should the Time A responses chosen for the rescore set be distributed? For dichotomous responses, should the same number of correct and incorrect responses be rescored, or should some other, disproportionate distribution be used?
- For dichotomous items, does reweighting the proportion of correct and incorrect responses in the sample improve detection of rater drift?

In this report, analytic relationships are derived between the two indexes of trend scoring quality (percent agreement and *t*-statistics of mean differences) and the properties of the trend rescore set of responses. To keep the algebra tractable, this report focuses on the case of dichotomous CR items, but the issues identified apply equally to polytomous items.

Theoretical Approach

Within-Year Reliability

For an arbitrary dichotomous CR item, Table 1 portrays the true (population) rescore probabilities at Time A.[2] The probability of a correct response is π. The percent exact agreement is the sum of the diagonal terms p_{11} and $p_{00} = 1 - 2\pi + 2\alpha$.

When the proportion of exact agreement p_A is fixed, all the values in the table are fixed, with $\alpha = \pi - \left[\left(1 - p_A\right)/2\right]$.

Trend Reliability

Table 2 shows the probabilities in the trend rescore table. The parameter δ reflects the change in mean item ratings (proportion scored "correct") from Time A scoring to Time B scoring. The parameter η reflects the difference in proportion exact agreement between within-year comparisons and trend rescore comparisons. Note that the changes keep the Time A margins the same as in Table 1. The proportion exact agreement for Table 2 is therefore $1 - 2\pi + 2\alpha - \eta$.

Rescore Papers

To study the properties of the rescore process, we consider taking a sample of size N of Time A papers and rescoring them at Time B. N_1 of these were given a score of 1 at Time A, and $N - N_1 = N_0$ were given a score of 0. The results are shown in Table 3.

Table 1 Population proportions: within-year rescore table

	Time A rescore		
	1	0	Total
Time A score			
1	α	$\pi - \alpha$	π
0	$\pi - \alpha$	$1 - 2\pi + \alpha$	$1 - \pi$
Total	π	$1 - \pi$	1

Table 2 Population proportions: across-administration rescore table

	Time B rescore (Y)		
	1	0	Total
Time A score (X)			
1	$\alpha - \frac{\delta}{2} - \frac{\eta}{2}$	$\pi - \alpha + \frac{\delta}{2} + \frac{\eta}{2}$	π
0	$\pi - \alpha - \frac{\delta}{2} + \frac{\eta}{2}$	$1 - 2\pi + \alpha + \frac{\delta}{2} - \frac{\eta}{2}$	$1 - \pi$
Total	$\pi - \delta$	$1 - \pi + \delta$	1

Note. Proportions other than 1/2 for δ change the proportion agreement. Similarly, proportions other than 1/2 for η for the proportion agreement also affect the marginal means.

Table 3 Sample sizes: across-administration rescore table

	Time B score (Y)		
	1	0	Total
Time A score (X)			
1	n_{11}	n_{10}	N_1
0	n_{01}	n_{oo}	N_0
Total	$n_{11} + n_{01}$	$n_{10} + n_{00}$	1

Sampling Models for Rescores

We assume that the population probabilities underlying these papers are given in Table 2. Because Table 2 is a 2×2 table, one approach would be to assume that $P\left(n_{11}, n_{10}, n_{01}, n_{00}\right) \sim$ multinomial $P\left(N, \pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}\right)$. In this case, the probability of observing the rescore table is

$$\begin{aligned} P\left(n_{11}, n_{10}, n_{01}, n_{00}\right) = & \frac{N!}{n_{11}!n_{10}!n_{01}!n_{00}!} \\ & \times \left(a - \frac{\delta}{2} - \frac{\eta}{2}\right)^{n_{11}} \left(\pi - a + \frac{\delta}{2} + \frac{\eta}{2}\right)^{n_{10}} \\ & \times \left(\pi - a - \frac{\delta}{2} + \frac{\eta}{2}\right)^{n_{01}} \left(1 - 2\pi + a + \frac{\delta}{2} - \frac{\eta}{2}\right)^{n_{00}}. \end{aligned} \quad (1)$$

Multinomial sampling implies that there is a population of papers that have been scored and rescored and that we then take a sample of size N from that population of papers. In some cases (e.g., random sampling of responses from the Time A pool of responses), multinomial sampling may be a reasonable model.

Generally, however, this does not correspond to reality. Instead, N Time A responses, N_1 with scores of $X = 1$, and $N_0 = (N - N_1)$ with scores of $X = 0$ are selected for the trend scoring set. Often, N_1 will be chosen so that either $N_1 = N_0$ or $N_1/N = \widehat{\pi}$. In these cases, it is appropriate to consider the row (Time A) margins of the table fixed. Given that a response was scored a 1 at Time A ($X = 1$), the conditional probability that it was scored a 1 at Time B ($Y = 1$) is

$$P(Y = 1|X = 1) = \frac{\pi - \frac{\delta}{2} - \frac{\eta}{2}}{\pi} = \gamma_{1|1\cdot\cdot} \quad (2)$$

Similarly, given that the Time A score was a 0 ($X = 0$), the probability that the Time B score is 1 is

$$P(Y = 1|X = 0) = \frac{\pi - \alpha - \frac{\delta}{2} + \frac{\eta}{2}}{1 - \pi} = \gamma_{1|0}, \tag{3}$$

with $\gamma_{0|1}$ and $\gamma_{0|0}$ defined analogously. Therefore, we have two samples that are distributed binomially: $n_{11} \sim B\left(N_1, \gamma_{1|1}\right)$ and $n_{10} \sim B\left(N_0, \gamma_{1|0}\right)$. The probability of sampling a specific Table 3 is the product of these two independent binomial probabilities:

$$\begin{aligned} P\left(n_{11}, n_{10}, n_{01}, n_{00}\right) &= B\left(N, n_{11}, \gamma_{1|1}\right) \times B\left(N_0, n_{10}, \gamma_{1|0}\right) \\ &= \left(\begin{array}{c} N_1 \\ n_{11} \end{array} \right) \gamma_{1|1}^{n_{11}} \left(1 - \gamma_{1|1}\right)^{N_1 - n_{11}} \left(\begin{array}{c} N_0 \\ n_{01} \end{array} \right) \gamma_{1|0}^{n_{01}} \left(1 - \gamma_{1|0}\right)^{N_0 - n_{01}}. \end{aligned} \tag{4}$$

This is referred to as a product-binomial model. It is a special case of the product-multinomial model (Feinberg, 1980, p. 30), where each Time A score level response follows a multinomial distribution, and it is the appropriate model when the sampling has been stratified. Rescored trend responses follow a product-binomial model.

The difference between multinomial sampling and product multinomial may seem subtle, but the consequences of a test based on the wrong model can be substantial. As discussed shortly, the product-binomial model was used for all results presented in this report. The results, however, are easy to extend to the case of product-multinomial sampling.

Statistical Tests of the Rating Process

Several indexes of rater agreement/quality of scoring are available (see, e.g., Abedi, 1996). In this report, we focus on two very widely used indexes: the proportion exact agreement and the comparison of the mean item score in the original and rescored data. For both, the null hypothesis is that the rating process is the same at Time B as it was at Time A. Under this hypothesis, the probability structure of Table 1 would be obtained (i.e., both δ and η are 0). Under the alternative hypothesis of rater drift, however, the structure of Table 2 is obtained, with either $\delta \neq 0$, $\eta \neq 0$, or both. For example, NAEP's 2002 criteria specified that a change in exact agreement must be less than .07 and that the absolute value of the t-statistic must be less than or equal to 1.5.[3]

Exact Agreement

The sample estimate of the exact agreement is

$$p_A = \frac{n_{11} + n_{00}}{N}, \tag{5}$$

and the hypothesis is

$$\eta \geq .07 \rightarrow p_A < p_{\text{crit}}. \tag{6}$$

For dichotomous items, NAEP typically uses a minimum acceptable p_A of .85, so $p_{\text{crit}} = .78$. Therefore, each table can be classified as exhibiting acceptable exact agreement ($p_A \geq p_{\text{crit}}$) or not ($p_A < p_{\text{crit}}$). Adopting a sampling model in Equation (1) or Equation (4), we can compute the probability of each possible rescore table. These probabilities are summed over tables to determine the overall probability that the exact agreement is acceptable. We examine two cases, either under the null hypothesis that $\eta = 0$ or under the alternative that $\eta = .07$.

The *t*-Statistic to Compare Item Means

Detecting a change in the average item score is another important component of evaluating trend scoring. Even if there is a high rate of agreement, if all the disagreements are in the same direction, the item can show a shift in overall mean. A paired t-statistic is commonly used to determine whether the means differ:

$$\begin{aligned} t &= \frac{n_{10} - n_{01}}{\sqrt{n_{10} + n_{01} - Nd^2}} \\ d &= \frac{n_{10} - n_{01}}{N}. \end{aligned} \tag{7}$$

Appendix A presents a brief proof that there are two cases in which the expected value of the t-statistic is zero: (a) when the proportion of 1 s selected for rescore $P = \pi$ and (b) when it happens that $\alpha = \pi^2$. As in the case of the percent exact agreement, we find that there is an interaction between π and P. NAEP uses a critical t-value of $|t| \geq 1.5$; rescore tables with $|t| \geq 1.5$ are deemed to exhibit scoring drift. As was the case with the exact agreement p_A, by adopting a sampling model in Equation (1) or Equation (4), we can compute the probability of each rescore table. We will examine two cases: under the null hypothesis that $\delta = 0$ or under the alternative that $\delta \neq 0$. Finally, note that although the *t-statistic* is being used, no reference is made to the *t-distribution*. Instead, we are computing the exact probability under the product-binomial model; the situation is analogous to using a permutation test or Fisher's exact test.

Results

Exact Agreement

For the exact agreement, Equation (1) was used to compute the probability that the proportion exact agreement between Time A and Time B ratings in the rescore table differed from that at Time A by more than .07. The original (within Time A) proportion exact agreement was assumed to be .85. Two factors were varied: (a) Prob(correct response) for Time A, π, and (b) P, the proportion of the rescored responses that were correct (i.e., $P = N_1/N$). We assume that item mean score is unchanged ($\delta = 0$). For power calculations, the rate in the rescore table was 7% lower ($\eta = 0$). Figure 1 presents these results.

Figures 1 and 2 show that there is a strong interaction between the value of π and both the Type I behavior and power of the test to detect a change in the percent exact agreement. When $\pi = .5$, there is virtually no effect of P. The Type I error rate ($N = 300$) is essentially .00 for all values of P. Similarly, the power is effectively .467 for $N = 300$ and .477 for $N = 600$, for all values of P. When $\pi = .25$, however, there is a strong interaction between π and P. For $P > .40$, the Type I error rate increases rapidly and is over .90 for $P = .80$. The correct rejection rate also increases with P, but at a much faster rate. Note that there is relatively little effect of sample size between $N = 300$ and $N = 600$; it is dwarfed by the effect of P. The preceding analyses assume that there has been no change in the average item score (i.e., $\delta = 0$). Our work also indicates that the direction of δ interacts with the values of π and P.

The *t*-Statistic to Compare Item Means

Figure 3 illustrates the relationship between π, the item mean at Time A, and P, the proportion of rescored responses that were correct at Time A, using a critical t-statistic of ± 1.5. We use the critical value of the t-statistic as an effect size and adopt the product-binomial sampling model developed earlier.

For each value of π, the Type I error rate attains its minimum at $\pi = P$. It increases sharply as P moves away from π in either direction. The minimum is close to 1, and $p(|t|) > 1.5$. Figure 4 shows the probability of correctly rejecting the null hypothesis when $\delta = -.05$. The appearance is similar to Figure 3, but the curves are shifted to the left.

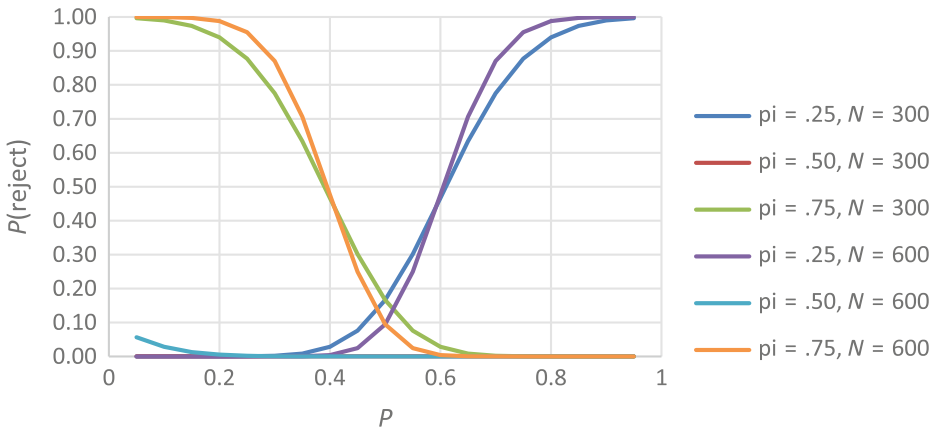

Figure 1 Proportion exact agreement: proportion identified as "significant" decrease, null case.

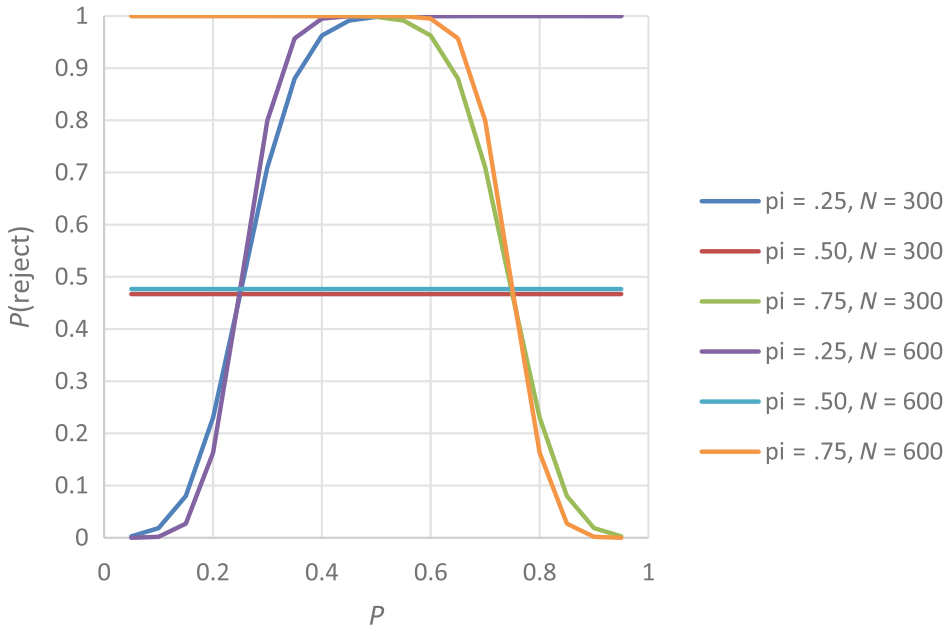

Figure 2 Proportion exact agreement: proportion identified as "significant" decrease, nonnull case.

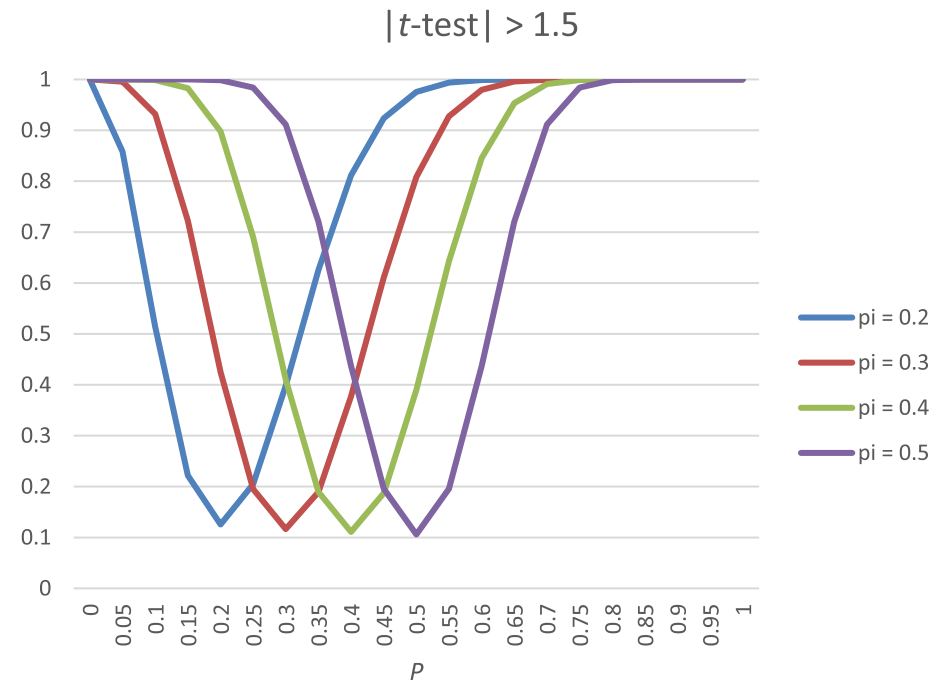

Figure 3 Type I error rate, *t*-statistic, for varying values of π and *P*.

The effect of ignoring the interaction between π and *P* is illustrated in Tables 4–6. In each case, 600 responses are selected to be rescored, where $\pi = .4$. Table 4 displays results when $\pi = P$. When the null hypothesis is true (see Figure 3), there is an approximately 90% chance of correctly failing to reject (recall that the critical *t*-value is ±1.5). When the null hypothesis is false ($\delta = -.05$; Figure 3), the false null hypothesis is correctly rejected slightly more than 80% of the time.

Tables 5 and 6 demonstrate the impact of selecting $\pi \neq P$. In Table 5, *P* was 5% higher than π (45% vs. 40%). The probability of correctly rejecting a false null hypothesis is 80%. But the Type I error rate ballooned to 29%. This level of Type I inflation is clearly unacceptable. Table 6 shows the reverse situation: *P* was 5% lower than π (35% vs. 40%). In this

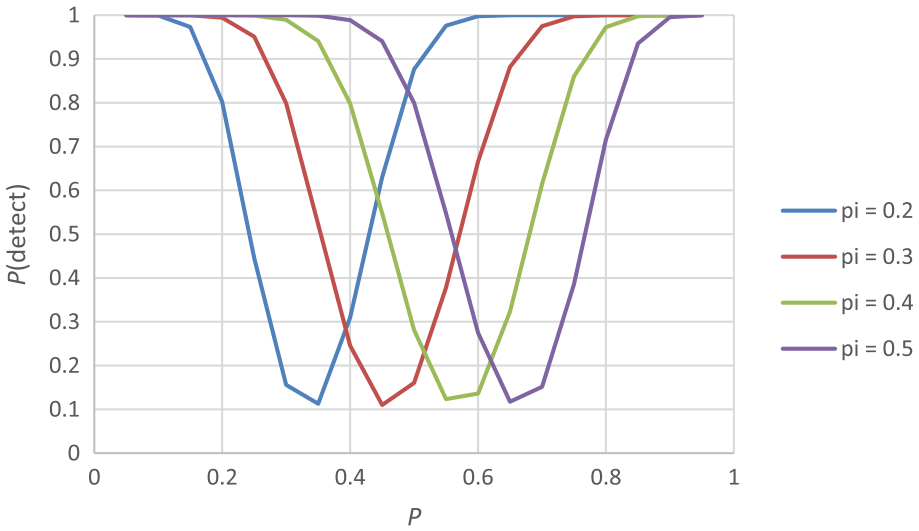

Figure 4 Correct detection rates, t-statistic, for varying values of π and P.

Table 4 $N = 600, \pi = .4, P = .4, \delta = -.05$

	True ho (%)	False ho (%)
Fail to reject	89.7	3.2
Reject	10.3	96.8

Table 5 $N = 600, \pi = .4, P = .45, \delta = -.05$

	True ho (%)	False ho (%)
Fail to reject	71.1	20.2
Reject	28.9	79.8

situation, the Type I error rate is similar to that in Table 5 (30% vs. 29%). Moreover, the probability of correctly rejecting a false hypothesis increased to 99.8%.

Reweighting the Rescore Table

Given the results of the previous section, it is natural to wonder to what extent the poor performance of the proportion of exact agreement and the t-statistic can be ameliorated through clever data analysis. As a first step toward examining this question, this section examines the possible benefits of collecting data with one allocation of Time A scores (i.e., N_1 correct responses) but then reweighting the results to yield an unbiased t-statistic. Because the rescore sample is artificially selected/constructed from the Time A responses, it is not appropriate to assume that $P(X = 1) = \pi$ the proportion correct in the population at Time A; the rescore sample may differ for the Time A population in overall performance on the item.

What one can assume as invariant are the values of $P(Y|X)$, the conditional probability of Y given X. These are given by the parameters $\gamma_{1|1}$ defined in Equation (2) and $\gamma_{0|1}$ defined in Equation (3). It is these conditional probabilities that really define the properties of the rescore. The goal, then, is to reweight the table so that the marginal $P(X = 1) = \pi$ while maintaining the conditional probability of the second score Y given the value of the first score X, that is, the parameters $\gamma_{1|1}$ and $\gamma_{0|1}$. This is accomplished by reweighting each of the rows separately. For correct responses, the cell counts are multiplied by π/P, and cell counts for incorrect responses are multiplied by $(1 - \pi)/(1 - P)$. These adjustments maintain the conditional probability $P(Y|X)$ while adjusting the marginal probability $P(X = 1) = \pi$. Table 7 summarizes the reweighting. Appendix B demonstrates that reweighting in this manner gives the t-statistic and expected value of 0. Appendix B also shows that the expected value of the proportion of exact agreement is equal to the Table 1 value, $1 - 2\pi + 2\alpha$. However, the

Table 6 $N = 600$, $\pi = .4$, $P = .35$, $\delta = -.05$

	True ho (%)	False ho (%)
Fail to reject	70.4	0.2
Reject	29.6	99.8

Table 7 Sample sizes: reweighted rescore table

	Time B score (Y)		
	1	0	Total
Time A score (X)			
1	$r_{11} = \frac{\pi}{P} n_{11}$	$r_{10} = \frac{\pi}{P} n_{11}$	$\frac{\pi}{P} N_1 = \pi N$
0	$r_{01} = \frac{1-\pi}{1-P} n_{01}$	$r_{00} = \frac{1-\pi}{1-P} n_{11}$	$\frac{1-\pi}{1-P} N_0 = (1-\pi)N$
Total	πN	$(1-\pi)N$	N

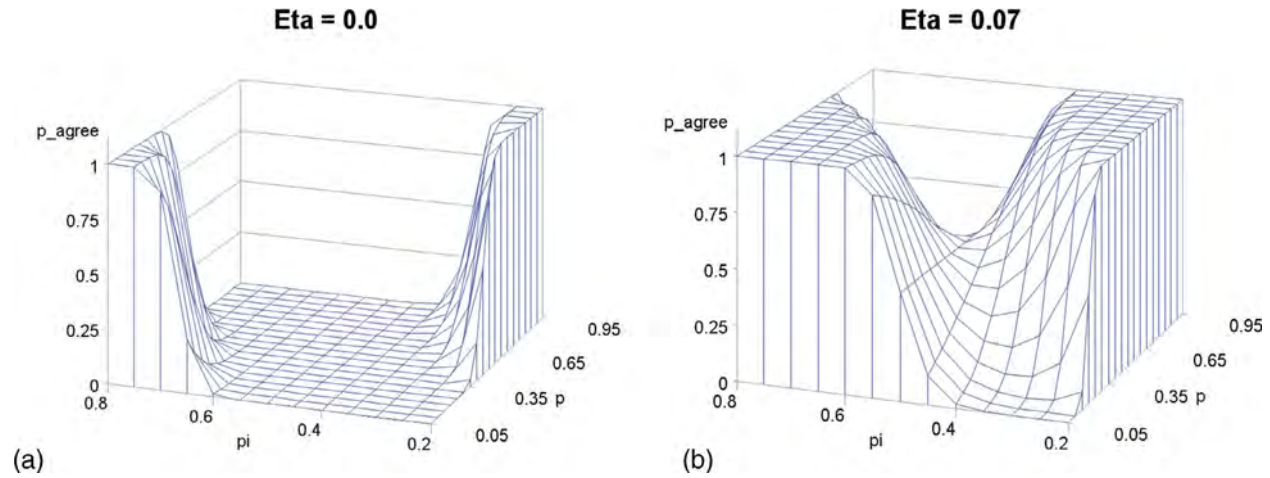

Figure 5 Detection for percent exact agreement in original tables.

results in Appendix B show only that the tests are unbiased. T here is no guarantee that the statistics will have the behavior usually expected of them.

Results

The reweighting approach was applied to tables for a wide range of values of π and P. The rescore sample size was held constant at 300. Figure 5 shows the effect of reweighting upon the test of exact agreement. When there is no change in the proportion of exact agreement (Figure 5a), the unweighted tables give good results when π and P are similar; this is indicated by the broad valley in the middle of the figure, in which the Type I error is essentially 0. When π and P differ, however, the Type I error rate can increase dramatically, as shown by the hills at the corners of the figure. When the reliability has changed ($\eta = .07$; Figure 5b), the surface has similar shape, but the valley is much narrower.

Figure 6 shows results for the proportion of exact agreement when the rescore table is reweighted. If the null hypothesis is true (Figure 6a), the surface is flat and uniformly close to 0 for all but the most extreme mismatches of values of π and P. When the null hypothesis is false ($\eta = .07$; Figure 6b), again, the surface is flat. In this case, however, the elevation is approximately .5, indicating moderate power that is insensitive to the value of π or P.

Comparison of Figures 7 and 8 shows the effect of reweighting on the t-statistic. Figure 7 gives unweighted results, and Figure 8 gives reweighted results. For the original, unweighted table, Figure 7a shows the probability of rejection when the null hypothesis is true, that is, Type I error. There is a narrow trough around the line $\pi = P$, in which the Type I error is well controlled. Outside of this trough, however, there is gross Type I error inflation, with the rate reaching 1.0 for most

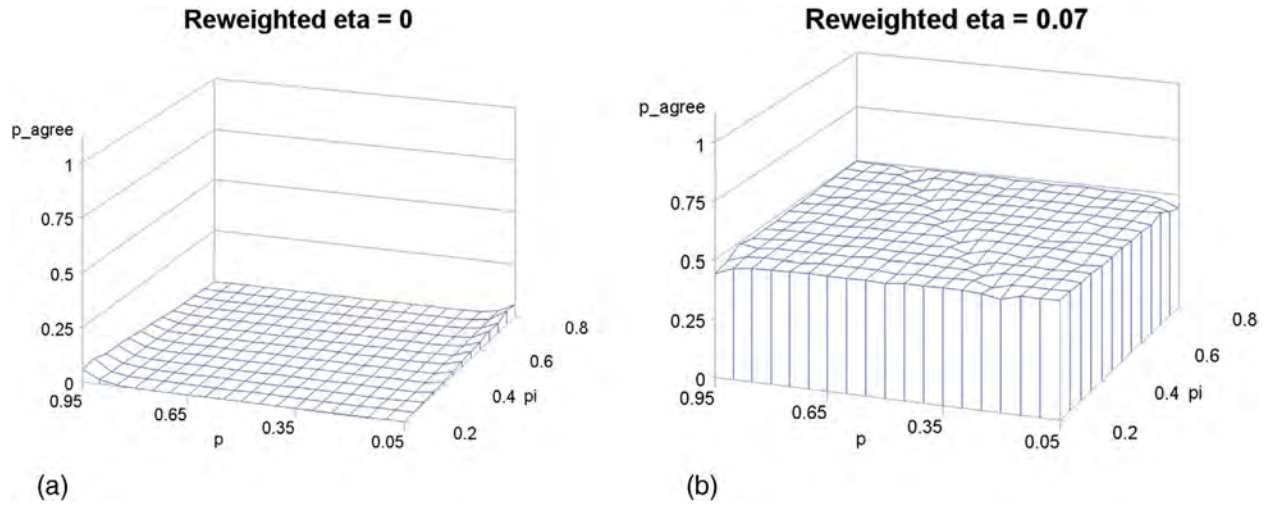

Figure 6 Detection for percent exact agreement in reweighted tables.

combinations of values. In the nonnull case in Figures 7b,c, the overall shape of the surface is similar to that in the null case, but the trough has been shifted in each case. Although these figures depict a high degree of correct detection of true differences, it cannot be termed "power," except in the regions where the Type I error was controlled, that is, around the line $\pi = P_1$.

Figure 8 shows results for the t-statistic when the tables are reweighted. Figure 8a plots the Type I error rates for different values of π and P. Although there is some Type I error rate inflation when π and P differ substantially, the surface is much lower and has a broad, flat range in which the Type I error rate is well-controlled (recall that a critical t-statistic of 1.5 was used).

Figure 8b,c shows detection for the t-statistic when $\delta = \pm 0.05$, for reweighted tables. The surface is uniformly high, with an average value between .6 and 1.0. In contrast to the original, unweighted tables, the Type I error control was relatively good for many combinations of π and P in the reweighted tables (although when the two differ substantially, the Type I error was still severely inflated), allowing for valid comparison of powers of different designs.

Discussion

Simplifying Probability Computations

It is worth noting that it is possible to simplify the computations described here, making analysis of rescore designs simple and faster. Under the product-binomial model, it is easy to prove that for the exact agreement P_A, the rejection region is coordinate-wise convex, if (n_{11}, n_{00}) falls in the rejection region and $(n_{11}, n_{00} - 1)$ also falls in the rejection region. Indeed, any smaller value of n_{00} will fall in the rejection region. Thus, for any given value of n_{11}, we can determine the maximum value of n_{00} that falls in the rejection range and then compute the cumulative binomial probability that k is less than or equal to n_{00}, rather than computing the value for every value of n_{00}.

For the t-statistic, the rejection region is the area above a parabola. The parabola is oriented about a line of positive slope (i.e., tipped over). In this case, for a given value of $_{11}$, there is a minimum and maximum value of n_{00} between which the t-statistic will not be significant. Again, the cumulative binomial probability can be computed for each, greatly reducing the amount of computation.

Making use of these kinds of results can greatly reduce the amount of brute force computation, while still allowing the exact probability to be computed. In addition, making use of continuous approximations to the discrete binomial distribution (i.e., the widely used normal approximation) can further reduce computation, although at the cost of some approximation error. Given the simplicity of the requisite computations, it would seem prudent to do them prior to executing any rescore study.

The effect of reweighting the table was unexpectedly effective in ameliorating the problems identified in analysis of the unweighted tables. This raises the possibility of dual-purpose designs, such as $N_1 = N_0$, to get the most stable estimates

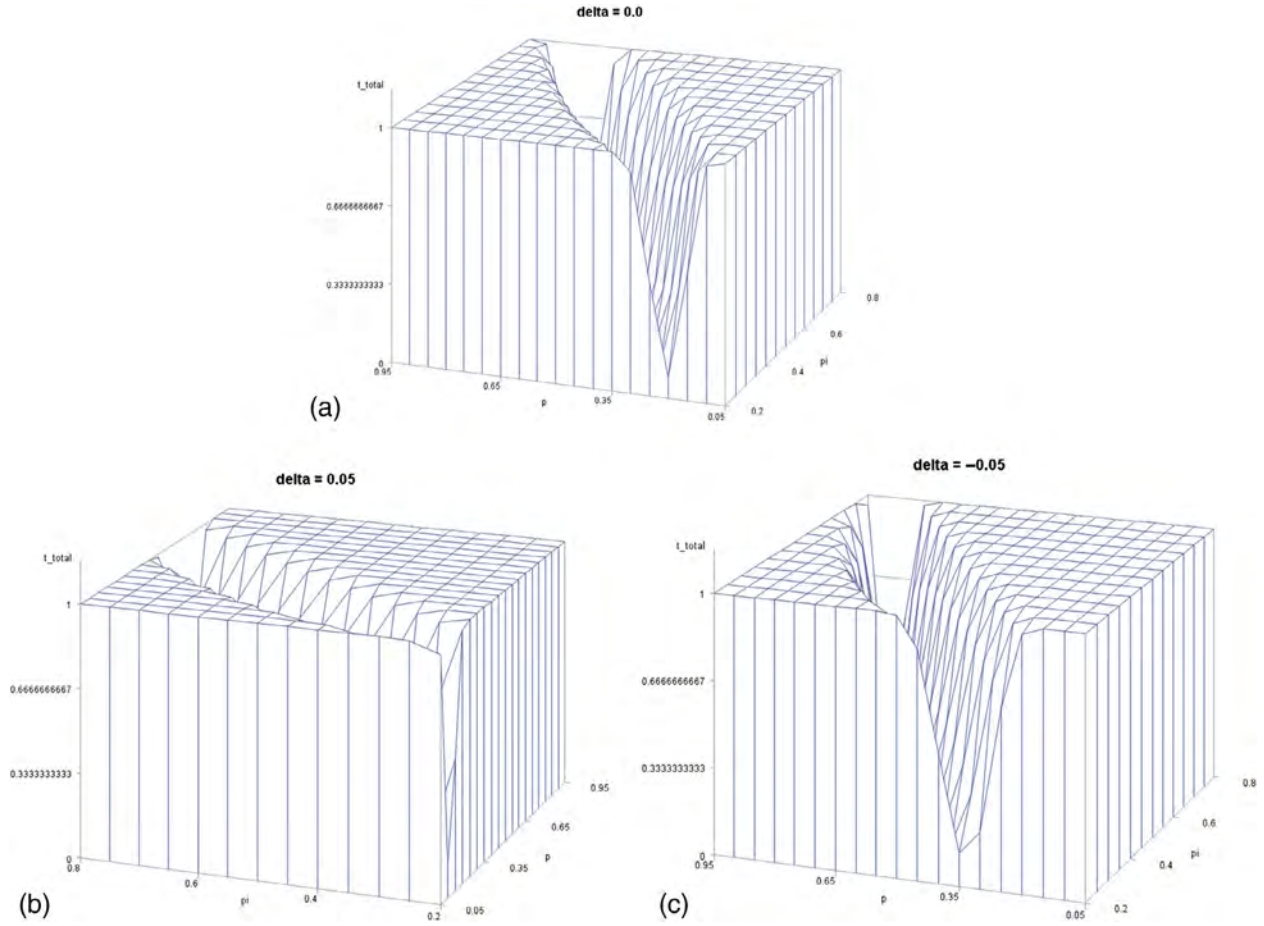

Figure 7 Detection for t-statistic in original tables.

of $\gamma_{1|1}$ and $\gamma_{1|0}$, and then reweighting the resulting table to get better tests for the proportion exact agreement and the t-statistic.

Extensions

This study was a first look to determine whether the design of a rescore study can have much effect on the statistical properties of the rescore statistics. Clearly, the answer is that it can. As this is an initial, exploratory study, a proper treatment of the extensions and future work can take longer to describe than the actual study did. Following are some of our ideas.

Other Measures of Association/Drift

This report examined the two most widely used measures of scoring quality: the exact agreement and the t-statistic. Other measures are certainly of interest, especially the (log) odds ratio and kappa statistic, including quadratically weighted kappa (Cohen, 1968) and the proportion of reduction in mean square error (Haberman, 2020), which are intended to be less sensitive to the margins of the table or to address the shortcomings of older, simpler measures.

Analytic/Computational Approaches

As was noted, making use of continuous approximations to the discrete distributions can simplify computing. It may also allow us to develop models that can facilitate analytic solutions, for example, to the problem of determining the optimal rescore design for a specific item.

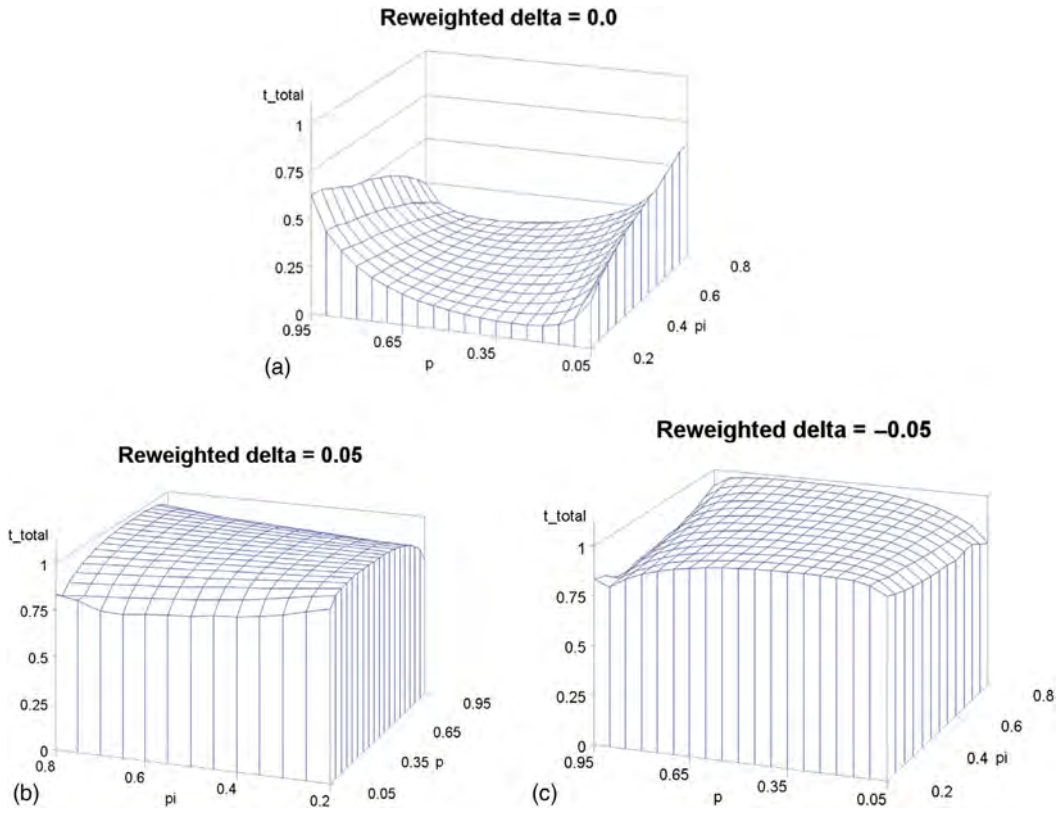

Figure 8 Detection for *t*-statistic for reweighted tables.

Polytomous Items

A limitation of the present work is that it focuses strictly on dichotomous items. Yet many CR items are scored polytomously. Extending the procedures here to polytomous items is straightforward (although the algebra is more complicated). For example, the product-binomial model becomes the product-multinomial sampling model.

Optimal Study Design

Clearly, a next step for this work is to identify a set of procedures with which one can determine the optimal design, in terms of Type I error control and power, for specific CR items. This could be done through a straightforward application of the computational approaches examined here. Alternatively, one could assign explicit costs to different types of errors and use the tools of Bayesian decision theory to explicitly minimize the expected loss of the rescore design.

More Sophisticated Models

The model (or, more properly, lack of model) for raters used here is consistent with the proportion of exact agreement and the *t*-statistic. However, more sophisticated approaches are certainly available and would prove interesting avenues for further work. Generalizability theory; models of contingency tables, such as log-linear models; and rater-based item response theory models, such as the FACETS model (e.g., Linacre, 1994) and hierarchical rater model (DeCarlo et al., 2011; Donoghue & Hombo, 2002; Patz et al., 2002), may provide interesting perspectives on the problem of design of a rescore study.

Conclusion

Detecting rescore problems is important; failing to do so can result in unnecessarily losing trend information or, in extreme circumstances, losing all the data from an item. On the other hand, Type I errors are *very* expensive in this context.

Declaring the Time B scoring process out of line can result in the item being retrained and rescored, often at a cost of thousands of dollars. The tools developed in this report allow the rescore studies to be designed in more principled ways and the implications of design decisions to be examined and made on a more rational basis.

Finally, it should be reemphasized that the results in this report apply to *any* situation in which a selected set of responses is rescored. The discussion has been framed in the terminology of trend, but the principles apply equally to the situation when a set of papers scored by human raters is then rescored by an automated scoring engine or when scores produced by one team of scorers or by a automated scoring engine is compared to that of another team or engine. Appropriate use of the product-binomial and product-multinomial coupled with reweighting model allows for improved inferences.

Acknowledgments

Catherine A. McClellan's work on this report was conducted while she was employed by ETS. Melinda R. Hess's work on this report was conducted while she was a predoctoral intern at the University of South Florida in Tampa. The work reported herein was supported under the National Assessment of Educational Progress (grant R999G50001, CFDA number 84.999G) as administered by the Office of Educational Research and Improvement, U.S. Department of Education. Portions of the report were prepared for presentation at the 2003 AERA/NCME program, Chicago, IL.

Notes

1 Although still based largely on proportion exact agreement and *t*-statistic plus Cohen's kappa, NAEP's current scoring monitoring procedures are more complicated and multifaceted, with different levels and color coding. Mirroring these rules would greatly complicate this report's presentation without affecting the issues described. To keep the discussion focused, we use the simpler historical rules in this report.
2 The rater who assigns the score and the rater who assigns the rescore are typically selected randomly from the pool of raters. Because every rater plays both roles, the rescore tables are usually symmetric.
3 As noted earlier, NAEP's current criteria are much more complex but largely based on the same basic approach.

References

Abedi, J. (1996). The interrater/test reliability system (ITRS). *Multivariate Behavioral Research*, *31*(4), 409–417. https://doi.org/10.1207/s15327906mbr3104_1

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

DeCarlo, L. T., Kim, Y., & Johnson, M. A. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333–356. https://doi.org/10.1111/j.1745-3984.2011.00143.x

Donoghue, J. R., & Hombo, C. M. (2002, April 2–4). *Detecting rater drift using an IRT hierarchical rater's model* [Paper presentation]. National Council on Measurement in Education annual meeting, New Orleans, LA, USA.

Feinberg, S. E. (1980). *The analysis of cross-classified categorical data* (2nd ed.). MIT Press.

Haberman, S. J. (2020). *Application of best linear prediction and penalized best linear prediction to ETS Tests* (Research Report No. RR-20-08). ETS. https://doi.org/10.1002/ets2.12290

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Patz, R. J., Junker, B. W., Johnson, M. A., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. https://doi.org/10.3102/10769986027004341

Appendix A

Cases in Which the Expected Value of the *t*-Statistic Is 0

$$E(Y=1) = E\left(\frac{N_1}{N}P(Y=1|X=1)\right) + E\left(\frac{N_0}{N}P(Y=1|X=0)\right),$$

$$E(Y=1) = \frac{N_1}{N}E(Y=1|X=1) + \frac{N_0}{N}E(Y=1|X=0),$$

$$E(Y=1)=\frac{N_1}{N}\gamma_{1|1}+\frac{N-N_1}{N}\gamma_{1|0},$$

$$E(Y=1)=\frac{N_1}{N}\frac{\alpha}{\pi}+\frac{N_0}{N}\frac{\pi-\alpha}{1-\pi}.$$

Case 1

The proportion of papers selected from Time A matches the marginal probability of $X=1$ at Time A:

$$P_1=\frac{N_1}{N}=\pi, P_0=\frac{N_0}{N}=(1-\pi),$$

$$E(Y=1)=\pi\frac{\alpha}{\pi}+(1-\pi)\frac{\pi-\alpha}{1-\pi},$$

$$E(Y=1)=\alpha+\pi-\alpha,$$

$$E(Y=1)=\pi.$$

Case 2

If the proportions do not match, $P_1=\frac{N_1}{N}=\pi+\varepsilon$, $P_0=\frac{N_0}{N}=(1-\pi-\varepsilon)$ for $\varepsilon\neq 0$, then the expected value is 0 only if $\alpha=\pi^2$:

$$E(Y=1)=(\pi+\varepsilon)\frac{\alpha}{\pi}+(1-\pi-\varepsilon)\frac{\pi-\alpha}{1-\pi},$$

$$E(Y=1)=\alpha+\frac{\alpha\varepsilon}{\pi}+\pi-\alpha-\frac{\varepsilon\pi-\varepsilon\alpha}{1-\pi},$$

$$E(Y=1)=\pi+\frac{\alpha\varepsilon-\pi\alpha\varepsilon+\alpha\varepsilon\pi-\varepsilon\pi^2}{\pi(1-\pi)},$$

$$E(Y=1)=\pi+\frac{\alpha\varepsilon-\varepsilon\pi^2}{\pi(1-\pi)},$$

$$E(Y=1)=\pi+\frac{\varepsilon\left(\alpha-\pi^2\right)}{\pi(1-\pi)},$$

$$E(Y=1)=\pi+\frac{\varepsilon\left(\pi^2-\pi^2\right)}{\pi(1-\pi)},$$

$$E(Y=1)=\pi.$$

Appendix B

Properties of Reweighted Tables

Proof that reweighted table has $E(t)=0$

We show that

$$E\left(r_{11}+r_{10}\right)=N\pi,$$

$$E\left(r_{11}+r_{10}\right)=E\left(\frac{\pi}{P}n_{11}+\frac{1-\pi}{1-P}n_{10}\right),$$

$$E\left(r_{11}+r_{10}\right)=\frac{\pi}{P}E\left(n_{11}\right)+\frac{1-\pi}{1-P}E\left(n_{10}\right),$$

$$E\left(r_{11}+r_{10}\right)=\frac{\pi}{P}N_1\frac{\alpha}{\pi}+\frac{1-\pi}{1-P}\left(N-N_1\right)\frac{\pi-\alpha}{1-\pi},$$

$$E\left(r_{11}+r_{10}\right)=\frac{N\alpha}{N_1}N_1+\frac{(\pi-\alpha)N}{N-N_1}\left(N-N_1\right),$$

$$E\left(r_{11}+r_{10}\right)=N\alpha+(\pi-\alpha)N,$$

$$E\left(r_{11}+r_{10}\right)=\pi.$$

Proof that reweighted table has the same percent exact agreement as the year 1 table, $1-2\pi+2\alpha$

$$p_{\text{agree}}=E\left(\frac{r_{11}+r_{00}}{N}\right),$$

$$p_{\text{agree}}=\frac{1}{N}\left[E\left(\frac{\pi}{P}n_{11}\right)+E\left(\frac{1-\pi}{1-P}n_{00}\right)\right],$$

$$p_{\text{agree}}=\frac{1}{N}\left[\frac{\pi}{P}E\left(n_{11}\right)+\frac{1-\pi}{1-P}E\left(n_{00}\right)\right],$$

$$p_{\text{agree}}=\frac{1}{N}\left[\frac{\pi}{P}N_1\gamma_{1|1}+\frac{1-\pi}{1-P}N_0\gamma_{0|0}\right],$$

$$p_{\text{agree}}=\frac{N_1}{N}\frac{\pi}{P}\frac{\alpha}{\pi}+\frac{N_0}{N}\frac{1-\pi}{1-P}\left(\frac{1-2\pi+\alpha}{1-\pi}\right),$$

$$p_{\text{agree}}=P\frac{\alpha}{P}+(1-P)\frac{1-2\pi+\alpha}{1-P},$$

$$p_{\text{agree}}=1-2\pi+2\alpha.$$

Suggested citation:

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

Action Editor: Brent Bridgeman

Reviewers: Adrienne Sgammato and Jodi Casabianca-Marshall