On the Performance of the Marginal Homogeneity Test to Detect Rater Drift

Applied Psychological Measurement

2018, Vol. 42(4) 307–320

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617730390

journals.sagepub.com/home/apm

Adrienne Sgammato¹ and John R. Donoghue¹

Abstract

When constructed response items are administered repeatedly, “trend scoring” can be used to test for rater drift. In trend scoring, raters rescore responses from the previous administration. Two simulation studies evaluated the utility of Stuart’s Q measure of marginal homogeneity as a way of evaluating rater drift when monitoring trend scoring. In the first study, data were generated based on trend scoring tables obtained from an operational assessment. The second study tightly controlled table margins to disentangle certain features present in the empirical data. In addition to Q , the paired t test was included as a comparison, because of its widespread use in monitoring trend scoring. Sample size, number of score categories, interrater agreement, and symmetry/asymmetry of the margins were manipulated. For identical margins, both statistics had good Type I error control. For a unidirectional shift in margins, both statistics had good power. As expected, when shifts in the margins were balanced across categories, the t test had little power. Q demonstrated good power for all conditions and identified almost all items identified by the t test. Q shows substantial promise for monitoring of trend scoring.

Keywords

rater agreement, monitoring scoring, rater effects, constructed response scoring, trend scoring

Introduction

When constructed response (CR) items are administered repeatedly, “trend scoring” can be used to test for rater drift. In trend scoring, raters rescore responses from the previous administration. This simulation study evaluates the utility of Stuart’s (1955) Q measure of marginal homogeneity as a way of evaluating rater drift when monitoring trend scoring. To ensure the data reflect realistic kinds of drift, the data generation is based on trend scoring tables obtained from an operational assessment. In addition to the lesser known Q , the paired t test was included because of its familiarity and widespread use in monitoring trend scoring.

The article is organized as follows. The next section contains background on monitoring of CR scoring, and the additional challenges in trend scoring. Next, computation of the paired t test and Q is described. Two simulations are then presented. In the first, focus is placed on the utility of Q and t test to detect realistic kinds of drift, and so the data generation was based on tables obtained from trend scoring in an operational assessment. In the second simulation, the table

¹Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

John R. Donoghue, Educational Testing Service, 660 Rosedale Road, Mail Stop 02-T, Princeton, NJ 08541, USA.

Email: jdonoghue@ets.org

margins are tightly controlled to unconfound some features present in the empirically based generating interrater agreement tables used for the first simulation. This is followed by discussion and final conclusions.

Background

CR items are widely used in educational assessment. Although there has been substantial progress made in the field of automated scoring, many testing programs such as National Assessment of Educational Progress (NAEP) and the State of Texas Assessments of Academic Readiness (STAAR) contain items that must be scored by human raters. In a typical case, raters assign a (ordered) categorical designation to a student response to an open-ended or essay item based on a scoring rubric. The purpose of scoring rubrics is to define each score level of a particular item. Raters are trained on how to interpret scoring rubrics and consistently apply each of the score categories. To achieve high-quality scoring, clear rubrics and exemplars are beneficial. However, it is necessary to monitor scoring to ensure that the raters remain “on rubric.” Myford and Wolfe (2009) describe several ways (e.g., changes in severity) in which raters’ application of the rubric may change over time.

A number of useful models exist to investigate the quality of scoring/rater effects. The original multifaceted Rasch model (Linacre, 1989), as well as its extensions (e.g., PARSCALE’s rater effect model (Muraki & Bock, 2003) for generalized partial credit model/graded response model [GPCM/GRM] items or the Markov Chain Monte Carlo [MCMC] version; Wang & Yao, 2013), allows scaling of multiple choice and of CR items, while modeling rater severity/leniency effects. These models are appropriate for multiple ratings of the same item and can be used to model a variety of rater effects (Myford & Wolfe, 2009). A criticism of these models is that multiple ratings appear as independent contributions to the likelihood (i.e., additional ratings are treated as additional items). Mariano (2002) showed that because the model does not reflect dependence between ratings of the same response, this has the erroneous implication that increasing the number of ratings of a single response allows examinee ability to be known exactly. To deal with this issue, the hierarchical rater model (Patz, Junker, Johnson, & Mariano, 2002) and its signal detection version (DeCarlo, Kim, & Johnson, 2011) impose a hierarchical structure on the severity (leniency) in multiple ratings of the same response. Related models include Wilson and Hoskens’s (2001) rater bundle model and Verhelst and Verstralen’s (2001) IRT model for multiple raters. These models also incorporate adjustments to account for dependence across multiple ratings of a single item.

The models described are most useful when the entire assessment has been processed; that is, all items have been scored and merged with multiple-choice components of the same assessment. However, in the authors’ experience, “scoring monitoring” typically occurs earlier, before the full data are available. In many testing programs (e.g., NAEP), CR scoring is a separate, prior activity. Each item is treated as a discrete entity, scored by one or more teams of raters. In part, this isolation is intentional, to avoid cross-item effects, such as halo effects where the raters form a holistic view of the set or responses and assign similar (or the same) scores to all of the individual items. Different items are scored at different times by different teams. Therefore, scoring monitoring must be based on only the data for the single item, making the rater effects models discussed above unavailable. The goal of monitoring is to intervene (via retraining raters, or possibly rescore the entire item) when the scoring shows evidence of deviating from expectations.

When monitoring CR scoring, two types of monitoring are commonly used. The first type is to use “validity” papers, which have expert-defined “true” scores. Comparing the score the rater assigns to the “true” score assigned by the expert gives direct evidence about scoring

accuracy. However, validity scoring requires expert raters to assign the scores. In many cases, such experts may be difficult to secure, expensive, or both. The second type of monitoring is rescoring. In rescoring, operational papers are assigned to two or more raters, who independently assign scores. The result can be placed in a two-way cross-tabulation of scores. If raters cannot reliably understand and apply the scoring rubrics, then any subsequent analyses based on inconsistent scores will yield results that are subject to this rater error. Lack of rater reliability negatively impacts the validity of the study or assessment and the reliability of the final analysis results. It is important to bear in mind that it is possible for the raters to agree and both be wrong. Nonetheless, rescoring is the most common way of monitoring scoring quality.

Many measures exist to evaluate rater agreement. Popular statistics used in practical testing applications include percentage exact agreement, correlations, Cohen's (1960) kappa, and the intraclass correlation. Although these statistics and procedures are widely used, they are not without limitation. The percentage exact agreement, for example, has been described as an inflated measure of true agreement (Landis & Koch, 1975, 1977; Zwick, 1988) as there are likely some instances in which raters agree due to chance or random guessing. Much criticism exists surrounding the use of the Cohen's kappa statistics. It is well known that the unweighted kappa statistic is sensitive to the marginal distribution and to the prevalence of the construct in the population (Brennan & Prediger, 1981; Byrt, Bishop, & Carlin, 1993; Feinstein & Cicchetti, 1990; Uebersax, 1987). Therefore, it is possible for kappa to be acceptable in the presence of systematic differences among raters (i.e., bias). Other commonly used rater agreement statistics are reviewed by Stemler and Tsai (2008) who note the advantages and disadvantages associated with each. For example, the Pearson correlation coefficient can mask systematic disagreement among raters (i.e., mean shifts). The intraclass correlation (ICC), like kappa, is sensitive to score variability and of the homogeneity of the sample, and it is not comparable across populations.

Trend Scoring

The statistics just described are often used on cross-sectional data when evaluating agreement among raters at one particular time. When CR items are administered in multiple test administrations, an additional concern is that the current administration's scoring should not differ systematically from the scoring in the previous administration. For example, in large-scale educational assessments such as the NAEP or Program for International Student Assessment (PISA), a shift in rater severity is confounded with the change of student proficiency that is being measured. To measure shifts in scoring, another set of rescores can be collected, referred to as trend scoring. A set of papers that were scored at time *A* are presented to time *B* raters. In this case, there is heightened concern about score drift. Are the papers being scored differently at time *B* than at time *A*? Such a shift can arise from a variety of causes. For example, time *B* raters might not be drawn from the same pool of candidates as time *A* raters. Also, in programs that conduct live training of raters, it may not be reasonable to assume that scorers have had exactly the same training. Trend scoring is designed to detect such effects.

To assess severity in trend scoring an item at times *A* and *B*, the paired *t* test is commonly used:

$$t = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\text{Var}(\bar{X}_A) + \text{Var}(\bar{X}_B) - 2\text{Cov}(\bar{X}_A, \bar{X}_B)}}, \quad (1)$$

where $\bar{X}_A$ is the mean score of raters at time *A*, and $\bar{X}_B$ is the mean score of raters at time *B*.

If time *B* raters differ in the way they interpret and apply the scoring rubrics from the interpretation at time *A*, the mean scores for time *A* and *B* raters may differ. The null hypothesis is

Table 1. Contingency Table of Counts.

		Rater 2				
		1	2	3	...	Total (marginal)
Rater 1	1	n_{11}	n_{12}	n_{13}	n_{1j}	n_{1+}
	2	n_{21}	n_{22}			n_{2+}
	3	n_{31}		n_{33}		n_{3+}
	...	n_{ij}			n_{ji}	n_{i+}
Total (marginal)		n_{+1}	n_{+2}	n_{+3}	n_{+i}	$n_{++} = N$

that the difference is zero; that is, the average score should not differ between the two time’s scorers. Mean score differences also translate to differences in the marginal distributions.

However, other changes in the distribution of scores will *not* be detected by *t*. A more appropriate null hypothesis is that the distribution of scores across all of the categories is the same in the two times. For the two-way cross-time or trend rescore table, the appropriate null hypothesis is that the table’s margins are the same in the two administrations. Marginal homogeneity refers to the identity of population marginal proportions in two-way rescore tables. Consider a two-way contingency table. Each cell in the Table 1 contains the counts n_{ij} . The assumption of marginal homogeneity states that the distributions of scores show only stochastic, not systematic, variability. Testing for violations of marginal homogeneity is useful when there is interest in determining whether changes occur in the number of scores for each category, depending on whether the paper was scored at time *A* or at time *B*, that is, the “rater drift” question.

Several measures of marginal homogeneity exist (e.g., Bhapkar, 1966, 1979; McNemar, 1947; Stuart, 1955). The present investigation focuses on Stuart’s *Q* which tests the homogeneity of marginal probabilities in a two-way classification table. Stuart’s *Q* is applicable to two-way tables for variables with multiple levels or categories. It simplifies to McNemar’s (1947) *Q* in the case of 2×2 tables. The hypothesis associated with Stuart’s test is $H_0: \mathbf{v}_A - \mathbf{v}_B = \mathbf{0}$; that the population vector of marginal counts for time *A* scores is equal to the population vector for time *B* scores. The formula for Stuart’s *Q* is given in Equation 2.

$$Q = (\mathbf{n}_A - \mathbf{n}_B)^T \mathbf{V}^{-1} (\mathbf{n}_A - \mathbf{n}_B), \tag{2}$$

where $\mathbf{n}_A$ is the vector of row (time *A*) marginal counts, $\mathbf{n}_B$ is the vector of column (time *B*) marginal counts, and $\mathbf{V}$ is the $(k - 1) \times (k - 1)$ variance–covariance matrix of the counts, with elements $V_{ii} = n_{i+} + n_{+i} - 2n_{ii}$ and $V_{ij} = -(n_{ij} + n_{ji})$.

Under the null hypothesis of no difference in marginal counts, the *Q* statistic is asymptotically distributed as χ^2 with $k - 1$ degrees of freedom, for an item with *k* score categories.

Zwick (1988) noted that marginal homogeneity is a necessary but not sufficient condition for homogeneous rater behavior. Zwick recommends starting with tests of marginal homogeneity (specifically, Stuart’s *Q*) and proceeding to a measure of agreement only if marginal homogeneity holds. To date, tests of marginal homogeneity are not commonly employed in social science research and practice.

The Present Study

To the extent that rater drift is monitored in real time in practical testing situations, a paired *t* test is often used. However, *t* does not capture all aspects of rater disagreement that are important

practically. The t tests the hypothesis of group mean differences; however, practitioners might be more interested in the marginal distributions of each rater group—the hypothesis that Q tests. Therefore, the purpose of this investigation was to evaluate Stuart's (1955) Q statistic as a measure for monitoring cross-year or trend scoring.¹ Because the examination of changes in overall score distributions over time was of interest, and because it is widely used in this context, the paired t statistic was also included. These two statistics were compared to evaluate their relative performance in terms of Type I error and power. Given the differences in the hypotheses that each statistic tests, some differences in performance were expected by design. However, one important question was to see whether the two statistics provide complementary information, or if one statistic is sufficient for nonhomogeneous rescore table margins.

Type I error rate and power for the two statistics were of interest, and identifying situations in which either was more powerful than the other. It was hypothesized that the t test would have little or no power when changes to the margins did not change the means. However, the authors did not know if Q would have much power in that circumstance. Similarly, it was unknown whether Q would be competitive with the paired t test when the means did change. Hence, a simulation study was conducted to address these questions.

Method

In this simulation study, four factors were manipulated. The factors were as follows:

1. Number of responses to be rescored (i.e., sample size)—five levels (150, 500, 1000, 1500, 2000);
2. Number of score categories—four levels (2, 3, 4, and 5);
3. Kappa—two levels: high = $0.9 \times \kappa_{max}$ and low = $0.65 \times \kappa_{max}$ (where κ_{max} is the maximum possible value of kappa, given the margins); and
4. Marginal asymmetry—five levels defined below.

All factors were crossed, with the exception of combinations of two-category items with type of asymmetry. The 2×2 tables for the two-category condition only have 1 df , and so only symmetric and a single asymmetric condition (discussed below) were used. While the details of the simulation are described below, the general procedure follows that 34 population (or generating) tables were created; one for each combination of number of score categories $\times$ kappa $\times$ marginal asymmetry. Those population tables were used to generate data sets of varying sample sizes—five sizes for each of the 34 tables.

The methodological choices were informed by the operational practices from a large-scale assessment program. CR items from this large-scale assessment range from two to six categories. During scoring monitoring, the number of rescoring performed by the program varies across subjects and grades but typically include 600 to 1,600 student responses. The authors extended this range to reflect more realistic sample sizes for other testing programs, as well as to better understand the effects of changing numbers of rescoring. The marginal asymmetry conditions were informed by real scored data from a variety of subjects including math, reading, science, U.S. history, geography, civics, and economics.

One constraint on the data generation involved the value of kappa. Given the table margins and the need to allot nonzero probability to all off-diagonal cells (see the section “Data Generation”), constructing tables with the theoretic maximum levels of kappa was difficult/impossible in some conditions. Therefore, the maximum possible kappa value was multiplied by 0.9 (high) and 0.65 (low) to define the two levels of kappa.

Marginal Asymmetry

Five asymmetry conditions were defined:

- Symmetric—the margins were identical for the row and column.
- Asymmetry Low—asymmetry involving higher disagreement at the low end of the score scale as compared with the middle or high end.
- Asymmetry High—asymmetry involving higher disagreement at the high end of the score scale as compared with the middle or low end.
- Severe Asymmetry—the severe rater group assigns more scores in the lowest category. This condition is what is most often considered when rater drift is discussed.
- General Asymmetry—one rater group gives scores that are less spread out than the other group. General asymmetry is often seen as rater centrality, in which raters overuse the middle scores, avoiding the extremes.

As was noted above, for two-category items, only symmetric and a single asymmetric condition were used for the asymmetry factor. The asymmetry level used was Severe Asymmetry, as it satisfies the requirement that “one rater is more severe” definition of that condition.

Population Table Margins

To determine the margins, a large number of empirical rescore tables from a large assessment program were examined. Based upon these, tables that best exemplified each type of asymmetry were identified; the margins of these tables became the population margins for each condition. In the two-category condition, it was necessary to adjust the margins to allow for both noticeable differences between kappa levels and substantial asymmetry. This has the advantage of guaranteeing that the table margins were “realistic.” However, it has the drawback that some of the tables did not completely exemplify the corresponding asymmetry condition. The authors address this in Study 2. Margins for the tables used in each condition and the full population tables are presented in Online Appendix A (descriptive and effect size information is in Online Appendix B).

Data Generation

Data generation took place in two steps. First, for the margins defining each condition, a population rescore table was formed. Second, each replication consisted of drawing a sample table and computing the paired t test and Q . Data for the individual cells of the rescore table were generated, rather than directly generating the margins. This guaranteed that the correlations between the margins would be similar to those observed in the context of a rescore table. Using the table margins that had been selected, the population tables were generated for each combination of number of categories, asymmetry, and kappa level. Population tables were constructed to have the desired margins, have the specified kappa, minimize exact disagreement, and have nonzero entries for all cells.²

To generate the population table, the maximum kappa possible, κ_{max} , given the table margins, was computed. This served as the basis for the high kappa ($\kappa_{high} = \kappa_{max} \times 0.90$) and low kappa ($\kappa_{low} = \kappa_{max} \times 0.65$) conditions. Next, minimum off-diagonal probability was assigned to each cell (0.0125 for a score difference of 1, 0.010 for a score difference of 2, 0.007 for a score difference of 3, and 0.005 for a score difference of 4). Next, the diagonal entries were filled, proportional to the product of the corresponding row and column margins (if this allocation

would make either the row or column sum exceed the desired margin, it was reduced and the other diagonal cells proportionally increased). Finally, off-diagonal cells were filled to minimize disagreement (allocating to cells with discrepancies of 1 score point received priority over discrepancies of 2, etc.). This procedure was automated for all but the two-category item tables, which had to be adjusted manually³ to achieve the desired result.

For the data generation, cells of the tables were mapped to bins, that is, regions of the unit line segment, with length proportional to the cell's probability. For each case in a replication, a uniform $[0, 1)$ random deviate U was drawn. The authors determined into which of the bins U fell, and the corresponding cell's count was incremented by 1. Data generation was done using the Python 2.7.8 random library, which uses the Mersenne Twister (Matsumoto & Nishimura, 1998) as its uniform generator. The stream was seeded once, and then all replications were completed in a single run.

For each population table, a sample of the indicated size was drawn and agreement statistics computed for the resulting sample table. For each condition, 1,000,000 independent replications were performed. Q and paired t for each replication were compared to the nominal ($\alpha = .05$) critical value. Significance rates were obtained for each design condition.

Results

The results are presented in terms of sample size, asymmetry condition, and number of categories. The fourth variable manipulated, true kappa, had a very small effect and so is not discussed. All results presented are aggregated over the levels of kappa.

In evaluating the results, the authors focused on three research questions:

Research Question 1: Is the Type I error rate of each statistic close to the nominal α ? Does it vary by condition?

Research Question 2: Do the statistics have power to detect the violations of marginal homogeneity? Does this differ by type of asymmetry?

Research Question 3: When comparing the two statistics, do they identify the same tables? Where the two differ, does one dominate the other, or do the two yield complementary information?

Type I Error Rate

Type I error of the Q and t statistic were evaluated using data from the symmetric (i.e., homogeneous) margins. The nominal $\alpha = .05$ was used for significance testing. Type I error was computed separately for each cell in the design.

To understand which factors (sample size, number of score categories, and their interaction) affect the differences between Q and t , an ANOVA was conducted. The proportion of cases significant for Q and significant for t in each cell of the design was calculated. An arcsine transformation was applied to the difference of the proportions between Q and t . The ANOVA was conducted to determine which factors had the greatest impact on these differences in proportions. Because the purpose of the ANOVA was descriptive in nature, the F statistic and p value are intentionally not presented. Partial ANOVA results and η^2 are contained in Table 2. Based on η^2 , sample size has a larger influence on the variability in differences in proportions than number of item score categories.

Type I error results for sample size and number of categories are presented in Table 3. Sample size by number of categories is presented in Figure 1. (Type I error rates for the Null

Table 2. ANOVA Results for Sample Size and Number of Categories: Homogeneous Margin Sample.

Source	<i>df</i>	Sum of squares	η^2
Sample size	4	0.00020	0.59
Score categories	3	0.00005	0.14
Sample Size $\times$ Score Categories	12	0.00009	0.25

Table 3. Type I Error Results (in percent metric) for Sample Size and Number of Categories: Homogeneous Margin Sample.

	Sample size					Number of score categories			
	150	500	1,000	1,500	2,000	2	3	4	5
<i>Q</i>	4.46	4.98	4.97	5.00	4.99	5.26	4.73	4.74	4.80
<i>t</i>	5.08	5.12	5.05	5.04	5.03	5.27	5.03	4.99	4.98

Figure 1. Type I error rates for *Q* and *t* by sample size and category.

condition [Symmetric] for kappa level by number of score categories and by sample size are presented in Online Appendix C). From Table 3, for the sample size condition, the expectation of about 5% significant was met for *t*, but the *Q* statistic was conservative, rejecting less frequently, especially for the smallest sample size.

About 5% of the cases were significant for the *t*, while fewer were significant for the *Q*, especially for three-, four-, and five-category items, where the rejection rate was 4.7% to 4.8%. With the exception of the small sample size condition for three, four, and five categories, the *Q* and *t* function similarly and exhibit expected Type I error.

Figure 1 displays the interaction of the sample size by number of categories. The figure reveals interactions for small sample sizes (150 or 500), with the largest differences seen in the three-category item case. Again, the sensitivity of the *Q* for the smallest sample size is evident.

The Type I error rate results for both statistics were encouraging. While the *Q* is clearly more conservative at the smallest sample size, overall, both the paired *t* test and *Q* exhibit good Type I error behavior. In the two-category condition, a slight inflation of Type I error rate was observed for both statistics; otherwise, deviations were of the form of rejecting at too low a rate.

Table 4. ANOVA Results of Arcsine Differences in Proportions ($Q - t$) for Sample Size and Number of Categories: Heterogeneous Margin Sample.

Source	<i>df</i>	Sum of squares	η^2
Asymmetry condition	3	13.723	0.42
Sample size	4	1.576	0.05
Asymmetry Condition $\times$ Sample Size	12	1.688	0.05
Score Categories	3	2.551	0.08
Asymmetry Condition $\times$ Score Categories	6	6.012	0.18
Sample Size $\times$ Score Categories	12	2.267	0.07
Asymmetry Condition $\times$ Sample Size $\times$ Score Categories	24	3.680	0.11

Table 5. Power Results (in percent metric) for Sample Size, Number of Categories, and Asymmetry Condition: Heterogeneous Margin Sample.

	Sample size					Number of score categories				Asymmetry condition			
	150	500	1,000	1,500	2,000	2	3	4	5	High	Low	Severe	General
<i>Q</i>	49.86	84.46	94.05	96.64	97.93	92.67	94.21	89.41	68.13	87.40	86.81	90.10	72.23
<i>t</i>	29.15	51.46	61.99	66.06	68.61	92.67	65.13	63.35	28.57	47.07	62.02	88.70	12.94

Power

The power to detect true marginal differences for each statistic was examined. Specifically, the effect of sample size, number of score categories, asymmetry condition, and their interaction were evaluated. One reviewer noted that without accurate Type I rates, “power” cannot be strictly determined. As noted above, Type I error for Q was conservative for the smallest samples, and so it can be argued that, as a result, power cannot be determined for these conditions. For completeness, the authors present the detection results for all conditions, but the conservative nature of Q for some conditions should be borne in mind.

As was done for Type I error, an ANOVA was conducted on the arcsine-transformed differences in proportions between Q and t . The results are displayed in Table 4. The largest effects are due to asymmetry condition and the interaction of asymmetry condition with number of score categories. The main effect of number of score categories has a moderately large η^2 . Sample size had the smallest main effect. Given these results, each individual effect will be examined in turn. Table 5 contains power results for main effects of asymmetry, number of categories, and sample size conditions.

From Table 5, it is clear that the Q and t are equally powerful in detecting true differences when rater disagreement is systematic as defined by the Severe Asymmetry condition (i.e., when the time A raters are consistently more severe/lenient than those at time B). As expected, the t test performs well in this particular situation because directionally consistent differences in scoring result in mean differences between the two groups of raters. Somewhat surprisingly, Q has power equivalent to the t test in this case. When the disagreement is not consistently in the same direction, however, as in the remaining three asymmetry conditions, the Q is more powerful than the t test at detecting marginal differences.

For the number of score categories (Table 5), Q is more powerful than t for all category levels except for the two-category. This result is logical based on the previous results that the Q

Figure 2. Power for Q and t by sample size and number of score categories.

and t perform equally for the Severe Asymmetry condition. Recall that the only asymmetric condition associated with two-categories was Severe Asymmetry. Therefore, the results for two-categories represent only the Severe Asymmetry condition. Even though the η^2 table shows that categories contribute to the variability in differences in proportions between the Q and t , the result associated only with the two-categories is contributing to that estimate. If the two-category represented additional asymmetry conditions, or if all two-category items were removed, the η^2 for number of categories would be substantially lower; the ANOVA with the two-category items removed resulted in a decrease in η^2 from 0.078 to 0.006 for the number of score categories. Thus, the asymmetry condition is largely responsible for the differences.⁴

For the sample size condition (Table 5), Q is consistently more powerful than t regardless of sample size. This result is somewhat surprising, in that Q showed conservative Type I error rate for $N = 150$ and $N = 500$ cases. To investigate this further, Figure 2 contains the interaction of sample size and number of score categories. For the 150 sample size, Q demonstrates least power for the five-category items. This could be due to poorly estimated individual probabilities given the 5×5 table (with 25 cells) with only 150 sample size. In addition, the distributional properties of Q were evaluated for the interaction of sample size and number of score categories. Those results suggest that Q might not exactly follow a chi-square distribution for all of those interactions that involve a sample size of 150. In those instances, the variance of Q is slightly smaller than that of the nominal $\chi^2(4)$.

Figure 3 gives the interaction of asymmetric condition by the number of score categories. The power of paired t test drops substantially for the Asymmetry High and Asymmetry Low conditions for five-category items. This finding is examined further in the "Discussion" section.

One noteworthy feature is the difference in power between the t and Q for the General Asymmetry condition, across all categories. Because the disagreement is bidirectional, the mean difference between raters is relatively small, thus not triggering a significant t test. However, the margins are heterogeneous enough to produce a significant result for Q . As expected, Q is much more powerful than the t at detecting this type of disagreement between the two groups of raters.

Finally, across the number of score categories, the t test and Q are almost equally powerful in detecting marginal differences within the Severe Asymmetry condition. As expected, the consistent differences result in different means between the two raters, triggering a significant t

Figure 3. Power for Q and t by asymmetry condition and number of score categories.

Table 6. Percentage of Correct Rejections by Statistic.

		<i>t</i>	
		Significant	Not significant
<i>Q</i>	Significant	54.42	30.17
	Not significant	1.04	14.37

statistic. What was not obvious a priori was that the Q is *at least as powerful* as the t in detecting Severe Asymmetry disagreement, as is corroborated by results in Table 5.

Comparison of Items Identified

Operational testing programs have limited time and budget. A question of some interest is “Do the t test and Q identify the same items?” If the information is largely redundant, using one statistic to monitor scoring will suffice. If not, use of both allows testing programs to detect a broader range of scoring issues. To understand the overall difference between Q and t in terms of the items identified, Table 6 contains the percentages of significant and nonsignificant cases for both statistics when the hypothesis of marginal homogeneity was false. From the simulated data set, about 54% of the cases in asymmetric conditions were significant for both Q and t ; 14% were not significant. About 30% of the cases were significant for Q but not for t , and about 1% were significant for t , but not Q . Of the 30% that were significant for Q but not t , nearly all came from the General Asymmetry, Asymmetry High, or Asymmetry Low conditions. As for sample size, those 30% were fairly evenly distributed across the five levels. As for score categories, just less than half of that 30% were five-category items while none were two-category.

From Table 6, the authors conclude, within the parameters of the simulation, that monitoring marginal homogeneity using Q alone is likely sufficient; one gains little by using both Q and the t test as about 1% of additional items were correctly identified as having marginal asymmetry. On the contrary, Q provides a substantial improvement over just using the t test, correctly identifying 29% more items. Despite the conservative Type I error rate for Q , from Table 6, Q has a better overall hit rate than t .

Discussion

While tests of marginal homogeneity are not novel, they are not used in educational research as often as the more common correlation, kappa, or t statistic for evaluating rater agreement. Given that measures of marginal homogeneity have been recommended (e.g., Zwick, 1988), this study examined the performance of Stuart's (1955) Q . For reference, the behavior of Q was compared with that of the widely used paired t statistic. Real trend rescore tables from an educational achievement test data were used as the basis for the simulation. The purpose was to evaluate the performance of Q , and then compare Q with the more commonly used paired t test to identify asymmetry cases that one statistic detects but that the other does not.

In examining the Type I and power results, several patterns emerged. First, few differences were found between the two statistics in the Type I error results. The Q and t behaved similarly across the conditions, with the exception that the Q was somewhat conservative with the smallest sample size. This is likely due to the sensitivity of chi-square distributed statistics to sparse expected cell frequencies; the table was not dense enough for the asymptotic distribution to hold. Haberman (1988) showed that chi-square tests of significance can demonstrate bias when the expected sample size per cell is small and when the number of cells in the contingency table is large. Haberman demonstrated that even when the smallest expected cell count is at least 5, the power of chi-square tests can be quite low. Therefore, with small sample sizes (e.g., $N = 150$) and large contingency tables, the consistency of the chi-square statistic can be questionable. In addition, as the sample size increased, this conservative Type I error rate of the Q decreased. Overall, both statistics exhibited good Type I error rates.

Second, for power, type of asymmetry and number of score categories had the largest influence on the difference between Q and t . Specifically, Q was more powerful for General Asymmetry, Asymmetry High, and Asymmetry Low conditions. The two performed nearly identically for the Severe Asymmetry condition, although Study 2 (available online) found a slight power advantage for paired t in small samples.

Third, Q generally was also more powerful than the t test for larger numbers of score categories; this was especially true for five-category items. The Q and t test performed similarly for two-category items due to the fact that the only asymmetry condition simulated with two-categories was the Severe Asymmetry condition; the two statistics functioned similarly for the Severe Asymmetry condition across the design. When aggregated across all sample size conditions, Q was more powerful than the t test at detecting marginal heterogeneity.

Finally, the relatively lower power for both statistics with five-category items in Population table deserves mention. To some degree, this may be an artifact of the empirical data used to choose the marginal proportions. Most of the five-category items were from a mathematics assessment, where rater agreement is typically very high. Thus, any asymmetry in the margins probably was not heterogeneous enough to trigger mean differences. Even the power of Q drops slightly for the five-category Asymmetry High and Asymmetry Low conditions. For the five-category items, in General Asymmetry, the power of Q again drops, much like it did in the Asymmetry High and Asymmetry Low conditions. This too is likely a function of the lack of heterogeneity in the real data upon which the Population table simulation was based.

Conclusion

The Q statistic does not exceed the nominal Type I error rate. In fact, it is conservative. Even though it is conservative, Q remains at least as powerful as the t test and, in many scenarios, is more powerful than the paired t test. Clearly, Q captures differences in the margins that the t test does not. One exception involves the Severe Asymmetry case in which the two test

statistics performed similarly. In this situation, where asymmetry was unidirectional, then either the Q or t test would suffice.

Given that empirical, cross-time rescore data were used as the basis of the simulation (Population table), it is unrealistic to assume that all rater disagreement follows the pattern of consistent directional differences; nondirectional patterns *were* observed in the empirical data. While the t test appropriately detects true differences under the Severe Asymmetry condition, the Q is clearly better at detecting other patterns of disagreement that are likely to be present in real data. As a result, the Q statistic identifies more cases of true rater disagreement than does the t test. Therefore, Q should be considered in place of the t test to evaluate marginal homogeneity for raters in educational measurement settings.

Author's Note

The opinions expressed are those of the authors and do not represent views of the Educational Testing Service or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported under the National Assessment of Educational Progress (ED-IES-13-C-0017) to Educational Testing Service (ETS) as administered by the National Center for Education Statistics (NCES) within the Institute of Education Sciences, U.S. Department of Education.

Supplemental Material

Supplementary material is available for this article online.

Notes

1. This statistic is not limited to rescoring, but the factors and levels were selected to mirror conditions seen in the context of cross administration, rescore monitoring.
2. In the authors' experience, extreme disagreements are rarely (or never) observed in practice. However, in the population table, any 0 cell becomes a structural 0 for that condition, making it impossible to *ever* observe such a disagreement. Thus, the authors required that each cell in the population table have a nonzero probability of occurring.
3. Probability was moved from one diagonal cell to the other. Then off-diagonal cells were filled.
4. In the Population table margins (based on empirical rescore tables), asymmetry condition and number of categories are confounded; the difference in marginal means varies across both factors making the cause of the results ambiguous. This issue is addressed in Study 2 available online.

References

- Bhapkar, V. P. (1966). A note on the equivalence of two criteria for hypotheses in categorical data. *Journal of the American Statistical Association*, 61, 228-235.
- Bhapkar, V. P. (1979). On tests of marginal symmetry and quasi-symmetry, in two- and three-dimensional contingency tables. *Biometrics*, 35, 417-426.

- Brennan, R., & Prediger, D. (1981). Coefficient kappa: Uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687-699.
- Byrt, T., Bishop, J., & Carlin, J. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46, 423-429.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48, 333-356.
- Feinstein, A., & Cicchetti, D. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43, 543-548.
- Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, 83, 555-560.
- Landis, R., & Koch, G. (1975). A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). *Statistica Neerlandica*, 29, 151-161.
- Landis, R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago, IL: MESA Press.
- Mariano, L. T. (2002). *Information accumulation, model selection and rater behavior in constructed response student assessments* (Unpublished doctoral dissertation). Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
- Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8, 3-30.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12, 153-157.
- Muraki E., & Bock R. D. (2003). PARSCALE 4: IRT item analysis and test scoring for rating scale data [Computer program]. Chicago, IL: Scientific Software.
- Myford, C., & Wolfe, E. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46, 371-389.
- Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment. *Journal of Educational and Behavioral Statistics*, 27, 341-384.
- Stemler, S., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In L. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Thousand Oaks, CA: SAGE.
- Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, 42, 412-416.
- Uebersax, J. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101, 140-146.
- Verhelst, N., & Verstralen, H. (2001). IRT models for multiple raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89-108). New York, NY: Springer-Verlag.
- Wang, Z., & Yao, L. (2013). *The effects of rater severity and rater distribution on examinees' ability estimation for constructed-response items*. ETS RR-13-23. Princeton, NJ: Educational Testing Service.
- Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26, 283-306.
- Zwack, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103, 374-378.