Journal of MultiDisciplinary Evaluation
Volume 8, Issue 19

ISSN 1556-8180
http://www.jmde.com

Utilizing Generalizability Theory to Investigate the Reliability of the Grades Assigned to Undergraduate Research Papers

Mihaiela Ristei Gugiu, Ph.D.
The Ohio State University

P. Cristian Gugiu, Ph.D.
The Ohio State University

Robert Baldus, M.A.
Central Michigan University

Background: Educational researchers have long espoused the virtues of writing with regard to student cognitive skills. However, research on the reliability of the grades assigned to written papers reveals a high degree of contradiction, with some researchers concluding that the grades assigned are very reliable whereas others suggesting that they are so unreliable that random assignment of grades would have been almost as helpful.

Purpose: The primary purpose of the study was to investigate the reliability of grades assigned to written reports. The secondary purpose was to illustrate the use of Generalizability Theory, specifically the fully-crossed two-facet model, for computing interrater reliability coefficients.

Setting: The participants for this study were 29 undergraduate students enrolled in an introductory-level course on Political Behavior in Spring 2011 at a Midwest university.

Intervention: Not applicable.

Research Design: Students were randomly assigned to one of nine groups. Two-facet fully crossed G-study and D-study designs were used wherein two raters graded four assignments for 9 student groups—72 evaluations in total. The universe of admissible observations was deemed to be random for both raters and assignments, whereas the universe of generalization was deemed to be mixed (random for two raters but fixed for four assignments).

Data Collection and Analysis: The semester-long project was assigned to groups consisting of an annotated bibliography, survey development, sampling design, and analysis and final report. Four grading rubrics were developed and utilized to evaluate the quality of each written report. Two-facet generalizability analyses were conducted to assess interrater reliability using software developed by one of the authors.

Findings: This study found a very high interrater reliability coefficient (0.929) for only two raters who received no training in how to use the four grading rubrics.

Keywords: *grading; reliability; Generalizability Theory; writing*

Introduction

Writing is an essential component and a fundamental goal of education. According to the National Commission on Writing in America's Schools and Colleges, "writing is a complex intellectual activity that requires students to stretch their minds, sharpen their analytical capabilities, and make accurate and valid distinctions" (2003, p. 13). However, a recent study (Arum & Roksa, 2011) reported that, despite the fact that writing is a critical aspect of higher education, there is an alarmingly low level of writing assignments required in colleges nowadays. Moreover, the authors found that "many freshmen report little academic demand in terms of writing, half of seniors report that they have not written a paper longer than twenty pages in their last year of college" (Arum & Roksa, 2011, p. 37). Given the crucial role writing plays in the development of critical thinking and analytical skills along with the ability to effectively communicate one's ideas to others, it is imperative that faculty provide as many writing opportunities as possible to foster their development (Reed & Burton, 1985; Pare & Joordens, 2008).

The use of writing assignments in a course is, however, beset by several challenges. In addition to being "resource and labor intensive" for instructors, the grading of written assessments is plagued by subjectivity and uneven variability (Bell, 1980; Anatol & Hariharan, 2009). The variations in the grades assigned by the same instructor or different instructors to the same paper may be the result of several factors. According to William E. Coffman (1971), there are three categories of explanations for the variation in grades assigned to written assignments. First, instructors may employ different standards in their ratings, with some being more lenient or severe than others. Second, some instructors distribute their scores over a greater portion of the rating scale, whereas others tend to concentrate their scores around a specific value. And third, instructors may differ in the criteria employed for rating the papers. Hence, if criteria are not pre-specified in the form of a grading rubric, grades may vary even if the same instructor grades the paper twice. The relative difference in the level of difficulty of essay questions may also contribute to the variation among different instructor ratings, while also compromising the impartiality of the rating process (Barrett, 1999). Moreover, other scholars found that factors such as the student's first name and gender, the presentation of the written assignment, the language used in the essay, and the order of the paper in the pile of essays to be graded may also influence the instructors' judgments (Branthwaite, Trueman, & Berrisford, 1981; Brown, 2010). The culmination of these problems is that the reliability of grades assigned to written assignments is often very low (Hopkins, 1998).

Studies investigating the reliability of the grades assigned to written assignments date back to 1930, when, as part of a study (Eells, 1930), 61 teachers graded the same set of papers at an interval of 11 weeks. The study found that the Pearson product-moment correlations for the repeated ratings varied between 0.25 and 0.51 and, consequently, concluded that they were highly unreliable (Eells, 1930). Six years later, a different study reached a similar conclusion and reported that the agreement between pairs of five instructors rating history honors essays varied between -0.41 and 0.85 with an average of 0.44 (Hartog, Rhodes, & Burt, 1936). An even more alarming conclusion was drawn by G.M. Bull's (1956), who reported that the grading of a typical final examination essay was so unreliable that a random assignment of grades would have been almost as helpful in differentiating among the examinees. A different study (Blok, 1985) investigated the reliability of grades assigned by 16 raters, who independently graded 105 essays on two separate occasions using scale from 1 (very poor) to 10 (excellent). The study found that the estimated correlations among the scores of different raters ranged between 0.415 and 0.910, indicating a significant variability existed in the rank-order of the grades assigned by different raters to the same papers. Fair levels of interrater agreement were also reported in a study that employed data from 13 examiners and 233 answer papers (k=0.385) (Anatol & Hariharan, 2009). Similarly, the overall reliability, based on Cronbach's alpha coefficient, was 0.672.

Given the problems reported in the literature, William E. Coffman (1971) recommended the use of two raters to grade the same essay so as to improve the reliability of scoring. Rebecca Cannings and her colleagues (2005, p. 302) made a similar recommendation based on the results of two study cohorts (1990-2000 and 2002-2003), which found that the reliability of the scores assigned to student essays were 0.38 and 0.39, respectively. [1] Additionally, weighted Cohen's

[1] Reliability coefficients less than 0.7 are generally considered unacceptable. However, this level may be too low for certain decisions (Nunnally, 1978), such as assigning grades to students.

Kappa was used to measure agreement among examiners' ratings, which produced a coefficient of 0.42 between the examiners of the first cohort and 0.62 between the examiners of the second cohort.[2] In contrast, Frijns et al. (1990) found a generalizibility coefficient of 0.80 for open-ended responses marked by physician-raters, if two raters received between four and six hours of training—as reported by Kuper (2006)[3].

One way in which the reliability of writing assignments could be improved is through the use of rubrics. In addition to saving time in providing feedback (Barringer, 2008), rubrics describe the various aspects of a task, inform students about the degree of mastery required for each level of the task, and highlight the criteria upon which they will be graded on (Reed & Burton, 1985; Luft, 1997; Popham, 1997; Hafner & Hafner, 2003; Stevens & Levi, 2005). Furthermore, by providing a description of the scoring criteria in advance, rubrics may positively impact interrater reliability (Moskal & Leydens, 2000). However, even when rubrics are employed, the reliability of grading may not necessarily be very high. For example, Rennee Williams and her colleagues (1991) investigated the interrater reliability of scores obtained by physical therapy and occupation therapy tutors in rating their students' final papers. The eight raters were provided rubrics to assist them in the grading process and were asked to rate all papers on a 12-point scale[4]. The interrater reliability—ICC(2,1)[5]—of their scores was 0.79, with a 95% confidence interval of 0.49-0.93. In other words, "if a tutor in the study graded a student's written paper as a 9 (B+) using the 12-point scale, the grades of the other tutors for the same paper would be between 6 (C+) and 12 (A+) 95% of the time" (p. 684).

Given the difficulty in producing highly reliable scores in the rating of written assignments, it is not surprising that faculty "indicated more concern about the grading or marking of student assignments than about any other aspect of their jobs, with the exception of tenure and salary... [and] they felt the marks given [by faculty] were often unreliable and frequently inaccurate" (Orpen, 1980, p. 567). However, written assignments are an essential component of higher education. Therefore, it is imperative that faculty, teachers, instructors find ways to minimize deficiencies in the grading of essays and to maximize the likelihood of producing highly reliable scores. The primary purpose of this study is to further investigate the reliability of ratings assigned by instructors to student written assignments. While past studies have focused mostly on essay examinations, the present study focused on written research projects employed in a political science course. The secondary purpose of this study is to introduce readers to modern measures of reliability. Correlations between raters and indices of agreement have fallen out of favor with psychometricians. Presently, psychometricians recommend utilizing intraclass correlations (ICCs) to compute reliability. Of the numerous existing ICCs methods, Generalizability Theory—particularly the two-facet model illustrated in this study—is considered among the most power due to its ability to account for multiple sources of error (known as facets) and its flexibility in modeling the universe of generalizability of interest to the test developer. Furthermore, Generalizability Theory enables one to estimate the number levels (sample size) necessary for each facet in order to attain a desired reliability level.

Brief Overview of Reliability Estimators

Since the concept of reliability was first introduced by Charles Spearman in his treatise on classical test theory over 100 years ago (Alexopoulos, 2007), many reliability estimators have been developed based on the general definition: the ratio of the true score variance to the observed score variance. These various conceptualizations of reliability can be classified into one of four groups (Crocker & Algina, 1986; Hopkins, 1998; Gugiu, 2011): stability, internal consistency, interrater reliability, and criterion reliability, where the first

[2] By convention, Kappa coefficients in the range of 0.41-0.60 are considered moderate, whereas coefficients in the range 0.61-0.80 are considered substantial (Landis & Koch, 1977). Although these coefficients are acceptable, Kappa is a measure of agreement, not reliability. Moreover, Kappa coefficients are subject to numerous constraints (Sadler & Good, 2006).

[3] A copy of the original Frijns et al. study could not be obtained by the present authors.

[4] Although the goal of the rating scale employed by Williams, Sanford, Stratford, and Newman (1991) was intended to correspond to a 12-letter grading system ('A', 'A-', 'B+', and so on to 'F'), the raters did not actually assign letter grades to papers. Hence, there is no way to know the impact of actually assigning a letter grade on reliability.

[5] Based on the work of Shrout & Fleiss (1979), an ICC(2,1) is a reliability estimate for a single rater selected at random from the population of all possible raters.

two represent classical definitions of reliability and the latter two represent modern measures of reliability.

Stability estimators (i.e., test-retest method, coefficient of equivalence, and alternative form method) are designed to measure the ability of a test or method to yield consistent results measured at two points in time. A test is considered reliable if it produces similar results, as determined by a Pearson or Spearman-rank correlation. Since the development of modern measures of reliability, however, psychometricians no longer recommend the use of correlation coefficients to measure interrater reliability. Correlation coefficients were designed to measure the degree to which two variables are linearly related. Therefore, two variables (or in this case raters) can attain a high correlation if the rank-order of the cases are relatively invariant. That is, the mean grade assigned by an instructor to an essay, for example, is highly correlated to the grade assigned by another instructor if the rank-order of all the essays graded by the first instructor is similar to that of the second one. In fact, regardless of the discrepancy between the two instructor grades, the correlation will equal unity if the intervals between all the rank-orders are equal for the two groups.

Internal consistency estimators, such as coefficients alpha (Cronbach, 1951) and omega (McDonald, 1999), measure the degree to which test items or a set of indicators are interrelated based on a single administration of the test or survey. Although these indices assume the items or indicators are continuous, which is not always the case, recent developments in measurement theory have resulted in ordinal versions of alpha and omega (Zumbo, Gadermann, & Zeisser, 2007). However, although these reliability estimators are well-suited for measuring the degree to which test items are sampled from the same content domain, their interpretability with regard to multiple raters is unclear.

Consequently, psychometricians prefer to use modern estimators of reliability due to their ability to directly measure the ratio of the true score to observed score variance. These modern estimators include interrater reliability and criterion reliability estimators. *Interrater reliability* estimators measure the degree to which the ratings made by different raters evaluating the same test agree with each other. Classically, the percentage of agreement among raters, Cohen's Kappa for two raters (Stemler, 2007), and multiple-rater Kappa (Fleiss, 1981) have been employed to measure interrater reliability. Although informative, these estimators have numerous limitations and are conceptually different than the accepted definition of reliability. Agreement rates are impacted by inconsistencies in the definition of what constitutes agreement (exact agreement versus agreement within a margin of error), sensitivity to the number of grading categories (the more categories, the lower the agreement rate), probability of agreement due to chance, and an insensitivity to the magnitude of individual differences (Sadler & Good, 2006). Although the Kappa statistics improve upon agreement rates by accounting for chance, the other objections still hold true. Hence, nowadays, psychometricians advocate the use of either the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979) or the generalizability coefficient (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001a), which are both derived from the ANOVA model. In general, these estimators are only appropriate for relative decisions, such as is the case when the relative standing of scores are compared to each other (i.e., when rank-order matter).

The *criterion reliability* estimators, such as the index of dependability (Brennan & Kane, 1977), measure the degree to which the classification decision of a method is consistent with the decisions that would result from replicating the study under parallel conditions. In contrast to ICCs and generalizability coefficients, these estimators are appropriate for absolute decisions, as is the case in criterion-referenced situations and mastery tests. That is, criterion reliability is employed to determine the consistency of a classification system or test relative to a single fixed cut-score. For example, one would employ criterion reliability to assess the likelihood that given a student's unknown true score, their observed score would fall below 60 percent (standard demarking pass/fail) if the test were replicated under parallel conditions. Obviously, the most appropriate reliability estimator is dictated by its intended use and interpretation, as is the case with all instruments and methods. For the present study, Generalizability Theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001a) was employed to compute both interrater and criterion reliability estimates for instructor ratings.

Methodology

Participants

The participants for this study were 29 undergraduate students enrolled in an introductory-level course on Political Behavior in

Spring 2011 at a Midwest university. The class was almost equally divided by gender, with 51.9 percent of students being male. Not surprisingly, 59.3 percent of students were freshman, 22.2 percent were sophomore, and the remaining 18.5 percent were junior. Moreover, the average ACT score for the class was 21.21, which was in line with the national average ACT scores in 2011 (ACT, 2011). Students were randomly assigned to one of nine groups that remained unchanged throughout the semester-long research project. Hence, nine groups produced a paper for each of the four tasks—72 evaluations in total.

Instrument

The semester-long project was composed of four assignments. The first assignment was to produce an annotated bibliography consisting of 20 scholarly articles on a topic of relevance to Political Science that could be measured via a survey administered to undergraduate students at the same Midwestern University. The second assignment was to design a 20-item questionnaire on the topic selected by the group and approved by the instructor (MRG), to pilot test and refine the survey instrument. The third assignment was to develop either a probability (systematic, simple, stratified, or cluster) or non-probability (quota, convenience, purposive, or snowball) sampling strategy and survey administration protocol. And, the fourth project required students to administer the survey instrument using the approved sampling design protocol, descriptively analyze, and write-up the results. Each of these assignments resulted in a group paper ranging in length from 2.5 to 10 pages (double-spaced).

Four grading rubrics (see Appendix) were developed (by MRG and PCG) to assess the quality of each assignment. Each of the rubrics broke down the tasks pertaining to each assignment into objective criteria based upon the guidelines found in the literature on development if rubrics (Stevens & Levi, 2005). All papers were scored on a 100-point scale and then converted to letter grades, where the range [93,100] denoted the grade A, [90,93) denoted an A-, [87,90) denoted a B+, [83,87) denoted a B, [80,83) denoted a B-, [77,80) denoted a C+, [73,77) denoted a C, [70,73) denoted a C-, [67,70) denoted a D+, [63,67) denoted a D, [60,63) denoted a D-, and [0,60) denoted an E (failure). Note that square brackets include the adjacent number while parentheses exclude the adjacent number from the range.

Procedure

The two instructors (raters) assigned to evaluate all the papers (MRG and RB) were provide with the description of each assignment and its accompanying grading rubric. No additional training was provided to either rater. Due to the limited number of available analytical software options for performing a Generalizability Theory analysis, one of the authors (PCG) developed a new analytical program based on the PROC IML SAS 9.2 platform.[6] This program was validated against the psychometric theory and examples provided in Cronbach et al. (1972), Crocker and Algina (1986), and Brennan (2001b).

Study Design

The generalizability study was composed of two parts: a G-study and a D-study (Brennan, 2001a). A G-study is similar to a pilot study that utilizes a specific study design (e.g., fully crossed) and is conducted under a set of conditions, known as the *universe of admissible observations*, defined by the investigator based on his or her assumption of whether the model variables are fixed, random, or mixed. The D-study represents the study design and conditions, known as the *universe of generalization* (i.e., the population and conditions to which the researcher wants to generalize the results) under which the study was conducted in the future. Based on these conditions and the variance estimates obtained in the G-study, the researcher can compute a generalizability (reliability) coefficient.

The design employed in this study conforms to what is known in generalizability terminology as a two-facet fully crossed G-study design ($p\times\alpha\times\beta$), where p denotes the object of measurement (i.e., the nine group topics), α denotes the two instructors or raters assigned to evaluate the papers, and β denotes the four tasks or assignments. The cross symbol indicates that each group paper is rated by the same set of randomly selected instructors for all four tasks. This study only presents the results for a fully crossed D-study design ($p\times A\times B$), although the use of a fully crossed G-study design would have allowed computation of reliability estimates for six two-facet designs. Furthermore, the rater facet (A) was assumed to be random, while the task facet (B) was assumed to be fixed. This makes sense given that, in the experience of the present authors, it is

[6] SAS code is available from the second author upon request.

far more likely that raters will change in future studies than for one to drastically alter the grading rubrics once they have been created. Note, the object of measurement, in this case the groups (p), is always considered to be random.

Results

The grades assigned by the two instructors to each group for the four assignments are summarized in Table 1. An examination of the means of these grades (67.9 for assignment 1, 88.3 for assignment 2, 84.1 for assignment 3, 79.1 for assignment 4, and 79.85 for the grand mean) revealed that significant variability in the scores assigned by instructors existed across the four assignments. This finding is in line with Robert L. Brennan's (2000, p. 348) conclusion that "virtually all available research on [performance assessments] suggests that generalizing over tasks is an error-prone activity, no matter how well the tasks are designed." Nonetheless, instructors who employ multiple written assignments encounter this problem every time they assign students a single grade at the end of the course. Hence, Generalizability Theory was used to investigate the impact of the assignment and rater facets on the reliability coefficients.

Table 1
Grades Assigned by Instructors to Each Group for Each Assignment

	Assignment 1		Assignment 2		Assignment 3		Assignment 4	
Group	R1	R2	R1	R2	R1	R2	R1	R2
1	48.0	30.0	83.0	76.8	94.0	83.0	58.0	66.0
2	73.0	75.8	91.0	85.5	95.0	92.0	86.0	92.0
3	51.0	44.3	97.0	87.3	60.0	65.3	73.0	70.0
4	77.8	67.8	95.0	81.5	90.0	89.5	92.0	84.0
5	76.0	64.5	89.0	84.0	87.0	86.5	82.0	75.0
6	71.0	74.0	93.0	89.0	87.0	76.5	83.0	78.0
7	78.8	81.5	97.0	90.3	90.0	96.0	87.0	88.0
8	76.5	74.8	87.0	79.0	92.0	95.0	62.0	73.0
9	77.0	81.3	92.0	92.0	61.0	74.0	90.0	84.0

Table 2 presents the results of the two-facet Generalizability analysis. The top part of the table summarizes the ANOVA results for the ($p\times\alpha\times\beta$) G-study design (top left) and the ($p\times A\times B$) D-study design (top right). The D-study design numbers were computed based on 2 instructors, 4 assignments, and the assumptions that the instructors would differ across courses (i.e., facet A is random), whereas the same four grading rubrics would be employed in future classes (i.e., facet B is fixed). The lower half of the table presents true score and error variances (bottom left) and reliability coefficients (bottom right). More precisely, $\sigma^2(\nu_\tau)$ represents the true score variance in grades for the nine groups after the error variance associated with the variability in assignments and raters is removed. The relative error variance, denoted by $\sigma^2(\delta)$, is used for norm-referenced comparisons (i.e., comparison of the grade of one group with the grade of another group), whereas the absolute error variance, denoted by $\sigma^2(\Delta)$, is used for criterion-referenced comparisons (i.e., comparison of the grade of one group to a single fixed standard). The variability in the observed grades for the nine groups is denoted by $\mathbb{E}S^2(p)$ and the error variance for the class average across all nine groups and four assignments is denoted by $\sigma^2(\varepsilon)$ (i.e., the error variance one would employ to construct a confidence interval on the class grand average). Finally, $S/N(\delta)$ and $S/N(\Delta)$ are measures of the amount of signal to noise.

The aforementioned variance estimates can be used to compute two reliability coefficients: the generalizability coefficient $\mathbb{E}\rho^2$ and the dependability index $\Phi(\lambda)$. The generalizability coefficient is the equivalent of the parallel test reliability used in classical test theory. That is, the generalizability coefficient is equal to the ratio of the true score variance $\sigma^2(\nu_\tau)$ to the observed score variance $\mathbb{E}S^2(p)$. The index of dependability is a measure of criterion reliability and denotes the probability that the absolute decision, resulting from a comparison of a group's grade to a standard (λ), would replicate if the written assignments were graded *ad infinitum* by a random set of instructors under parallel conditions. Therefore, the index of dependability is a function of the location of the standard (Brennan & Kane, 1977), where the closer the standard is to the grand mean, the lower the index (likelihood) will be that the unknown universe score underlying the

Table 2
Two-Facet Generalizability Analysis

Source	SS	DF	MS	GVAR	Percent	Source	DVAR	Percent
p	3,457.24	8	432.156	32.333	16.3	*p*	32.333	44.8
α	125.22	1	125.215	1.966	1.0	*A*	0.983	1.4
β	4,179.12	3	1,393.040	66.978	33.7	*B*	16.745	23.2
*p*α	246.94	8	30.868	2.411	1.2	*pA*	1.206	1.7
*p*β	3,932.35	24	163.848	71.312	35.9	*pB*	17.828	24.7
αβ	134.42	3	44.805	2.620	1.3	*AB*	0.328	0.5
Error (*p*αβ)	509.38	24	21.224	21.224	10.7	Error (*pAB*)	2.653	3.7
Total	12,584.66	71	177.249	198.844	100.0	Total	72.072	100.0

Note 1: *p*=Group; α, *A*=Instructors; and β, *B*=Assignments.
Note 2: GVAR represents the (*p*×α×β) G-study variance, where α and β are random.
Note 3: DVAR represents the (*p*×*A*×*B*) D-study variance, where *A* is random, *B* is fixed, n'_{α}=2, and n'_{β}=4.

Model & error variances		Reliability coefficients	
$\sigma^2(\nu_\tau)$	50.161	Generalizability $\mathbb{E}\rho^2$	0.929
$\sigma^2(\delta)$	3.859	Dependability Φ(λ=79.85)	0.907
$\sigma^2(\Delta)$	5.169		
$\mathbb{E}S^2(p)$	54.019		
$\sigma^2(\varepsilon)$	7.313		
S/N(δ)	13.000		
S/N(Δ)	9.704		

Note 4: Tau $\sigma^2(\nu_\tau)$ represents the universe score variance.
Relative error variance $\sigma^2(\delta)$ is used for norm-referenced comparisons.
Absolute error variance $\sigma^2(\Delta)$ is used for criterion-referenced comparisons.
$\mathbb{E}S^2(p)$ represents the expected observed score variance for the *p* mean scores if randomly parallel forms of the procedure are administered.
Mean error $\sigma^2(\varepsilon)$ represents the error variance for the grand mean across all the assignments and groups.
S/N(δ) represents the relative signal-to-noise ratio.
S/N(Δ) represents the absolute signal-to-noise ratio.
Note 5: Generalizability is a norm-referenced reliability coefficient.
Dependability is a criterion-referenced reliability coefficient, where the cut-score λ is set to the mean.

composite average of a randomly selected group would be correctly classified relative to the standard, and vice versa.

An inspection of the results from Table 2 confirmed that the assignment facet (β) had a significant contribution to the total G-study variance, $\sigma^2(\nu_\beta)+\sigma^2(\nu_{p\beta})+\sigma^2(\nu_{\alpha\beta})$=66.999+71.316 +2.623=140.938 (70.9%). However, if one averages the grades over the number of levels in the corresponding D-study facet, this variability is reduced to $\sigma^2(\nu_B)+\sigma^2(\nu_{pB})+\sigma^2(\nu_{\alpha\beta})$=16.750+17.829 +0.328=34.907 (48.4%). Additionally, the error variance was reduced from 10.7% in the G-study to 3.7% in the D-study. Averaging across the levels of the two facets and eliminating the variance attributed exclusively to the assignments leads to an increase in the variance attributed to the object of measurement, *p*, comparative to the observed score variance. That is, the ratio of the true score variance, $\sigma^2(\nu_p)+\sigma^2(\nu_{pB})$=32.324+17.829=50.153, to the total score variance, $\sigma^2(\nu_p)+\sigma^2(\nu_{pB})+\sigma^2(\nu_{pAB})$ =50.153+2.652=52.805, equals to 0.929, known as the generalizability coefficient $\mathbb{E}\rho^2$. Thus, if one removes the error variance related to instructors and assignments, the estimate of reliability obtained is not only higher but also more precise. Note then, by averaging across the levels for the two facets and computing $\mathbb{E}\rho^2$, one can generate Figure 1.

Although faculty are more likely to make norm-referenced comparisons since the goal is to assign grades that are able to discriminate among student performance on written assignments, sometimes it is beneficial to estimate the reliability of a decision with respect to a fixed standard. By similar reasoning to that utilized in computing $\mathbb{E}\rho^2$, one can estimate the index of dependability Φ(λ). Specifically, when the cut-score λ is set at the sample mean (79.85), $\Phi(\lambda)\equiv\sigma^2(\nu_\tau)/[\sigma^2(\nu_\tau)+\sigma^2(\Delta)]$ =50.153/(50.153+5.169)=0.907.

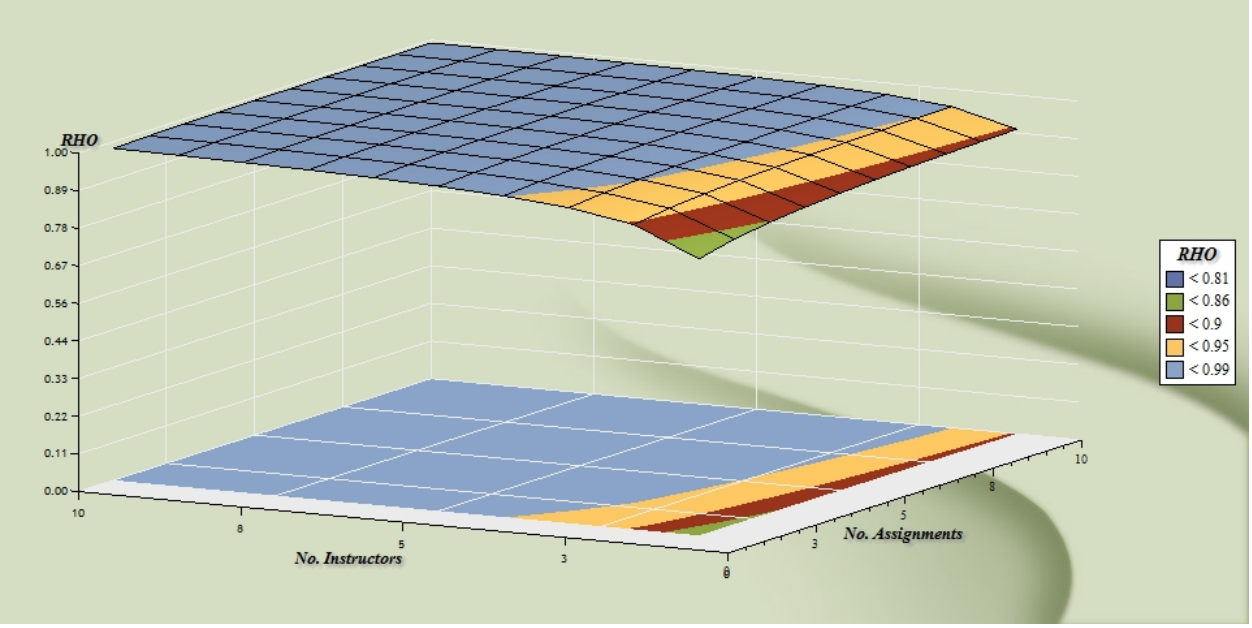

Figure 1. Objective Tree and Dashboard Format: an Objective Tree that Breaks Down the Objectives in Order to Evaluate

Note: Rho denotes the reliability (generalizability) coefficient.

Likewise, one can compute the error-tolerance ratio (Brennan, 2001b). However, faculty may be interested in setting the cut-score at 60, which usually demarks the line between passing and failing a test. One can estimate the reliability that the unknown composite true score falls below or above the fixed standard using the following formula, $\Phi(\lambda) \equiv [\sigma^2(\nu_p)+(\bar{X}-\lambda)^2-\sigma^2(\varepsilon)]/[\sigma^2(\nu_p)+(\bar{X}-\lambda)^2+\sigma^2(\Delta)-\sigma^2(\varepsilon)]$. In the case of the cut-score set at the pass/fail standard ($\lambda=60$), the index of dependability was equal to $\Phi(60)=[32.324+(79.85-60)^2-7.313]/[32.324+(79.85-60)^2+5.169-7.313]=0.988$. As this index indicates, the probability of misclassifying the unknown universe grade for a chosen group was extremely low with respect to the standard. Furthermore, by imputing various levels for the cut-score λ, one can generate a graph similar with the one in Figure 2.

An examination of the criterion reliability graph showed that it exhibited the classic V-shaped curve and that as the standard was set further away from the mean, the criterion reliability increased. Nonetheless, even at the lowest point (i.e., the sample mean of 79.85), the criterion reliability exceeded the highest reliability reported in the literature of 0.79 (Williams, Sanford, Stratford, & Newman, 1991). The noise to signal graph exhibited the reversed V-shape and it indicated that as the standard approached the sample mean, the inability of the grading rubrics to correctly classify groups based on the observed composite scores decreased.

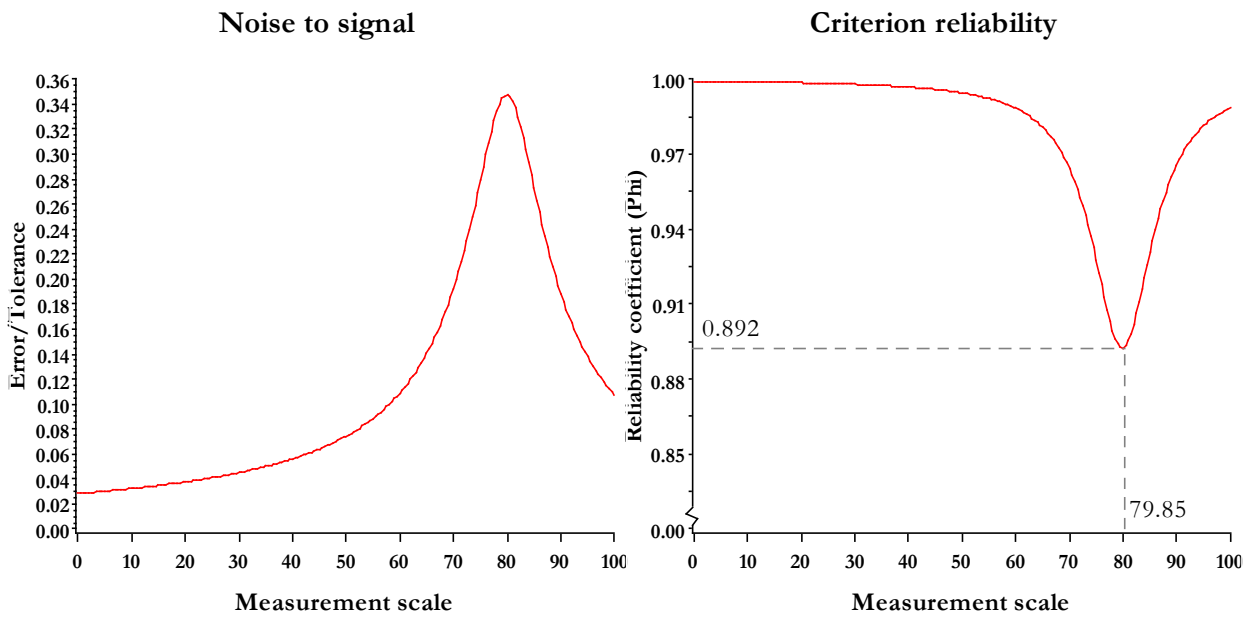

Figure 2. Error-Tolerance and Criterion-Reliability for the Two-Facet Instructor Rater Study

Discussion

As highlighted in the introduction, a number of studies have investigated the reliability of the scores or grades assigned to written papers. The results of these studies have been inconclusive with some studies reporting low levels of reliability (Eells, 1930; Cannings, Hawthorne, Hood, & Houston, 2005; Anatol & Hariharan, 2009), a few studies reporting good levels of reliability (Williams, Sanford, Stratford, & Newman, 1991; Frijns, van der Vleuten, Verwijnen, van Leeuwen, & Wijnen, 1990), and yet other studies reporting mixed levels of reliability (Hartog, Rhodes, & Burt, 1936; Blok, 1985). This uncertainty has led several researchers to argue against the use of written assignments. The present study contributes to this body of knowledge by providing a plausible explanation for some of the contradictory results. Namely, most of the previous reliability studies employed analytical techniques that, by today's psychometric standards, are antiquated (e.g., kappa, correlations, Cronbach's alpha). Generalizability Theory, particularly the two-facet model, is arguably the most (or one of the most) sophisticated method(s) for estimating reliability. Yet, a review of the literature revealed that only two studies reported an intraclass correlation coefficient, both of which were equivalent to a one-facet generalizability coefficient[7]. Furthermore, both studies used a scoring system as an index for a grading system. However, it is conceivable the psychological act of assigning grades has a different impact on reliability than the act of assigning scores. Therefore, the reliability coefficient reported herein is a "cleaner" measure of the reliability involved in *grading* students because it controlled for two sources of measurement error (raters and assignments).

The present study demonstrated the grading of written assignments can be very reliable (0.929) even when only two instructors (i.e., raters) are employed, provided that clear grading criteria are used (in the form of a rubric) and an appropriate study design is implemented. It is important to stress that, unlike the results reported in previous studies, neither of the raters employed in this study received prior training on how to use the rubrics. Hence, it is conceivable (in fact, quite probable) that a slightly higher reliability coefficient could have been attained had both raters received such training.

As expected, reliability was a function of both the number of raters and the number of written assignments. Figure 1 illustrates that increasing in the number of written assignments and the number of raters produces higher levels of reliability. Therefore, instead of using a single major written project during a semester, faculty

[7] Note, a one-facet generalizability coefficient is generally equivalent to coefficient alpha.

should divide it into several smaller assignments. This approach has the advantage of not only improving reliability, but it gives students the opportunity to practice their writing and analytical skills multiple times during the semester and to see their progress from one assignment to the next. Of course, faculty can assign only a limited number of written assignments per semester since they are time-consuming to grade and the increase in reliability is very small when more than five assignments are used per semester. One should also employ as many raters (e.g., Teaching Assistants) as possible. That said, the reliability attained for these four rubrics was so high, one could even consider employing a single rater with only a minor degradation resulting in the generalizability coefficient (0.867).

Written assignments are an important part of academic education and, despite their inherent weaknesses and the decreasing use in recent years, they remain an important tool for assessing educational achievement. Therefore, it is imperative that means of increasing the reliability of scoring such assignments are found. This study does not claim to have found *the* solution to the problem, but it successfully showed that under proper conditions and by employing the appropriate study design, very high levels of reliability can be attained for grading written assignments. Furthermore, it illustrated the use of Generalizability Theory for estimating interrater reliability. Hence, it would behoove researchers not familiar this analytical technique to explore its many benefits.

References

ACT. (2011). *2011 ACT National and State Scores*. Retrieved December 23, 2011, from ACT: http://www.act.org/newsroom/data/2011/profilereports.html

Alexopoulos, D. S. (2007). Classical Test Theory. In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (pp. 140-143). Thousand Oaks: CA: SAGE Publications.

Anatol, T., & Hariharan, S. (2009). Reliability of the Evaluation of Students' Answers to Essay-type Questions. *West Indian Medical Journal, 58*(1), 13-16.

Arum, R., & Roksa, J. (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.

Barrett, S. (1999). Question Choice: Does Makers Variability Make Examinations a Lottery? *Cornerstones : What Do We Value in Higher Education?* (pp. 1-17). Melbourne, Australia: HERDSA Annual International Conference, July 12-15.

Barringer, S. A. (2008). The Lazy Professor's Guide to Grading: How to Increase Student Learning While Decreasing Professor Homework. *Journal of Food Science Education, 7*, 47-53.

Bell, R. C. (1980). Problems in Improving the Reliability of Essay Marks. *Assessment & Evaluation in Higher Education, 5*(3), 254-263.

Blok, H. (1985). Estimating the Reliability, Validity, and Invalidity of Esssay Ratings. *Journal fo Educational measurement, 22*(1), 41-52.

Branthwaite, A., Trueman, M., & Berrisford, T. (1981). Unreliability of Marking: Further Evidence And a Possible Explanation. *Education Review, 33*(1), 41-46.

Brennan, R. L. (2000). Performance Assessments From The Perspective of Generalizability Theory. *Applied Psychological Measurement, 24*(4), 339-353.

Brennan, R. L. (2001a). *Generalizability Theory*. New York: Springler-Verlag New York, Inc.

Brennan, R. L. (2001b). *Statistics for Social Science and Public Policy*. New York: Springer.

Brennan, R. L., & Kane, M. T. (1977). An Index of Dependability for Mastery Tests. *Journal of Educational Measurement, 14*(3), 277-289.

Brown, G. T. (2010). The Validity of Examination Essays in Higher Education: Issues and Responses. *Higher Education Quarterly, 64*(3), 276-291.

Bull, G. M. (1956). An Examination of the Final Examination in Medicine. *The Lancet, 271*, 368-372.

Cannings, R., Hawthorne, K., Hood, K., & Houston, H. (2005). Putting Double Marking to The Test: A Framework to Assess If It Is Worth The Trouble. *Medical Education, 39*(3), 299-308.

Coffman, W. E. (1971). On the Reliability of Ratings of Essay Examinations in English. *Research in the Teaching of English, 5*(1), 24-36.

Crocker, L., & Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. Belmont, CA: Wadsworth Group/Thomson Learning.

Cronbach, L. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika, 16*(3), 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of*

Generalizability for Scores and Profiles. New York: John Wiley & Sons.

Eells, W. C. (1930). Reliability of Repeated Grading of Essay Type Examinations. *The Journal of Educational Psychology, 21*(1), 48-52.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2 ed.). New York: John Wiley & Sons.

Frijns, P., van der Vleuten, C., Verwijnen, G., van Leeuwen, Y., & Wijnen, W. (1990). The effect of structure in scoring methods on the reproducibility of scores of tests using open-ended questions. In W. Bender, R. Hiemstra, A. Scherpbier, & R. Zwierstra (Ed.), *Proceedings of the Third International Conference on Teaching and Assessing Clinical Competence* (pp. 466–471). Groningen: BoekWerk Publications.

Gugiu, P. C. (2011). *Summative Confidence*. Unpublished doctoral dissertation, Western Michigan University, Kalamazoo, MI.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative Analysis of The Rubric as an Assessment Tool: An Empirical Study of Student Peer-group Rating. *International Journal of Science Education, 25*(12), 1509-1528.

Hartog, P., Rhodes, E., & Burt, C. (1936). *The Marks of Examiners*. Londonh: Macmillan.

Hopkins, K. D. (1998). *Educational and Psychological Measurement and Evaluation* (8th ed.). Boston, MA: Allyn and Bacon.

Kuper, A. (2006). Literature and Medicine: A Problem of Assessment. *Academic Medicine, 81*(10), 128-137.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

Luft, J. (1997). Design Your Own Rubric. *Educational Leadership, 20*(5), 25-27.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Moskal, B., & Leydens, J. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research and Evaluation, 7*(10), 1-11.

National Commission on Writing in America's Schools and Colleges. (2003). *The Neglected "R": The Need for a Writing Revolution*. New York, NY: College Board.

Nunnally, J. C. (1978). *Psychometric Theory*. New York: McGraw-Hill.

Orpen, C. (1980). What Lecturers Like And Dislike About Their Jobs. *Unpublished manuscript*. Johannesburg, South Africa: Department of Psychology, University of Witwatersrand.

Pare, D. E., & Joordens, S. (2008). Peering Into Large Lectures: Examining Peer And Expert Mark Agreement Using PeerScholar, An Online Peer Assessment Tool. *Journal of Computed Assisted Learning, 24*, 526-540.

Popham, J. W. (1997). What's Wrong--and What's Right--with Rubrics. *Educational Leadership, 55*(2), 72-75.

Reed, M. W., & Burton, J. K. (1985). Effective and Ineffective Evaluation of Essays: Perceptions of College Freshmen. *Journal of Teaching Writing, 4*(2), 270-283.

Sadler, P. M., & Good, E. (2006). The Impact of Self- and Peer-Grading on Student Learning. *Educational Assessment, 11*(1), 1-31.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 86*(2), 420-428.

Stemler, S. E. (2007). Cohen's Kappa. In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (pp. 164-165). Thousand Oaks, CA: SAGE Publications.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to Rubrics*. Sterling, VA: Stylus Publishing.

Williams, R., Sanford, J., Stratford, P. W., & Newman, A. (1991). Grading Written Essays: A Reliability Study. *Physical Therapy, 71*(9), 679-686.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal Versions of Coefficient Alpha and Theta for Likert Rating Scales. *Journal of Modem Applied Statistical Methods, 6*(1), 21-29.

Appendix

Assessment Rubric for Group Project 1

Dimension	Evaluation criteria	Deductions
General Format Requirements ***4 points***	1 inch margin all around (1 pt.)	____ out of 1
	Double space (1 pt.)	____ out of 1
	Font: Times New Roman (1 pt.)	____ out of 1
	Font size: 12 (1 pt.)	____ out of 1
Quality of writing ***20 points***	*Writing style and organization* Deduct 1 point for every poorly written paragraph. For example, deduct a point if the paragraph is difficult to read, or the writing does not flow, or lacks a logical framework, or is "fluffy," as well as anything else that you feel negatively impacted the writing of the paragraph. (max 10 pt.)	___ out of 10
	Spelling and punctuation Deduct 1 point for every spelling and punctuation mistake. (max 5 pt.)	____ out of 5
	Grammar Deduct 1 point for every grammar mistake. (max 5 pt.)	____ out of 5
Citation Style ***20 points***	*Chicago Manual of Style* Deduct 1 point for every incorrectly cited, incomplete, or missing citation. (max 20 pt.)	___ out of 20
Content ***26 points***	*20 articles* Deduct 1 point for every missing article. (max 5 pt.)	____ out of 5
	Peer-reviewed journal Deduct 1 point for every article that is not from a legitimate journal. If you cannot tell by the journal title in the citation, you should Google it. Blogs, the internet, newspaper articles, movies, documentaries, and so on are not legitimate sources. (max 6 pt.)	____ out of 6
	Summary length Deduct 1 point for every summary whose length was not one paragraph. This goes for both too short or more than one paragraph. (max 5 pt.)	____ out of 5
	Adequate summary Deduct 1 point for every summary that did not state the main points or issues raised by the author. (max 5 pt.)	____ out of 5
	Coherent topic Deduct 1 point for every article that is unrelated to the stated topic of interest or the majority of other articles if topic is not stated. (max 5 pt.)	____ out of 5
Analysis ***30 points***	*Opinion* Deduct 1 point for every summary that did not include the group's opinion regarding at least 1 of the main points or issues raised by the article. (max 10 pt.)	___ out of 10
	Agreement with article For each article for which the group agreed with at least 1 of the main points, deduct 1 point if the summary did not explain why the group agreed. (max 10 pt.)	___ out of 10
	Counter examples/Logical weakness For each article for which the group disagreed with at least 1 of the main points, deduct 1 point if the summary did not provide a counter-example or an explanation of why the argument or logic supporting it was weak. (max 10 pt.)	___ out of 10

Total Score: 100 – Total Deductions = _____

Assessment Rubric for Group Project 2

Dimension	Evaluation criteria	Deductions
General Format Requirements ***5 points***	1 inch margin all around (1 pt.)	____ out of 1
	Double space (1 pt.)	____ out of 1
	Font: Times New Roman (1 pt.)	____ out of 1
	Font size: 12 (1 pt.)	____ out of 1
	Survey is limited to a single page (1 pt.)	____ out of 1
Quality of writing ***25 points***	*Writing style* Deduct 1 point for every poorly written paragraph in the body of the paper (this excludes the surveys). For example, deduct a point if the paragraph is difficult to read, or the writing does not flow, or lacks a logical framework, or is "fluffy," as well as anything else that you feel negatively impacted the writing of the paragraph. (max 10 pt.)	____ out of 10
	Organization Deduct up to 5 points if the paper does not have a clear introduction, body, or conclusion. (max 5 pt.)	____ out of 5
	Spelling and punctuation Deduct 1 point for every spelling and punctuation mistake in the paper (excluding the surveys). (max 5 pt.)	____ out of 5
	Grammar Deduct 1 point for every grammar mistake in the paper (excluding the surveys). (max 5 pt.)	____ out of 5
Content ***40 points***	*Number of questions* Deduct 1 point for every missing question (demographic questions do NOT count as part of the survey). (max 15 pt.)	____ out of 15
	Response scale Deduct 1 point for each question that uses a different response scale than the majority of survey questions. (max 5 pt.)	____ out of 5
	Subjects If the survey was pilot-tested on less than 10 people, deduct 1 point for each missing individual. If the number of persons to whom the survey was administered is not mentioned, deduct all 10 points. (max 10 pt.)	____ out of 10
	Coherent topic Deduct 1 point for every question that is unrelated to the stated topic of the survey. (max 10 pt.)	____ out of 10
Analysis ***30 points***	*Identification of Subjects* Deduct 5 points if no statement is provided regarding the process by which individuals were recruited for participating in the pilot study. (max 5 pt.)	____ out of 5
	Feedback Summary Deduct 5 points if no summary is provided of the feedback they received from respondents. (max 5 pt.)	____ out of 5
	Changes to Pilot Survey Deduct 1 point for every survey question that was significantly revised without providing an explanation as to why the revision was made. (max 5 pt.)	____ out of 5
	Final Survey For each question on the final survey, deduct 1 point if it contains a spelling or grammar mistake, is redundant with another question on the survey, is biased, potentially offensive to readers, or unrelated to the stated purpose of the survey. (max 15 pt.)	____ out of 15

Total Score: 100 – Total Deductions = _____

Assessment Rubric for Group Project 3

Dimension	Evaluation criteria	Deductions
General Format Requirements ***5 points***	1 inch margin all around (1 pt.)	____ out of 1
	Double space (1 pt.)	____ out of 1
	Font: Times New Roman (1 pt.)	____ out of 1
	Font size: 12 (1 pt.)	____ out of 1
	Deduct 1 point if the paper is 1/2 page shorter than or longer than 2 pages. (1 pt.)	____ out of 1
Quality of writing ***25 points***	*Writing style* Deduct 1 point for every poorly written paragraph in the body of the paper. For example, deduct a point if the paragraph is difficult to read, or the writing does not flow, or lacks a logical framework. (max 10 pt.)	____ out of 10
	Organization Deduct up to 5 points if the paper does not have a clear introduction, body, or conclusion. (max 5 pt.)	____ out of 5
	Spelling and punctuation Deduct 1 point for every spelling and punctuation error in the paper. (max 5 pt.)	____ out of 5
	Grammar Deduct 1 point for every grammar mistake in the paper. (max 5 pt.)	____ out of 5
General Sampling Design ***20 points***	*Target population* Deduct up to 7 points if the target population is not clearly identified. (max 7 pt.)	____ out of 7
	Type of sampling Deduct 4 points if the type of sampling design is not specified and up to an additional 3 points if it cannot be deduced. (max 7 pt.)	____ out of 7
	Sample size If sample size is below 40, deduct 1 point for every missing subject. (max 6 pt.)	____ out of 6
Probability Sampling, ONLY! ***50 points***	*Type of Probabilistic Sampling* Deduct 4 points if the type of probabilistic sampling was not identified and up to another 4 points if it cannot be deduced. (max 8 pt.)	____out of 8
	Target Population List/ Deduct 4 points if the target population list is not included (e.g., copy of student directory, name of dorm/s and room numbers) and up to an additional 3 points if the mechanism for constructing the list is not specified. (max 7 pt.)	____ out of 7
	Method of Random Selection Deduct 10 points if the method of randomly selecting the participants is not identified (e.g., random generating table or software). (max 10 pt.)	____ out of 10
	Actual Sample Deduct 15 points if the list of observations selected from the target population is not provided. (max 15 pt.)	____ out of 15
	Non-response Rate Deduct 1 point for every number less than 50 (required sample size). (max 10 pt.)	____ out of 10
Non-Probability Sampling, ONLY! ***50 points***	*Type of Non-Probabilistic Sampling* Deduct 5 points if the type of non-probabilistic sampling was not identified and up to another 5 points if it cannot be deduced. (max 10 pt.)	____ out of 10
	Selection Bias Deduct up to 20 points for not identifying and discussing <u>all</u> the steps you feel should be taken to reduce selection bias. (max. 20 pt.)	____out of 20
	Representative Sample Deduct up to 20 points for not identifying and discussing <u>all</u> the steps you feel should be taken to ensure the sample is representative of the target population. (max 20 pt.)	____out of 20

Total Score: 100 – Total Deductions = _____

Assessment Rubric for Group Project 4

Dimension	Evaluation criteria	Deductions
General Format Requirement ***5 points***	1 inch margin all around (1 pt.)	____ out of 1
	Double space (1 pt.)	____ out of 1
	Font: Times New Roman (1 pt.)	____ out of 1
	Font size: 12 (1 pt.)	____ out of 1
	Deduct 1 point if paper was ½ a page shorter than 5 pages or longer than 7 pages.	____ out of 1
Quality of writing ***18 points***	*Writing style* Deduct 1 point for every poorly written paragraph (i.e., difficult to read, writing does not flow, or lacked a logical framework) in the body of the paper. (max 3 pt.)	____ out of 3
	Organization Deduct up to 4 points if the paper did not follow the outline provided in the project description. Deduct 2 points if the reference list was not included. (max 6 pt.)	____ out of 6
	Spelling and punctuation Deduct 1 point for every spelling and punctuation error in the paper. (max 3 pt.)	____ out of 3
	Grammar Deduct up to 4 pt. for grammar and 2 pt. if the past tense was not used. (max 6 pt.)	____ out of 6
Abstract ***6 points***	*Presence* Deduct 5 points if the abstract was missing. (max 5 pt.)	____ out of 5
	Length Deduct 1 point if the abstract was < 100 words or >150 words. (max 1 pt.)	____ out of 1
Literature Review ***20 points***	*Citation* Deduct 1 point for every incorrectly cited (Chicago) or missing citation. (max 5 pt.)	____ out of 5
	20 articles Deduct 1 point for every missing article. (max 5 pt.)	____out of 5
	Peer-reviewed journal Deduct 1 point for every article not from an academic journal. (max 5 pt.)	____out of 5
	Adequate presentation Deduct up to 5 pt. if the articles were not integrated into topic. Deduct 5 pt. if articles were summarized individually. Deduct 2 pt. if the lit. review was >2 pages.(max 5 pt)	____ out of 5
Methods ***20 points***	*Instrument* Deduct 4 points if no description of the survey was provided. Deduct 2 more points if a copy of the final version of the survey was not included. (max 6 pt.)	____ out of 6
	Procedure Deduct 2 pt. if the type of sampling design was not specified. Deduct up to 4 more pt. if information regarding the survey administration was not provided. (max. 6 pt.)	____ out of 6
	Participants Deduct 3 points if the target population was not specifically identified and up to an additional 5 points if the sample population was not described. (max 8 points)	____ out of 8
Results ***18 points***	*Sample size* Deduct 5 points if the sample size was not reported. (max 5 pt.)	____ out of 5
	Non-response rate Deduct 3 points if the non-response rate was not included. (max. 3 pt.)	____ out of 3
	Frequency response table Deduct up to 5 points if the table was incomplete (i.e., missing questions) and 5 points if the table was missing. (max 10 pt.)	____out of 10
Conclusion ***13 points***	Deduct up to 10 points if the results presented were not interpreted. Deduct 3 points if lessons learned or suggestions for future studies were not included. (max 13 pt.)	____ out of 13

Total Score: 100 – Total Deductions = _____