Who Makes the Grade?

Research Comparing Self, Peer and Instructor Grades in College

Kayla Atkinson[1], Carmen E. Sanchez[1], Alison C. Koenka[1], Hannah Moshontz[1] & Harris Cooper[1]

[1] Department of Psychology and Neuroscience, Duke University, Durham, USA

Correspondence: Harris Cooper, Department of Psychology and Neuroscience, Duke University, Durham, North Carolina, USA.

Received: May 30, 2018 Accepted: June 19, 2018 Online Published: July 17, 2018

doi:10.5430/irhe.v3n3p1 URL: https://doi.org/10.5430/irhe.v3n3p1

This research was supported by a grant from the WT Grant Foundation. The authors wish to thank Emily Richardson, Grace Hopkins, and Arielle Kahn for their assistance with this project.

Abstract

With the increase in large college classes and online education, student grading of their own work and that of peers is also increasing in frequency. This meta-analysis of 36 studies and 103 effect sizes examined several questions regarding the relationship between grades assigned by college students (either to themselves or peers) and those assigned by their instructors on the same assessment. On average, students graded themselves .41 standard deviations higher than their instructors. The grade distribution correlation between the two types of graders averaged r=.71. Inter-judge reliability estimates suggested that a range of 2-4 peer-graders are needed in order to attain the same level of reliability achieved by the instructor. Little research was found on the effect of student grading on subsequent student performance. Moderator analyses revealed that differences between graders appeared to be minimized when (a) students are grading a peer's work rather then their own, (b) they are in their freshman versus sophomore, junior or senior year, (c) some form of training is given, (d) assessment has lower stakes, (e) more objective tests are given, and (f) course content is English, social science, or professional versus science or engineering. These results have implications for what contexts best facilitate the use of students as graders, and bring to light areas where future research is needed.

Keywords: student grading, self grading, peer grading, meta-analysis, online education

1. Introduction

Student grading involves students judging how well they (or other students) have completed an assignment and giving a grade based on this judgment. The focus of this paper will be on college course grading (Note 1) and will include classes both large and small in size that use either in-class or online grading. (Note 2) The results should be informative for grading that takes place in other contexts as well, especially online courses taken for college credit.

The debate surrounding both self- and peer-grading (SPG) and its role in the classroom is anything but new. Although the learning benefits of SPG have been discussed for years, there are still numerous questions regarding how much of a role students should play in assessments of their work and that of others (Dunning, Heath, & Suls, 2005). Central to these issues is how student grades of their own work might differ from those of instructors.

In this meta-analysis, we examine the existing research in order to better understand how grades given by college student graders compare to grades given by instructors in the college classroom. First, we examine the mean grades given by self- and peer-graders and how this compares with mean grades given by instructors on the same assessment. We also explore the degree of similarity of (correlation between) a college student's position in a distribution when the grades are assigned by students and by instructors. These two research questions represent complementary but independent aspects of how instructor and student grading might differ. We examine not only these overall questions but also how individual differences and different grading contexts might influence their answers. In addition, we explore how inter-judge reliability compares across student and instructor grading. Finally, although the empirical research is limited, we examine the impact of student grading on the students' subsequent test performance.

1.1 SPG in the Current Context of Higher Education (Note 3)

In 2008, College Board Advocacy issued a report titled “Report of the Commission on Access, Admissions and Success in Higher Education.” While the report focused primarily on institutional policies, it also highlighted some instructional changes needed in the objectives meant to improve completion rates among studenets in postsecondary classrooms. Among these recommendations was to use empirical data to identify best practices. Four years later, the American Association of Community Colleges (2012) proposed an action plan, related to what has come to be called “the completion agenda.” The plan included that institutions should “redesign curriculum and instruction to reflect contemporary pedagogical practices” (p. 13) and that the emerging questions involved how to best engage students in learning, especially students taking online courses (p. 19).

1.2 Conceptual Grounding of SPG

In traditional views of education, whether in college or primary and secondary schools, students played a passive role in their own evaluation: taking tests and waiting (often for long periods of time) for feedback, with little or no involvement in the assessment process. However, in the last half-century, educators have come to view students as active learners (e.g., Piaget, 1976). Dochy, Segers, and Sluijsmans (1999) captured well this change in approach: “Research has shown that the nature of assessment tasks influences the approaches which students adopt to learning. Traditional assessment approaches can have effects contrary to those desired” (p. 333). One tangible way of implementing these new ideas was through the use of self- and peer-assessment in the classroom.

1.3 The Increasing Need for SPG

Further, as the demand for higher education continues to increase, self- and peer-assessment may be implemented in the college classrooms out of necessity. For example, more people than ever are continuing education beyond high school (United States Department of Education, National Center for Education Statistics, 2015) and college class sizes are increasing (NBC News, 2007). Additionally, there has been a dramatic increase in both the use of and enrollment in online classes (Allen & Seaman, 2011; Marklein, 2012). Some of these classes enroll thousands of students across numerous colleges and universities and, assuming the course is taken for credit, all students must be graded. These changes create new challenges for educators. Utilizing students as graders may provide a solution that allows instructors to manage a large volume of students in a timely manner, while at the same time allowing variations in assessment types rather than forcing instructors into only multiple-choice options.

1.4 Definition

Many definitions for SPG can be found, (e.g. Falchikov, 1995; Fallows & Chandramohan, 2001; Topping, 1998) but the one we have adopted is similar to that stated by Boud (1991). Student grading involves students judging how well they (and others) have completed a set assignment, and giving a grade based on this judgment. Often, but not always, grading criteria or rubrics are used to guide this process. This involvement requires students to take a more active role in their own learning, moving some of the responsibility away from the instructor (Dochy et al. 1999). With this shift in focus comes the disentanglement of the traditional hierarchy seen within higher education, and the emergence of a new dynamic relationship between the instructor and students (Fallows & Chandramohan, 2001).

It is important to note that student grading gives rise to the question of grading “accuracy” as it relates to the correspondence between student and instructor grades. We use the terms “mean difference,” “correspondence,” and “correlation” when we make our comparisons. While instructor grades are most certainly the performance assessments of record, the issue of how accurate (and valid) the instructor and/or student grades may be is an issue for another review.

1.5 Positive and Negative Effects of SPG

The benefits of student involvement in assessment have been largely accepted (Pearce, Mulder, & Baik, 2009). For example, some argue that traditional grading might diminish learning by placing the focus on the instructor’s particular expectations rather than acquired knowledge (Beckwith, 1991). Involving students in assessment has been argued to shift the focus to students and allows them to learn from their mistakes (Fallows & Chandramohan, 2001). In terms of formative assessment, the utility of self- and peer-assessment has been largely uncontested, with clear learning benefits observed (Topping, 2003). Students may reap several metacognitive benefits from involvement in assessment. Both self- and peer-grading involves the students in the entire learning process, requiring a deeper approach to learning that extends beyond the simple accumulation of facts. The assessment process, from development to application of grading criteria, is a cognitively demanding task that may serve as a reinforcement of material previously learned and may strengthen understanding by presenting material to student graders in a different context (Topping, 1998). With the new role as a self- or peer-grader, students not only receive instruction in the classroom, but also revisit the material in every step of the grading process. Other possible benefits include enhancing the sense of ownership that students feel over their schoolwork, augmenting their level of personal responsibility, increasing motivation levels, and changing their emotional state, such as a newfound empathy for their peers (Topping, 1998).

Finally, student graders acquire skills that are transferable to many other contexts (Topping, 1998). In general, students who serve as assessors may be able to enter the work force with enhanced communication skills, increased ability to negotiate, and a deeper understanding of how to both give and accept criticism from peers (Falchikov & Boud, 1989).

Instructors also benefit from allowing students to engage in assessment activities. First, there are potential logistical advantages to sharing some of the grading responsibility with students. With less time spent marking tests, instructors may be freed to devote more attention to other activities, such as lesson planning and engaging students (Sadler & Good, 2006). In addition, each student grader will likely have more time to grade and provide thorough and useful feedback than a single instructor who has the responsibility of assessing a large number of students.

This shared responsibility also allows for more timely feedback. In the traditional college classroom, a delay often exists between turning in an assignment and receiving a grade. This delay is worrisome for instructors because when students are left unaware of their errors, they begin to consolidate these errors and apply them to subsequent work. With the use of student graders, students receive more timely feedback (Sadler & Good, 2006).

Another important benefit that instructors may experience is an improvement in their relationship with students due to the demystification of the grading process (Sadler & Good, 2006). Students often feel they put forth effort on an assignment only to send it to the abyss, where it is arbitrarily and subjectively graded. Conversely, when SPG is used, instructors are encouraged to provide explicit and detailed guidelines for grading. Expectations for performance become more transparent (Edwards, 2007). This experience may result in a more favorable relationship between instructors and their students.

While the potential positive effects abound, possible negative effects exist as well, both for the student and the instructor. Some worry that including students in the assessment process is more a function of cost and time necessity but is not always in the best interest of the students (Topping, 2003). Involving students as graders may promote competition between classmates, which may encourage the development of maladaptive academic motivation profiles (Ames, 1992; Meece, Anderman, & Anderman, 2006). Peer grading raises questions of confidentiality (Sadler & Good, 2006), and some students may find this practice to be unfair. Generally, students may have less confidence in their classmates than in instructors to assign a grade that is a true reflection of their level of learning. Self-grading can also be imprecise; some argue that it is impossible for students to grade themselves without bias (Dunning, Heath, & Suls, 2004). Bias can be present in peer-assessment, with grading potentially influenced by relationships, friendship or otherwise, with classmates (Topping, 2009). Others argue that having peers assess each other can cause learners to face social discomfort. Because of this, it is often observed that with peer-assessment there tends to be a central tendency, meaning that the majority of peers rate each other's performance as falling near average (Topping, 2005).

Furthermore, instructors may have difficulty accepting student grades as accurate. They may fear that students inflate grades because they possess lower standards for grading themselves and give more weight to effort in the determination of the final grade (Strong, Davis, & Hawks, 2004). They may also fear that social loafing will occur, and some students will participate more fully than others, exacerbating accuracy concerns (Topping, 2009). Additionally, although instructors may assume that using students as graders will save them time, this assumption may not be accurate. In fact, the use of SPG in the classroom may actually be more time consuming than traditional grading (Topping, 1998). Time must be spent establishing grading criteria as well as training students, which in turn may reduce instruction time for other material.

Taken together, there are mixed opinions about the use of SPG as a source of summative assessment in the classroom. While there are numerous potential positive benefits of student grading, with the possibility of enhanced student learning paramount among them, clear negatives also exist. These relate primarily to the validity of student-assigned grades. This uncertainty makes further research of SPG in higher education and a synthesis of the existent research both worthwhile and necessary.

2. Summary of Past Research Synthesis

This is not the first research synthesis to be performed on SPG in higher education. Twenty-nine years ago, Falchikov and Boud (1989) conducted a meta-analysis on student self-assessment in the college classroom. Their synthesis included 57 studies that compared grades assigned by the self with those assigned by instructors. The average standardized mean difference was *d* = 0.47, suggesting that on average students graded almost one-half standard deviation higher than instructors on the same assessment. Furthermore, the average correlation between student and instructor grades was *r* = 0.39. This early synthesis identified several variables that may predict the degree of difference between self and instructor assessment. Falchikov and Boud (1989) found that while the specific year in school of the college student did not predict how similar to instructors they could grade themselves, enrollment in more advanced classes did predict higher grading consistency between students and instructors. This finding suggests that the increased knowledge of the subject matter that is presumed to come with more advanced courses enhances one's ability to self-assess. Also, this synthesis found that those studies identified as being "higher quality" (which included both research design and test condition variables) tended to produce results with better agreement between self- and instructor-graders.

Nearly a decade later, Topping (1998) narratively reviewed 42 studies on peer-assessment in higher education. Topping found that in general, peers were more reliable graders compared to students who graded their own work. Furthermore, this review highlighted the importance of grading criteria and the benefits of including students in the development of grading rubrics.

In 2000, a meta-analysis was conducted on peer grading in higher education, summarizing 48 studies (Falchikov & Goldfinch, 2000). The overall effect size was *d* = .24, suggesting peers gave higher grades than instructors. Compared with the results found in Falchikov and Boud's (1989) earlier meta-analysis on self-assessment, students seemed to assign grades that were more similar to instructor grades when they were grading their peers rather than themselves. Similarly, a higher correlation (*r* = 0.69) was found between peers and instructors than was observed in the earlier meta-analysis of self-graders. Also, peers appeared to give grades more similar to instructors when they were involved in developing the grading criteria. Unlike the synthesis conducted on self-grading, these results did not find the level of course difficulty to have a significant effect on the similarity of peer and instructor grades. Both meta-analyses found that those studies that were of "higher quality" contributed to better grading agreement between student and teacher.

2.1 Factors That May Influence Effects of Grading

In the literature, numerous variables have been offered as possible influences in student grading. Of course, one of the most prominent influences pertains to whether the student is marking their own paper or that of a classmate. The present meta-analysis contributes to this body of literature by being the first to compare and contrast self- and peer-grading, both in terms of the differences in average grades given and in the similarity of grade distributions.

Second, an influence that may be crucial in understanding SPG is the presence and structure of grading criteria. Rubrics are implemented to ensure objective grading and reduce bias during scoring and feedback (McMillan, 2012). While previous research has suggested that rubrics enhance the correspondence between student grades and instructor grades (Falchikov, 1986; Stefani, 1994), we were interested in examining whether the presence, specificity, and training for use of rubrics leads to more comparable grades between students and instructors. Thus, we examined through moderator analyses several variables involving grading criteria.

Third, similarities and differences in students' and instructors' grades are thought to vary with the subject of study. In past research, greater correspondence between peer and instructor grades has been found in science and engineering classes than in the social sciences and humanities (Falchikov & Boud, 1989). Yet, in Falchikov's (2000) meta-analysis on peer-assessment, the subject matter was not found to affect peer-instructor differences. We re-investigate this issue by looking at subject matter as a moderator.

In the past, literature on first-year college students involved in SPG has been scarce relative to other years in college. This gap may exist in part because first-year students are thought to possess underdeveloped skills in self-reflection and a general lack of knowledge about grading in college (Nulty, 2011). Past research has shown that students taking higher-level courses are more accurate graders (Falchikov & Boud, 1989). We hope to test this hypothesis by conducting a moderator analyses to systematically examine how a students' year in school might moderate their ability to grade similarly to instructors.

2.2 The Current Research Synthesis

Part of the rationale for a new meta-analysis pertains to the need to update current knowledge regarding student grading: the last syntheses of SPG were conducted decades ago. Past reports have found high levels of agreement between student and instructor graders (Fry, 1990; Rushton, Ramsey, & Rada, 1993) while others found results that are less positive about the differences in peer assessment (Orsmond & Reiling, 1996; Zoller & Ben-Chaim, 1997). Given these conclusions that appear inconsistent, an update is particularly important. Notably, the current synthesis examines self-grading and peer-grading together in one meta-analysis, allowing for a direct and systematic comparison between these graders. Furthermore, the classification of both self and peer graders as collectively, “student graders” allows the power needed for an examination of moderator variables that might affect student grader outcomes as a whole.

3. Method (Note 4)

3.1 Literature Search Procedures

We used six different strategies to search the literature. Each strategy is described below.

Electronic databases. We searched two electronic reference databases: ERIC (Educational Resources Information Center) and PsycINFO. The databases were accessed through the EBSCO search software on April 28, 2013 and searches were not restricted by date or year.

The term *grades (scholastic)* was used to search titles and abstracts in intersection with the following subject (SU) terms: “evaluation methods,” “evaluation criteria,” “test methods,” “measurement technique,” “peer evaluation,” “self evaluation,” and “multiple choice tests.” After the initial search, a second search was performed in the same databases using the same subject terms combined with the term *grading (educational)*. Searches were conducted sequentially, with overlapping documents excluded from the yield of each subsequent search.

Three searchers independently judged whether each report contained empirical data on one or more of the research questions. In total, 1,153 abstracts were examined and we obtained the full text for the 158 documents that were judged to contain empirical research on *college grading strategies,* along with numerous other documents that we used as background information. Using the inclusion criteria explained below, 10 reports met all necessary criteria to qualify for this meta-analysis. Our searches were not restricted by language.

Backward search. We examined the reference lists of all reports that met the inclusion criteria. After the reference database search, 21 additional qualifying reports were retrieved.

Forward search. We conducted citation analyses on three documents that were frequently cited in relevant reports: one review article (Topping, 1998) and two meta-analyses that focused on SPG (Falchikov & Boud, 1989; Falchikov & Goldfinch, 2000). We found 7 additional qualifying reports through this method.

Direct contact with researchers. We employed several direct contact strategies to tap sources that might have access to additional SPG research. The documents we received this way were already in our database.

Serendipity. We received documents from colleagues who were conducting searches on different but related research topics. One document retrieved in this way was a unique contribution that was included in the meta-analysis (DeGrez, Valcke, & Roozen, 2012).

Direct contact with authors. Several reports found through the previously mentioned search strategies presented data in an unusable way (i.e. the data provided were not amendable for computing an effect size). Unfortunately, the authors were unable to provide us with the data needed to include these reports.

3.2 Criteria for Including Studies

A study had to meet several criteria to be included in our meta-analysis. First, we coded reports of studies that focused on the comparison between instructor grades and student grades in college classrooms using a measure of achievement as the outcome. These comparisons could be either instructor versus peer-graders or instructor versus self-graders. In order to be included in the statistical integration, studies needed to have employed a within-tests design, meaning that the instructor and student must have both graded the same test or other assessment task completed by the students. Studies that assigned some participants to receive student grading and others to receive expert grading (so that the same tests were not marked by both) were not included in the meta-analysis, but were examined separately. In addition, we coded studies that correlated student and instructor grades on the same test or other performance by the student. Finally, we retained but did not code studies that examined the inter-rater reliability of grades assigned by instructors, the self and peers. These studies were analyzed separately.

We employed four additional screens. First, given the earlier meta-analyses, we restricted inclusion to reports appearing between 1990 and April 2014. Second, we required that participants be exclusively undergraduate students. Third, studies had to answer at least one of the two main research questions (mean difference or correlation) in order to be fully coded. Multiple reports were coded for both questions. Lastly, we eliminated reports that did not provide enough information to calculate a standardized mean difference. (Note 5)

3.3 Information Retrieved From Studies

Numerous characteristics of each study were retrieved during the coding process. These characteristics encompassed five broad distinctions among studies: (a) *report characteristics* included basic information about authorship and date of report appearance; (b) *study characteristics* included information about the setting and cultural context, details about the classroom environment, and specifics of the experimental design; (c) *sample information* detailed the demographic characteristics of the different samples that were coded; (d) *grader comparison* information included general grading instructions, rubric specificity and any form of training that occurred before students took part in the grading task; (e) lastly, specifics about the type of assessment, as well as any data needed for effect size calculation were recorded in the *outcome measures*. As is true in most meta-analyses, many of the variables we coded were either not reported often enough or occurred with too little variability across studies to be examined through moderator analyses.

3.4 Effect Size Estimation

We used the standardized mean difference, or *d*-index, to estimate grader differences in assigned grades (Cohen, 1988). The *d*-indexes were computed using Wilson (2018, http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-SMD-main.php), which subtracts the mean instructor grade from the students' mean grade and divides this difference by their weighted standard deviation. Positive *d*-indexes indicated that when students grade either their own work or that of a peer, they assign, on average, higher grades than instructors do when grading the same outcome measure.

For those reports contributing correlations between students and instructors, correlation coefficients were coded exactly as reported, with larger and positive correlations indicating higher levels of agreement between graders with regard to test-takers positions in the distribution of grades.

3.5 Coder Reliability

Each report was coded independently by the first and second authors. All independently coded variables were examined side-by-side for discrepancies, and the two coders discussed any disagreement until a consensus was reached. If disagreements were still unresolved, the principal investigator was consulted. Because all studies were independently coded twice and disagreements were resolved by a third independent coder, the effective reliability of codes is very high (Rosenthal, 1987) and an estimate of reliability (which would involve two new coders and an independent disagreement resolver) is not called for (Appelbaum et al., 2018).

3.6 Methods of Data Integration

Identification of statistical outliers. First, we examined the distribution of effect sizes, separately for *d*-indexes and *r*-values, to determine if any were statistical outliers. The "maximum normed residual test" was applied (Grubbs, 1950, also see Barnett & Lewis, 1984), using $p < .05$, two-tailed, as the significance level. Outlier values were set at the value of their next nearest neighbor. This same procedure was applied to the distribution of sample sizes.

Publication bias. We used Duval and Tweedie's (2000a, 2000b) Trim-and-Fill procedure to test whether the distribution of effect sizes used in the analyses was consistent with the variation of effect sizes that would be expected if the data were normally distributed. The Trim-and-Fill procedure imputes these missing values, thus permitting an estimate of the impact this possible data censoring has on the observed distribution of effect sizes.

Average effect size estimates. We used a weighting procedure to calculate average effect sizes across all comparisons (Borenstein, Hedges, Higgins, & Rothstein, 2006-18). In addition, 95% confident intervals (CIs) were calculated for each average effect. If zero was not included within the interval, the null hypothesis was rejected and it could be concluded that student-assigned grades (either self or peer) were significantly different from instructor grades.

Identifying independent hypothesis tests. One problem that arises in calculating effect sizes involves deciding what constitutes an independent estimate of effect. To address this problem, we used a shifting unit of analysis approach (Cooper, 2018). The shifting unit of analysis approach retains as much data as possible from each study while at the same time minimizing any violations of the assumption that all data points should be treated as independent. Additionally, effect sizes are assigned a weight based on sample sizes, so those studies with multiple samples but few participants are not given disproportionate weight to the precision of their estimates.

Tests for moderators of effects. Possible moderators of effect sizes were tested using homogeneity analyses (Cooper, Hedges, & Valentine, 2009). For each moderator analysis, a *Q*-statistic was calculated to test whether the difference between effect sizes was greater than would be expected by sampling error. Results were calculated using a random-effects models of error. The random-effects model allows for variability in population effect size, and thus greater ability for generalization to other populations. Because SPG is an educational intervention and it is expected that the magnitude of effect will vary within different contexts, we only report results obtained from the random effects model (Borenstein et al. 2006-18).

Each moderator was tested individually, against the full remaining error term, rather than in a meta-regression. Thus, along with the use of a shifting unit of analysis and random error model, this decision represents a conservative approach to data analysis. Meta-regression would have required us to use the individual effect sizes as though they were independent (Hedges, Tipton & Johnson, 2010).

Definitions of moderator variables. Most of the definitions of the moderators we tested should be self-evident (e.g., self versus peer grader, students' year in college) but a few warrant further explanation:

Geographical context. "English-Speaking" countries were defined as reports from studies conducted in the United States, Canada, Great Britain, Ireland, Australia, and New Zealand. "Asian" countries included those studies conducted in China, Japan, Singapore, and Taiwan.

Course content. Science and engineering classes were compared with a category labeled "Other." The Other category included courses in the social sciences, English, and professional subjects.

Classroom versus online grading. Not all studies used traditional methods of pen and paper for grading. Some asked students to do the grading via computer and online. These two modes of grading implementation were compared to one another.

Rubric specificity. While most of the assessments we found contained grading rubrics given to students, they varied along several dimensions. First, we labeled rubrics as either general or specific. "General rubric" provides some criteria, but ultimately gives the grader a high level of autonomy in grading the quality of the work in front of them. This type of rubric can be found in the study conducted by Stefani (1994). A "specific rubric" clearly details the criteria that should be used to judge the quality of performance. An example that guides students on how to assign both partial and full credit, can be found in Freeman and Parks (2010).

Examples or practice with grading. Students were labeled as having exposure to examples or practice prior to grading if they were provided with either (a) a sample assessment along with a rubric used to assign a grade or (b) a completed assessment and rubric form, from which they were able to see how grades had been assigned.

Discussion of criteria. In many cases, instructors spoke with students about SPG and how to best implement it in their classroom before any grading occurred. These discussions came in various forms, including tutorials and less structured dialogue about rubric criteria.

Rubric training index. The four rubric elements described above were combined post-hoc in order to create a "rubric training index." Each element was weighted equally. The rubrics described in reports were rated on a scale of 0-4, depending on how much training students received. No points were awarded for missing elements.

Test weight. This analysis looked at how much weight the graded test would be given to the student's overall grade in the class. We dichotomized this variable into studies in which the student-graded work was worth less than or equal to 15% of the overall grade and those studies in which it was worth more the 15% of the overall grade.

Software. All statistical analyses were conducted using the Comprehensive Meta-Analysis (CMA) statistical software package (Borenstein et al. 2006-18).

4. Results

The literature search located 36 reports with usable data that compared student grading to instructor grading (References to these reports can be found in Supplemental Files C and D, available from the authors.) In total, these reports provided 103 effect sizes. More specifically, our meta-analysis included reports that answered at least one of two main research questions: "How do mean college grades assigned by students compare to grades assigned by instructors on the same outcome measure?" or "What is the degree of similarity (correlation) of students' positions on a grade distribution when grades are assigned by students and by instructors?" Analyses were conducted separately for each research question. A list of all reports that were included in our analyses for both research questions can be obtained from the authors. Often, reports contributed effect sizes for both questions and thus their coded information was included in both sets of data. Those reports that contained only comparisons between self-graders and peer-graders, without an instructor comparison, were not included in either data set, but were analyzed separately.

The literature search uncovered two studies that compared the performance of college students on tests subsequent to the test(s) that involved student graders. Also, we found 12 studies that reported some form of inter-judge agreement among students and/or instructors. Inter-judge reliability statistics are reported separately.

4.1 Differences in Mean Grades

Table 1 summarizes the overall findings that examined mean differences between students and instructors in the grades they assigned. There were 31 reports that contained 81 effect sizes from 44 independent samples. Independent sample effect sizes can be found in Figure 1. Sample sizes ranged from 11 to 3,588 but after one statistical outlier, *n* = 3,588, was identified and adjusted, the sample sizes ranged from 11 to 232. Only one effect size among the 81 was identified as a statistical outlier (*d* = -2.53) and this value was adjusted to that of its next nearest neighbor (*d* = -1.46). Of the 81 effect sizes, 63 were in the positive direction and 18 were negative.

Using the independent sample as the unit of analysis (*n* = 41 after averaging within samples), effect sizes that contributed to the weighted overall mean difference ranged from *d* = -1.46 to 1.4. The average weighted *d*-index was 0.41 and the 95% confidence interval (based on a random effects error model) was *d* = .25 to .56 (Note 6). These results suggest that, on average, college students assigned grades about four-tenths of a standard deviation higher than the grades that instructors assigned, when grading the same assessment. The test for heterogeneity of effect sizes was significant, $Q(44) = 363.03$, $p < .001$, which means that the variability in *d*-indexes was greater than that which would be expected due to sampling error alone. The *tau* for this analysis was 0.46 and the I^2 was 88.2%, suggesting a large portion of the variance in effect sizes was due to heterogeneity among studies.

The Trim-and-Fill analysis, used to test for publication bias, found no evidence of missing effect sizes that were less than the overall mean and six missing effect sizes greater than the mean. Including the estimated missing values in the adjusted weighted overall mean difference raised the overall *d*-index estimate to 0.53.

Table 1. Comparisons of student and instructor mean grades

# of studies contributing comparisons: 31		
# of independent samples: 44		
# of comparisons (effect sizes): 81		
Range of sample sizes: 11-3588		
	Outliers: 3588	
		moved to next nearest neighbor: 232
Range of *d*-indexes using the outcome as the unit of analysis: -2.53 – 1.31		
	# Positives: 63	
	# Negatives: 18	
	Outliers: -2.53	
		moved to next nearest neighbor: -1.46
Range of *d*-indexes using the independent sample as the unit of analysis: -1.46 – 1.4		
	# Positives: 33	
	# Negatives: 11	
Weighted average *d*-index using the independent sample as unit: 0.41		
	CI 95% (random effects model)	
		High: 0.56
		Low: 0.25
	Tau: 0.46	
	I-squared: 88.16	

Study Name	Subgroup Within Study	ES	p-value
Bouzidi A (2009)	Combined	0.064	0.709
Bouzidi B (2009)	Combined	0.062	0.678
Cheng A (1999)	Combined	0.161	0.581
Cheng B (1999)	Combined	-0.267	0.362
Cheng C (1999)	Combined	-1.272	0.000
De Grez A (2012)	Combined	0.655	0.001
Falchikov A (1995)	16	0.380	0.497
Freeman A (2010)	Combined	0.629	0.000
Freeman B (2010)	Combined	0.727	0.000
Freeman C (1995)	22	-0.650	0.192
Freeman D (1995)	23	1.270	0.055
Freeman E (1995)	24	-0.300	0.621
Fry A (1990)	25	-0.290	0.632
Harris A (2011)	26	0.200	0.528
Harris B (2011)	27	0.210	0.580
Hughes A (1993)	28	-0.430	0.159
Karnilowicz A (2012)	29	0.210	0.402
Kriby A (2007)	30	0.910	0.000
Kirby B (2007)	31	1.220	0.000
Kwan A (1995)	32	-0.110	0.590
Leach A (2012)	33	-0.010	0.956
Lew A (2010)	Combined	0.932	0.000
Liang A (2010)	42	1.120	0.000
Liu A (2002)	Combined	0.156	0.003
Magin A (2001)	55	1.000	0.000

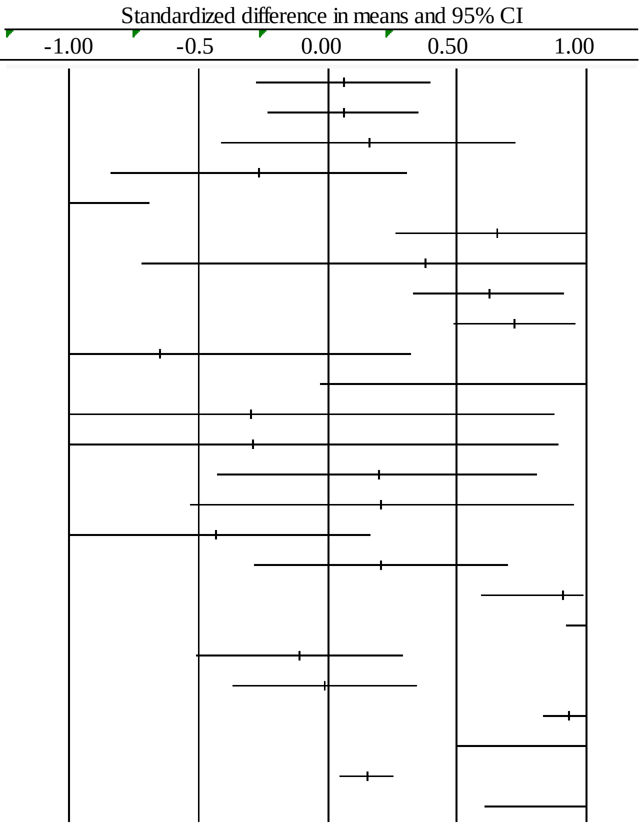


Figure 1. Effect size for mean difference between students and instructors for each independent sample included in the meta-analysis

Note. The CMA program was used for our analyses. In this program, each study is required to have a number, seen in the figures under "Subgroup Within Study." These numbers are completely arbitrary, having no impact on the analyses. When the value is combined, it means that the study contributed multiple effect sizes to our analyses, as multiple outcome measures were present. Because we conducted our analyses at the independent sample level, effect sizes were often statistically combined by the program so that each independent sample contributes only one effect size to the overall analyses.

4.2 Substantive Moderator Analyses of Mean Difference

We conducted analyses exploring substantive moderators of the effects of SPG on achievement grades. Table 2 presents the results of all the moderator analyses (the number of studies contributing to each effect size estimation differ because of differences in reporting). Below, we describe only those comparisons that reached or approached statistical significance.

Self versus peer-grader. The average weighted d-index was 0.54 for the self versus instructor comparisons and 0.36 for peer vs. instructor comparisons. This moderator analysis was not significant, $Q(1)$ 1.18, $p = 0.28$. Effect sizes from studies that directly compared self-graders to peer-graders revealed d-indexes ranged from -0.68 to 1.05. The weighted average d-index for this comparison was 0.37. The 95% confidence interval was $d = -0.05$ to 0.79, $p = 0.08$. This analysis suggested an equivocal finding in the direction that self-graders assigned higher grades than peer-graders.

Geographical context. The weighted average d-index was 0.52 for "English-Speaking" countries and 0.16 for countries from Asia. The difference between average d-indexes was close to significance, $Q(1) = 3.65$; $p = 0.056$, suggesting that students taking classes in countries from Asia may grade more similarly to their instructors than students studying in English-Speaking countries. (Note 7)

Students' year in college. The weighted average d-indexes were 0.12 for freshman students and 0.53 for all other years in college. This analysis revealed a significant positive effect ($Q(1) = 3.71$; $p = 0.05$), indicating freshman students graded more similar to their instructors than students who had been in college more years. (Note 8)

Course content. The weighted average *d*-indexes were 0.64 for science and engineering classes and 0.17 for other subjects. This analysis revealed a significant positive effect, *Q*(1) = 8.67, *p* = 0.003, suggesting that those students in science or engineering classes graded significantly higher compared to their instructors than students taking classes in other subjects.

Student participation in rubric development. The weighted average *d*-index for those students who were involved in the development of the rubric was *d* = 0.15. The *d*-index was 0.48 for those students not involved in rubric development. This analysis produced a equivocal non-significant effect, *Q*(1) = 3.53, *p* = 0.06, but one that suggested the involvement of students in rubric development might facilitate more similar grading between the student and instructor.

Rubric training index. The weighted average *d*-indexes were 0.57 for students receiving no form of training and 0.33 for those with any training at all. This analysis produced an effect that was in the predicted direction but not significantly so, *Q*(1) = 2.11, *p* = 0.15.

Test weight. This analysis produced a significant effect, *Q*(1) = 3.99, *p* = 0.046, suggesting that students graded more similar to instructors when assessments contributed less than 15% towards the students' overall grades.

4.3 Distribution Similarity

Table 3 summarizes the findings examining the correlation between student and instructor grades. There were a total of 28 reports that contained 62 correlations from 37 independent samples. Independent sample effect sizes can be found in Figure 2. Sample sizes ranged from 16 to 3588, but after two statistical outliers, from two independent samples, were identified and adjusted the sample size range was 16 to 230. No correlations were identified as outliers.

Table 2. Moderator analyses for mean grades and correlations between student and instructor grades

Moderator Variables			Difference in Mean Grades			Q_b	Correlations Between Grades			Q_b
			k	*d*	95% CI		*k*	*r*	95% CI	
Grader						1.18				0.03
	Self vs. Instructor		15	0.54	0.26/0.83		15	0.67	0.37/0.84	
	Peer vs. Instructor		39	0.36	0.20/0.53		28	0.69	0.59/0.77	
Context										
	Country[a]					3.65				1.04
		English-Speaking	28	0.56	0.37/0.74		21	0.69	0.57/0.78	
		Asia	11	0.16	-0.21/0.52		9	0.54	0.19/0.77	
	Student Year in College					3.71**				0.04
		Freshman	14	0.12	-0.25/0.48		11	0.73	0.47/0.88	
		Soph,Junior,Senior	21	0.53	0.31/0.76		20	0.71	0.60/0.79	
	Course Content					8.67***				5.32**
		Science/Engineering	19	0.64	0.41/0.88		14	0.80	0.70/0.87	
		Other	24	0.17	-0.04/-0.38		22	0.63	0.50/0.73	
	Classroom vs. Online					0.85				0.46
		Classroom	36	0.43	0.26/0.61		29	0.68	0.57/0.77	
		Online	8	0.29	0.05/0.54		8	0.74	0.57/0.85	

Grading Rubric								
Rubric Specificity				0.76				0.03
General	12	0.51	0.27/0.75		12	0.69	0.47/0.83	
Specific	25	0.36	0.12/0.61		22	0.67	0.57/0.76	
Participation in Development[b]				3.53				
Yes	9	0.15	-0.16/0.45		NA	NA	NA	
No	34	0.48	0.31/0.66					
Examples/Practice				0.17				0.05
Yes	22	0.45	0.22/0.67		20	0.68	0.57/0.77	
No	21	0.38	0.14/0.62		17	0.71	0.53/0.82	
Discussion				0.50				0.58
Yes	18	0.35	0.08/0.62		22	0.72	0.61/0.80	
No	27	0.47	0.28/0.65		15	0.65	0.49/0.78	
Rubric Training Index				2.11				2.28
0	11	0.57	0.32/0.83		9	0.56	0.30/0.73	
1,2,3,4	34	0.33	0.12/0.53		29	0.72	0.62/0.79	
Outcome Characteristics								
Test Weight				3.99**				0.35
≤15%	9	0.14	-0.16/0.44		8	0.59	0.31/0.78	
>15%	18	0.55	0.28/0.82		14	0.67	0.51/0.79	
Test Type				0.50				4.04**
Essay/Presentation	31	0.45	0.25/0.65		25	0.62	0.50/0.72	
Other	11	0.32	0.04/0.61		11	0.84	0.65/0.93	
Test Response				0.51				2.22
Oral	19	0.45	0.16/0.75		13	0.61	0.54/0.68	
Written	28	0.32	0.14/0.51		27	0.72	0.59/0.81	

Notes.

[a] "English-speaking" includes courses at colleges in United States, Canada, Great Britain, Australia, New Zealand, South Africa and Ireland. "Asia" includes China, Singapore, and Taiwan.

[b] Analyses for correlations based on student participation in rubric development could not be run because only 3 independent samples contributed correlations for this variable.

[c] For Test Type, "Other" includes formats involving fill-in-blank, short answer, and those tests involving multiple modes.

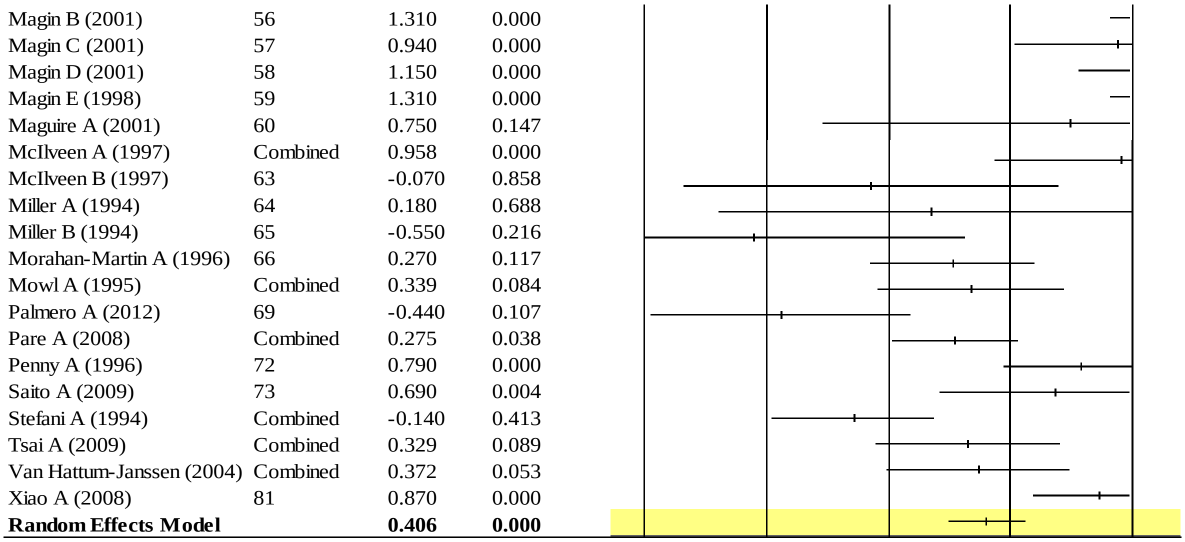
** indicates $p < 0.05$; *** indicates $p < 0.01$

The correlations between SPG and instructor grading ranged between -0.03 and 0.98. Of the 62 r-values only one value was in the negative direction. Using the independent sample as the unit of analysis (and a random effects error model), the average weighted r-value was 0.69 (95% CI = 0.60-0.77) (Note 9). The test for heterogeneity of effect sizes was significant, $Q(37) = 1164.75$, $p <.001$, which means that the variability in correlations was greater than that which would be expected due to sampling error alone

The Trim-and-Fill analysis found no missing effect sizes to the left of the mean correlation and six effect sizes to the right of the mean correlation. An estimate of the adjusted weighted overall mean correlation, including the identified missing values, raised the correlation estimate to 0.78. (Note 10) The *tau*-value using the independent sample as the unit of analysis was .562 and the I^2 was very high, 97%, suggesting that most of the variance in correlations stemmed from heterogeneity across studies.

Table 3. Correlations between student and instructor grades

# of studies contributing correlations: 28
# of correlations (effect sizes): 62
# of independent samples: 37
Range of sample sizes: 16-3588
    Outliers: 3588, 490
        moved to next nearest neighbor: both changed to 230
Range of correlations using outcome values as the unit of analysis: -0.03 to 0.98
    # Positives: 61
    # Negatives: 1
    Outliers: None
Range of correlations using the independent sample as unit of analysis: .10 - 0.98
    # Positives: 37
    # Negatives: 0
    Outliers: None
Weighted average *r*-index using the independent sample as unit: 0.69
    CI 95% (random effects model)
        High: 0.77
        Low: 0.60
    Tau: 0.56
    I-squared: 96.91



Magin B (2001)	56	1.310	0.000
Magin C (2001)	57	0.940	0.000
Magin D (2001)	58	1.150	0.000
Magin E (1998)	59	1.310	0.000
Maguire A (2001)	60	0.750	0.147
McIlveen A (1997)	Combined	0.958	0.000
McIlveen B (1997)	63	-0.070	0.858
Miller A (1994)	64	0.180	0.688
Miller B (1994)	65	-0.550	0.216
Morahan-Martin A (1996)	66	0.270	0.117
Mowl A (1995)	Combined	0.339	0.084
Palmero A (2012)	69	-0.440	0.107
Pare A (2008)	Combined	0.275	0.038
Penny A (1996)	72	0.790	0.000
Saito A (2009)	73	0.690	0.004
Stefani A (1994)	Combined	-0.140	0.413
Tsai A (2009)	Combined	0.329	0.089
Van Hattum-Janssen (2004)	Combined	0.372	0.053
Xiao A (2008)	81	0.870	0.000
Random Effects Model		**0.406**	**0.000**

Figure 2. Effect size for correlation between students and instructors for each independent sample included in the meta-analysis

Note. The CMA program was used for our analyses. In this program, each study is required to have a number, seen in the figures under "Subgroup Within Study." These numbers are completely arbitrary, having no impact on the analyses. When the value is combined, it means that the study contributed multiple effect sizes to our analyses, as multiple outcome measures were present. Because we conducted our analyses at the independent sample level, effect sizes were often statistically combined by the program so that each independent sample contributes only one effect size to the overall analyses.

4.4 Substantive Moderator Analyses of Correlation Differemce

We separately conducted the same moderator analyses for correlations between the student and instructor grade distributions as we did for mean differences in grades. These can be found in Table 2.

Course content. The weighted *r*-values were 0.80 for science and engineering classes and 0.63 for classes in all other subjects. The analyses produced a significant effect, $Q(1) = 5.32$, $p = 0.02$, with a higher correlation and therefore better agreement reported between students and instructors in science or engineering classes.

Test type. The weighted average *r*-values were 0.62 for essay and presentation outcomes and 0.84 for all other types of assessments (fill-in-the-blank, short answer, or multiple modes). The analyses produced a significant effect, $Q(1)$ $=4.04$, $p = 0.04$, suggesting that the correlation between student and instructor grades is lower when the assessment type is an essay or presentation.

No other outcome moderator analyses reached or approached significance.

4.5 Relation Between Mean Differences and Distributional Similarity

For exploratory purposes, we performed an analysis to see whether any relationship existed between our two main research questions. Correlations between student and instructor grades were grouped by the magnitude of difference between student and instructor mean grades in their study. Only those reports that examined both research questions could be used. A median split was conducted in order to group mean differences as either small or large. There were 34 independent samples contributing 54 correlations.

Overall, the weighted average *r*-value was 0.76 ($k = 14$) for independent samples reporting small mean differences (*d*-index of .2 or less) and 0.61 ($k = 20$) for those reporting larger effect sizes. The moderator analysis revealed a non-significant positive effect, $Q(1) = 3.29$, $p = 0.07$. This result does equivocally suggest that students who assigned mean grades that were more similar to the grades given by instructors also tended to reveal distributions of grades more similar to instructors.

4.6 Inter-judge Reliabilities

Table 4 presents the results of studies that examined the inter-judge reliability of peers or instructors. In general, studies comparing reliability measures amongst students and instructors have produced mixed results. For example, some studies have found low agreements between the two graders (Note 11) (Cho, Schunn, & Wilson, 2006), while others have reported a moderate level of agreement between students and teachers (Note 12) (Kovach, Resch, & Verhulst, 2009). Also, different measures of agreement have been used by different researchers. This makes it difficult to compare the estimates of inter-judge reliability both across studies and across multiple peers or instructors. In addition, Magin and colleagues (2001a; 2001b) conducted two studies that included reliabilities for both peers and instructors. The reports noted that according to the Spearman–Brown formula, it would require the averaging of scores of two to four students to attain the level of reliability found for single instructor ratings. This conclusion appears consistent with the general trend gleaned from examining the entries in the table.

4.7 SPG Effects on Subsequent Testing

We found two studies that looked at students' subsequent performance, meaning their performance on tests taken after they had participated in SPG. In these cases, the follow-up tests were all graded by a single grader who was not a student.

In Ozogul and Sullivan (2009), undergraduates in a teacher education program were assigned to a peer, self, or instructor group, which determined who would grade their lesson plans. Later, a post-test was administered to test the students' knowledge of the lesson plans. Results revealed that while generally scores did increase significantly from pre-test scores, there was no significant differences in improvement among the students whose plans were graded by themselves, peers or teachers.

A second study (Khonbi & Sadeghi, 2013) utilized a post-test design, but only compared self- and peer-graders, with no instructor comparison. In this report, peer-graders outperformed those who took part in self-grading on the post-test, suggesting that peer-grading significantly improved learning as compared to self-grading. In this report, learning was measured by tracking improvement on a knowledge test from pre- to post-test, with any gains assumed to be the result of different evaluation types (self- or peer-grading).

Table 4. Inter-judge reliabilities among peers and among instructors grading the same test

Peers			Instructors		
Author	Index of Agreement	Inter-judge Agreement	Author	Index of Agreement	Inter-judge Agreement
Magin (2001b)	r_{11} for average of 4.6 peers[b] with any 1 peer r_{nn} averaged from 4.6 peers[b]	.25 .60	Magin (2001b)	r_{11} for average of 4.8 instructors[b] with any 1 instructor r_{nn} averaged from 4.8 instructors[b]	.49 .82
Marin-Garcia (2008)	r_{11} for average of 43 peers with any 1 peer r_{nn} averaged from 43 peers	.47 .90	Marin-Garcia (2008)	r_{11} for average of 4 instructors with any 1 instructor r_{nn} averaged from 4 instructors	.46 .76
Magin (2001a) Study 1	r_{11} for average of 11 peers with any 1 peer r_{nn} averaged from 11 peers	.38 .84	Baker (1995) Liu (2002)	generalizability (g) analysis, for 4 instructors Correlation of two instructors	.53 .78
Magin (2001a) Study 2	r_{11} for average of 8.1 peers[b] with any 1 peer r_{nn} averaged from 8.1 peers[b]	.29 .75	Lin (2009)	Correlation of two instructors	.67
Cho (2006)	ICC[c] for 3-4 peers ICC[c] for 6 peers	.55 .78	Timmerman (2011)	generalizability (g) analysis, for 1 grader, calculated from 2.5 instructors[b] generalizability (g) analysis, for 3 graders	.66 .85
De Wever (2011) Study 1	ICC for 5 peers	.50			
De Wever (2011) Study 2	ICC for 5 peers	.59			
El-Mowafy (2014)	ICC for 3 peers	.78			
Xiao (2008)	ICC for 3 peers For 20 peers	.62 .75			
Hafner (2003)	ρ^2 for 2 peers	.50			
Kamp (2011)	ρ^2 for 3 peers For 4 peers For 5peers For 6 peers	*.72* *.77* *.81* *.83*			
Sluijsmans (2001) Study 1	$\rho^2$for 6 peers For 13 peers	.81 .86			
Sluijsmans (2001) Study 2	ρ^2 for 6 peers For 14 peers	.86 .92			

Note. The first two rows demark studies in which peer and instructor reliability was measured with the same outcome measure. r_{11} = individual rater reliability, r_{nn} = inter-judge reliability averaged across judges, ICC = intraclass correlation, ρ^2 = generalizability coefficient

[a] Intra-judge reliability is not reported.

[b] The number of peers grading an outcome measures was averaged.

[c] Labeled as "Effective Reliability" within the citation.

5. Discussion

Overall, college students appeared to grade their own papers about four-tenths of a standard deviation higher than the same papers graded by instructors. This estimate is roughly equivalent to those generated from meta-analyses conducted on studies appearing before 1990. It is also consistent with findings in other domains of behavior, such as health and work performance (Dunning, Heath, & Suls, 2004). Students grading their own papers marked themselves about a third of a standard deviation higher than their peers grading the same paper.

As suggested by Dunning et al. (2004) these differences are plausibly attributed to differences in available information. Perhaps students have trouble eliminating perceived effort as a criterion measure, relying on it more heavily than would instructors. They also may not be privy to alternative responses better than the ones they (and their peers) might have provided. Also, as we shall see, it may be that the importance of tests for students might influence their interpretation of the "correctness" of responses.

With regard to the placement of students within grade distributions, students and instructors grades correlated about $r = .70$, suggesting about 50% of the variance in grades was shared by students and instructors. The correlation between student and instructor grades did not vary as a function of whether the student was grading their own paper or that of a peer. Discussion of the grading rubric seemed to produce the largest effect for increasing the distributional similarity between student and instructor grades.

One of our initial research interests involved differences in inter-judge reliability. Our synthesis revealed two important limitations in studies of inter-rater reliability: use of a wide variety of reliability measures and lack of simultaneous reporting of student and instructor reliability. That said, this limited data suggests that averaging two to four student-assigned grades of the same test approached the reliability of a single instructor's grades.

An additional question that we sought to answer involved the impact of prior use of self- and peer-grading on subsequent tests. Only two reports looked at this issue in college classrooms, with varied results. Thus, it appears that this research topic remains largely unstudied in the college population (but less so in earlier grades, see Sanchez, Atkinson, Koenka, Moshontz, & Cooper, 2017). We suspect this inattention is due to the difficulty of conducting such studies, given that they require (a) equating of multiple sections of the same college course or (b) randomly assigning students to different conditions within a course section. More research attention should be devoted to exploring not only the characteristics of grades given by the self and peer, but also how this experience may change subsequent learning.

Several variables were found to moderate the impact of SPG, including year in school, course content, test weight, and test type. More advanced students (sophomores or later), science or engineering classes, and higher-stakes assessments all were associated with more positive grading by students, when compared to their instructors. Additionally, when assessments were essays or presentations, there seemed to be a weaker relationship between student and instructor graders.

Other moderator analyses were not statistically significant, but revealed findings that suggest more research could lead to interesting results. Students from Asian countries seemed to have closer agreement with instructor grades than students from countries that are typically English-speaking. Results that approached significance suggested that the addition of some form of training (rubric specificity, discussion, participation in development, examples/practice) prior to SPG might increase the similarity between student and instructor mean grades. Below, we discuss several of these moderators in greater detail.

5.1 Moderator Variables

Year in school. Freshman appeared to assign grades more consistent with their instructors than did other undergraduates. This conclusion is surprising, with past literature suggesting that first-year students may lack the knowledge and experience needed to grade (Nulty, 2011). The reality, however, seems to be the opposite. Perhaps freshmen, being new to college, are more apt to follow rules and thus be less biased in assigning grades. In addition, students in more advanced classes are more likely to have an increased awareness of how much their grades matter for their future, and thus may (even subconsciously) be positively biased in giving themselves and their peers the benefit of the doubt.

While this finding is unique and important, the freshmen samples include students from various contexts. An assumption has been made that freshman students are at similar levels, both in education and life experience. It is possible, however, that different studies have defined "freshmen" in different ways. For example, freshmen may be defined as a student's in their first year of college, by the number of credits earned, or by the difficulty level of the course, that is, a freshman course might be introductory but enroll students who have been in college for varying numbers of years. If this is true, than our analyses may have lumped together subsets of students that future research should keep distinct.

Course content. Students seem more likely to overestimate (relative to instructors) their grades in science and engineering classes. Perhaps students have a harder time grading when partial-credit is involved, something that is more likely in science-type classes (e.g., organic chemistry). In addition, science and engineering classes possess a unique dynamic between the student and instructor. In a pre-professional course, students may exhibit bias towards themselves, needing to achieve a certain grade in order to maintain a high GPA and be a competitive applicant for admission to professional schools. Instructors of such classes may be faced with a large group of students, many of whom will not ultimately achieve the next level of education.

Interestingly, student graders in science and engineering courses were found to have a higher correlation with instructor than did those students from other courses. Students from these classes may over-grade, but they also appear to be most capable of discerning the relative quality of work in a manner similar to their instructors.

Test weight. Not surprisingly, in high-stakes situations there was a larger discrepancy between student and instructor mean grades. It seems that in lower-stakes situations, students operate with less bias and are able to grade in a manner more similar to their instructors.

Test type. Tests involving a response that allows more subjectivity in grading (e.g., essays, presentations) appeared to produce less distributional correspondence than other forms of responses (e.g., fill-in-the-blank, short answer). One possible explanation is the subjective nature of these assessments. When grading essays or presentations, grading becomes more complicated and the criteria more ambiguous. One way to increase this correlation might be to use specific and detailed essay grading criteria in an effort to guide students in the grading. Providing students with examples and practice essays prior to actual grading may also help.

Grader training. Although not significant, there was some indication that rubric specificity and examples or practice, as well as other training elements (participation in rubric development, class discussions) could reduce the difference between student and instructor grades and grade distributions, though this evidence was far less impressive than we expected. A rubric may be implemented to reduce bias during scoring and therefore is a variable that is potentially crucial in our understanding of self and peer-assessment. Because of the low number of studies that did not provide any rubric to students, rubric variability used across the studies was limited in the meta-analysis. This lack of statistical power may have been a contributing factor to the nonsignificant results. However, there was some indication that involving the student in the development of the rubric (as opposed to having the instructor develop them alone) produced more similar mean grades. Future research should be done to examine experimentally the effects of training. This approach might increase the students' understanding of the criteria for grading, especially in the sense of creating consensus among students and instructors. It might also provide students with a fuller explanation and more time to study the tested material.

Geographical context. Students from countries in Asia (China, Taiwan, Singapore) were better at producing grades similar to their instructors than students studying in countries that are traditionally English-speaking. In fact, the students in the former category did not assign grades that were significantly different from their instructors. This finding may be the result of cultural differences that exist between these groups of students. Future studies that include more than one culture while holding other variations constant (e.g., subject matter, student level) could be informative about the effect culture has on SPG in the college classroom.

6. Conclusions, Limitations, Implications, and Future Research

The findings of these meta-analyses lead us to believe that SPG may have a valuable place in the college classroom, given the several potential gains to be made with its use. However, SPG should be used only in certain contexts if better alignment with instructor grades is to be reached. Student graders are most likely to be "successful" at SPG when: (a) they are grading a peer's work rather then heir own, (b) they are early in their college years, (c) some form of training is given, (d) assessment is of low-stakes, (e) more objective tests are given, and (f) course content is not science or engineering. Peer grading can be improved with the use of several peer-graders. If a classroom context fails to meet any one of these criteria, attention to the remaining criteria becomes more important in establishing the trustworthiness of SPG.

6.1 Limitations

As is always the case with meta-analysis, our analysis was limited by the level of completeness of reporting found in the studies we located through our literature search, and there was quite a bit of variability in the level of detail presented in each.

Another major limitation of the meta-analysis was an issue of statistical power. For several of our moderators, cell sizes were small and unequal. In particular, several of the moderators used to create the training index were not close to being equally distributed. The lack of statistical power present for several moderator analyses likely contributed to non-significant results. Still, in several of these analyses, results did point in interesting directions for future research (Borenstein et al., 2006-18).

6.2 Implications

The results of these research syntheses have some important implications for grading strategies and how instructors and colleges might develop policies for their courses, especially in this era of an expanding number of large and online courses that are given for college credit and therefore need some form of grading. First, it is clear that instructors of courses which use absolute grading criteria and student or peer graders will need to consider some form of re-centering of grades and/or adjustment to grade cut-offs to result in assigned grades similar to the grades the instructor might assign. Additionally, for grades marked on a curve (as well as on an absolute scale), the issue of correspondence regarding where students place in the class distribution will be an issue. Instructors will need to decide whether the correlation of about .7 justifies the use of students as graders (read, results in fair grades). Of course, instructors can use several techniques, including averaging multiple peer-assigned grades, giving more tests of lesser weight, perhaps more detailed scoring rubrics and intensive training of students, to strengthen the correspondence between the grades they would assign and those given by students. At the institutional level, training could be provided for instructors to learn best practices for using SPG and even some prescriptions regarding when it should and should not be used in courses giving credit toward graduation.

6.3 Future Research

While this meta-analysis shows that much research already exists regarding SPG and its usage on the college campus, more work needs to be done. Specifically, primary research needs to be devoted to further explore and experimentally manipulate several of the moderators we have identified in this report. For example, though our analyses did not produce significant results between self- and peer-graders, the direct comparisons suggested that important differences may exist. This effect needs more direct comparisons both for greater power and increased ability to identify what variables might moderate the difference. The same is critical for grader training studies.

Also, because of sample size issues and the fact that our initial self-versus-peer moderator analysis was not significant, we did not have a strong argument for analyzing these two types of student grader groups separately. When a sufficient database has been amassed, however, these two groups should be compared more closely, especially to see if moderator variables affect them differently.

Finally, there is a paucity of qualitative research that explores the meaning and implications for self- and peer-grading as experienced by the students themselves. Such research could provide important insights into SPG and how it is interpreted and influences those who take part in it.

Critically important, we have identified a substantial gap in the college literature. Few studies have examined subsequent learning that takes place as a result of prior SPG. The supposed learning benefits of both self- and peer-grading remain one of the strongest arguments for its use. However, this claim needs to be supported by more experimental data.

Similar to all educational practices, SPG can have both positive and negative implications for students and the validity of academic evaluation, depending on the context and manner of its use. This meta-analysis has identified what the current empirical base might be for “best practice” for SPG and what future research is needed.

References

Allen, I. E., & Seaman, J. (2011). *Going the distance: Online education in the United States, 2011*. Retrieved from http://files.eric.ed.gov/fulltext/ED529948.pdf

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*, 261-271. https://doi.org/10.1037/0022-0663.84.3.261

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*, 3-25. https://doi.org/10.1037/amp0000191

Barnett, V., & Lewis, T. (1984). *Outliers in Statistical Data (*2nd ed.). New York, NY: Wiley.

Beckwith, J. B. (1991). Approaches to learning, their context and relationship to assessment performance. *Higher Education, 22*(1), 17-30.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2006-18). *Comprehensive meta-analysis* version 3. Retrieved rom https://www.meta-analysis.com/?gclid=Cj0KCQjw37fZBRD3ARIsAJihSr1PXSDH5EwNkg1mrmIl8RZmQJyebXk2d9xhZoANW5VPVVHQDn9CfXYaAmIvEALw_wcB

Boud, D. (1991). *Implementing student self-assessment*. Higher Education Research and Development Society of Australasia (HERDSA).

Cho, K., Schunn, C., & Wilson, R. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*, 891-901. https://doi.org/10.1037/0022-0663.98.4.891

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. J. (2014). The MOOC phenomenon: Who takes massive open online courses and why? Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2350964

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

College Board Advocaacy. (2008). *Coming to our senses: Education and the American Future*. New York, NY: The College Board.

Cooper, H. (2018). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Thousand Oaks, CA: Sage.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

DeGrez, L., Valcke, M., & Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education*, *13*, 129-142. https://doi.org/10.1177/1469787412441284

Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24*, 331-350. https://doi.org/10.1080/03075079912331379935

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*, 69-106. http://dx.doi.org/10.1111/j.1529-1006.2004.00018.x

Duval, S., & Tweedie, R. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association*, *95*(449), 89-98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 276-284. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Edwards, N. M. (2007). Student self-grading in social statistics. *College Teaching*, *55*, 72-76. https://doi.org/10.3200/CTCH.55.2.72-76

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education*, *11,* 146-166. https://doi.org/10.1080/0260293860110206

Falchikov, N. (1995). Peer feedback marking: developing peer assessment. *Programmed Learning*, *32,* 175-187. https://doi.org/10.1080/1355800950320212

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59,* 395-430. https://doi.org/10.3102/00346543059004395

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70,* 287-322. https://doi.org/10.3102/00346543070003287

Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education*, *6,* 229-246. https://doi.org/10.1080/13562510120045212

Freeman, S., & Parks, J. W. (2010). How accurate is peer grading?. *CBE-Life Sciences Education*, *9*, 482-488. https://doi.org/10.1187/cbe.10-03-0017

Fry, S. A. (1990). Implementation and evaluation of per marking in higher education. *Assessment and Evaluation in Higher Education*, *15*, 177-189. https://doi.org/10.1080/0260293900150301

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, 27-58. https://doi.org/10.1214/aoms/1177729885

Hedges, L.V., Tipton, E. & Johnson, M.C. (2010). Robust variance estimation in meta-regression with dependent effect sizes. *Research Synthesis Methods, 1*, 39-65. https://doi.org/10.1002/jrsm.5

Khonbi, Z. A., & Sadeghi, K. (2013). The effect of assessment type (self vs. peer) on Iranian university EFL students' course achievement. *Procedia-Social and Behavioral Sciences*, *70,* 1552-1564. https://doi.org/10.1016/j.sbspro.2013.01.223

Kolowich, S. (2012). Who takes MOOCs. *Inside Higher Ed, 5*.

Kovach, A. R., Resch, S. R., & Verhulst, J. S. (2009). Peer assessment of professionalism: A five-year experience in medical clerkship. *Society of General Internal Medicine, 24,* 742-746. https://doi.org/10.1007/s11606-009-0961-5

Li, H.L., Xiong, Y., Zang, X.J., Kornbaher, M., Lyu, Y.S., Chung, K.S., & Suen, H.K. (2014, April). *Peer assessment in a digital age: A meta-analysis comparing peer and teacher ratings.* Presentation at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Lipsey, M.W., & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181-1209. https://doi.org/10.1037/0003-066X.48.12.1181

MacPhail, C.J. (2011). *The completion agenda; A call to action.* Washington, DC: American Association of Community Colleges.

Magin, D. J. (2001a). A novel technique for comparing the reliability and multiple peer assessments with that of single teacher assessments of group process work. *Assessment & Evaluation in Higher Education, 26,* 139-152. https://doi.org/10.1080/02602930020018971

Magin, D. J., & Helmore, P. (2001b). Peer and teacher assessments of oral presentation skills: how reliable are they?. *Studies in Higher Education*, *26*, 287-298. https://doi.org/10.1080/03075070120076264

Marklein, M.B. (2012). Online-education trend expands. *USA Today*. Retrieved from: http://www.usatoday.com/story/news/nation/2012/11/18/more-on-board-with-online-education-trend-of-moocs/1713079/

McMillan, J.H. (Ed.). (2012). *SAGE handbook of research on classroom assessment*. Thousand Oaks, CA: Sage Publications.

Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation, and academic achievement. *Annual Review Psychology*, *57*, 487-503. https://doi.org/10.1146/annurev.psych.56.091103.070258

NBC News. (2007). Monstrous class sizes unavoidable at colleges: Nobel prize-winning prof calls for reform, says huge classes cause damage. Retrieved from http://www.nbcnews.com/id/21951104/ns/us_news-education/t/monstrous-class-sizes-

Nulty, D. D. (2011). Peer and self‐assessment in the first year of university. *Assessment & Evaluation in Higher Education*, *36*, 493-507. https://doi.org/10.1080/02602930903540983

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education*, *21,* 239-249. https://doi.org/10.1080/0260293960210304

Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. *Educational Technology Research and Development*, *57,* 393-410. https://doi.org/10.1007/s11423-007-9052-7

Pearce, J., Mulder, R., & Baik, C. (2009). Involving students in peer review. Case studies and practical strategies for university teaching. Centre for the Study of Higher Education, University of Melbourne.

Piaget, J. (1976). *Piaget's theory* (pp. 11-23). Berlin, Germany: Springer.

Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis*. Cambridge University Press.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: A case study. *Journal of Computer-based Instruction*, *20*, 75-80.

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment*, *11*, 1-31. https://doi.org/10.1207/s15326977ea1101_1

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-Grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, 1049-1066. https://doi.org/10.1037/edu0000190

Stefani, L. A. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education*, *19*, 69-75. https://doi.org/10.1080/03075079412331382153

Strong, B., Davis, M., & Hawks, V. (2004). Self-grading in large general education classes: A case study. *College Teaching*, *52*, 52-57. https://doi.org/10.3200/CTCH.52.2.52-57

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, *68*, 249-276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. S. R. Segers, F. J. R. C. Dochy, & E. C. Cascallar (Eds.), *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55-87). Dordrecht, The Netherlands: Kluwer Academic. https://doi.org/10.1007/0-306-48125-1_4

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology*, *25*, 631-645. https://doi.org/10.1080/01443410500345172

Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48*, 20-27. https://doi.org/10.1080/00405840802577569

United States Department of Education, National Center for Education Statistics. (2015). Institutional retention and graduation rates for undergraduate students. Retrieved from http://nces.ed.gov/programs/coe/indicator_cva.asp

Zoller, Z., & Ben-Chaim, D. (1997, August). *Student self-assessment in HOGS science examination; is it compatible with that of teachers?*. Paper presented at the meeting of theEuropean Association for Research on Learning and Instruction, Athens, Greece.

Notes

Note 1. A separate research synthesis (Sanchez, et al., 2017) looks at similar research questions but for the 4-12 grade levels.

Note 2. We do not include Massive Open Online Courses (MOOCs). These deserve their own treatment (see Li et al. 2014). MOOCs allow unlimited participation and most people who take them are retired or do so for job advancement (Christensen et al. 2014; Kolowich, 2012) rather than working toward a postsecondary degree.

Note 3. A Supplemental File A is available from the authors describing the history of SPG in the context of higher education.

Note 4. A Method section that contains a complete description of the search strategy is available from the authors.

Note 5. A Supplemental File B presents a fuller description of the search procedures. Supplemental Files C and D contain all references included in Figures 1 and 2. Supplemental File E contains studies closely related but excluded from the synthesis for the stated reason. All supplement files are available from the authors.

Note 6. The analysis was also conducted without adjustment for statistical outliers. The average weighted *d*-index was 0.42 and the 95% confidence interval was $d = 0.26$ to 0.58. The test for heterogeneity was still significant ($Q(44) = 547.92$, $p < .001$).

Note 7. Nine effect sizes from seven independent samples contained within five reports were excluded from this analysis because they came from countries that did not fall in either of these groups. They included studies conducted in Spain ($d = -0.44$), Algeria ($d = -0.19$ and 0.33), Belgium ($d = 0.47$ and 1.24), Portugal ($d = 0.35$ and 0.39), and South Africa ($d = 0.91$ and 1.22).

Note 8. The year-in-school analysis was also conducted for freshmen and sophomores versus juniors and seniors. This analysis produced differences between mean *d*-indexes that were not significant, $Q(1) = 1.36$, $p = 0.24$, but there was better agreement between those in earlier years (freshman and sophomore) and instructors.

Note 9. The analysis was also conducted without adjustment for statistical outliers. The average weighted *r*-value was 0.60 and the 95% confidence interval was $r = 0.58$ to 0.79. The test for heterogeneity of effect sizes was still significant, $Q(37) = 2837.36$, $p < 0.00$.

Note 10. We examined whether the mean difference and/or correlation between student and instructor grades varied based on year of the report. A significant difference between means was found for those studies disseminated before 2000 (d = 0.10), and those disseminated after 2000 (d = 0.58), $Q(1)$ = 7.03, p = 0.008. Students in studies conducted in earlier years had greater similarity with instructor grades than those in later years.

Note 11. Cho et al. (2006) includes both undergraduate and graduate students. They give six estimates for undergraduates in a figure but not the precise values. However, it appears the values are consistent with the values in the main table and reliability seems to be peaking at a higher number of raters (n = 6). It seems there might be significant gains for having six rather than 2-4 raters.

Note 12. Kovach et al. (2009) has not been included in our analyses because students were in medical school.