Applied Research on English Language

V. 9 N. 4 2020
pp: 477-502
http://jare.ui.ac.ir/are

DOI: 10.22108/are.2020.120578.1516

The Relationship between Graded and Tested Achievement: Do Gender and Proficiency Level Make a Difference?

Majid Nowruzi [1], Majid Amerian [2*], Hooshang Yazdani [3], Alimohammad Mohammadi [4]

[1] *PhD Candidate, Department of English Language and Literature, Arak University, Arak, Iran*
[2] *Associate Professor, Department of English Language and Literature, Arak University, Arak, Iran*
[3] *Assistant Professor, Department of English Language and Literature, Arak University, Arak, Iran*
[4] *Assistant Professor, Department of English Language and Literature, Arak University, Arak, Iran*

Received: 2019/12/23 **Accepted: 2020/09/01**

Abstract: Grades represent one of the most common sources of evidence of student achievement in classrooms, though their relationship with test scores has remained understudied, particularly in settings such as in Iran, where English is taught as a foreign language. The purpose of this study was to investigate the relationship between graded and tested achievement with respect to gender and proficiency level differences. Teacher-assigned grades and standardized achievement test scores of 693 Iranian learners of English taught by 15 teachers were examined. Primary analyses focused on the validity of teacher grades and the subsequent Pearson correlation coefficients revealed that grades associated positively with externally-validated test scores obtained from reliable tests, an indication of the validity of teacher grading. Additionally, the results of independent-samples t-tests showed that female students outperformed male students on achievement tests, but with fluctuations across proficiency levels. Higher proficiency levels gave male participants an advantage over female participants in achievement tests. Moreover, male teachers were found to grade female participants more accurately than their female counterparts. Implications are discussed for informing teachers about the validation of their grading practices, as well as for teacher education programs and teachers' professional development.

Keywords: Assessment and Grading Practices, Graded Achievement, Tested Achievement, Grades, Achievement Test Scores, Gender, Proficiency Level, Correlation.

* **Corresponding Author.**

Authors' Email Address:
[1] **Majid Nowruzi** (m-nowruzi@phd.araku.ac.ir), [2] **Majid Amerian** (m-amerian@araku.ac.ir), [3] **Hooshang Yazdani** (h-yazdani@araku.ac.ir), [4] **Alimohammad Mohammadi** (a-mohammadi@araku.ac.ir)
ISSN (Online): 2322-5343, **ISSN (Print):** 2252-0198

Introduction

The 20th century has witnessed substantial reliance on high-stakes tests in order to measure academic achievement for a variety of purposes (McMillan, 2013, 2018). However, in recent decades, researchers have turned their attention to classroom assessment (CA) as a field of inquiry in its own right (McMillan, 2013). Grades, as the output of such assessment, are at the heart of teachers' everyday educational measurement practices, and grading decision-making influences the lives of millions of students around the globe (Brennan, Kim, & Wenz-Gross, 2001). Grades are the key determinants of students' school achievements and performance and are frequently used by various stakeholders to make important decisions about students (Brookhart et al., 2016; Guskey, 2015; Guskey & Link, 2018). Pattison, Grodsky, and Muller (2013) noted that "Grades are the fundamental currency of our educational system; they signal academic achievement and non-cognitive skills to parents, employers, postsecondary gatekeepers, and students themselves" (p. 259). Nevertheless, grades have not been adequately explored as far as the relationship between grades and test scores is concerned. Numerous grading research to date have probed composite report card grades as multidimensional measures of student achievement, effort, and participation (Bowers, 2009, 2011; Klapp Lekholm, 2011; Klapp Lekholm & Cliffordson, 2008, 2009), early studies of (un)reliability of teacher grading (Brimi, 2011; Healy, 1935; Hulten, 1925; Lauterbach, 1928; Silberstein, 1922; Sims, 1933; Starch & Elliott, 1912, 1913a, 1913b), and the (in)validity of grades (McMillan, 2001; McMillan, Myran, & Workman, 2002). This study aims to investigate the relationship between grades and test scores in relation to gender and proficiency level differences.

The commonly-held beliefs that teacher judgments and classroom assessments do not appear to be as accurate measures of academic achievement as standardized testing and that they are subject to frequent error and bias have been circulating among parents, students, and even some professionals, although not explicitly stated (Allen, 2005; Brophy, 1983; Egan & Archer, 1985; Hoge, 1983, 1984; Hoge & Cudmore, 1986; Schwab, Moseley, & Dustin, 2018). In line with such beliefs, Egan and Archer (1985) noted that "It is commonly argued that commercial tests provide teachers with valuable information about the abilities and deficiencies of their students, from which it follows that teachers who rate their students without such information will often be in error" (p. 25). Consequently, investigating the relationship between graded and tested achievement may help practitioners examine the credibility of such assumptions.

Additionally, as it appears that there are no independent and psychometrically-sound

measures of validating teacher grading because such grading practices are often restricted to the boundaries of the classroom and not applicable on a larger scale, the probable invalidity of such grading practices might pose concerns for teachers and other stakeholders (Allen, 2005; Schwab, Moseley, & Dustin, 2018). Disputes among the major stakeholders including parents, teachers, and students are more likely when teachers have no means of supporting the validity of their grading decision-making (McMillan, 2013). Therefore, equipping teachers with expertise to validate their own grading practices could help prevent or settle possible disputes. Consequently, investigating the way grades and test scores that are usually, though not necessarily, externally-validated may give teachers the toolbox they need to primarily validate the grades they assign using standardized tests as a criterion and then either support or refute their grading-related decisions. Such criterion-related validity could be easily estimated by correlating grades and pre-validated test scores and examining the extent of their association.

Moreover, the significance of studying grading boils down to issues such as the ubiquity of grades and popular perceptions of grades as measures of student achievement (Brookhart, 2015), their role as feedback providers to stakeholders (Elliott, Gresham, Freeman, & McCloskey, 1988; Gerber & Semmel, 1984; Hoge, 1983), their impact on student motivation and learning, and their effects on learners' future achievements and their willingness for participation in educational programs (Black & Wiliam, 1998; Crooks, 1988; Natriello, 1987). As a result, grades are important to study for a variety of reasons including a) their power in predicting significant future educational measures (Brookhart, 2015), b) their role as predictors of school dropouts (Bowers, 2010; Bowers & Sprott, 2012; Bowers, Sprott, & Taff, 2013), c) their predictive role in taking entrance exams for colleges (Bowers, 2010), d) their power in predicting admission to college, performance in college, and also graduation from college (Atkinson & Geiser, 2009; Thorsen & Cliffordson, 2012), and e) their influence in highly competitive selection settings (Sawyer, 2013). Despite changes in grades and the introduction of phenomena such as 'grade inflation', grades have still kept their signaling power throughout many years (Pattison et al., 2013).

The chief reason behind the study of gender-based variations in teacher grading pertains to the possibility of finding any probable bias on the part of the teachers when determining student grades. In line with this argument, Marmeleira et al. (2019), in their study of PE teachers' grading, noted that finding any probable disparity in grades of either male participants or female participants should alert researchers and educational authorities to look for reasons for the weaker performance of either sexes or their teachers' biases and to take measures to compensate for such shortcomings. Additionally, studying teacher grading in settings such as Iran where single-gender classes are the

norm in mainstream education and even in private EFL institutes can significantly contribute to the teachers' understanding by helping them more critically and analytically think about their own assessment and grading practices through self-reflection and reconsider such practices if needed. Rauschenberg (2014) states that studying gender, race, and other characteristics relating to student performance is important because they may affect teachers' grading as they also influence the way teachers view students. For example, common stereotypes such as male participants are better at math or female participants do better in literature or courses that require extensive memorization may result in score discrimination, whereby teachers assign grades at least in part on the basis of the stereotypes of students' innate features, rather than their actual performance.

Nevertheless, one of the areas of inquiry that has been explored the least is where graded achievement and tested achievement intersect. As Brookhart (2015) noted, graded achievement consists of the students' achievements of learning goals and classroom instructional objectives and is, therefore, subject-specific. It also highlights individual teachers' grading practices and aligns much better with tested achievement. The focus here is on student-centered achievement, not assessment, which is primarily teacher-oriented, without taking learners into account. Nowadays, more and more experts tend to recognize the need for teacher judgments and assessments of students' achievement of learning goals to be seen as psychological measures in their own right and, therefore, subject to the same psychometric standards that are often applied to other assessment measures such as tests and observations (Edelbrock, 1983; Gerber & Semmel, 1984; Gresham, 1981; Hoge, 1983, 1984; Hoge & Cudmore, 1986). Brookhart (2003) pointed to the lack of 'indigenous' measurement theory that accounts for the specific uses and purposes of classroom assessment without borrowing concepts and measurement tools from the large-scale assessment. Similarly, McMillan (2013, 2018) stated that little emphasis was placed on CA until late in the previous century. He argued that although a myriad of studies has been done on teachers' grading practices in the late 20th and early 21st century, "lack of theoretical grounding" is seriously felt in the majority of these studies to date, a gap that needs to be filled by conducting even more studies on various aspects of CA such as grading validity. This recognition heightens the importance of studying the probable interactions between teacher grades and achievement test scores, though they may not measure even the same underlying constructs (Brookhart, 2015).

Literature Review

Hoge and Butcher (1984) conducted a multiple regression analysis and found that of the variables of student gender, IQ, and student achievement, the last was a stronger predictor of

teachers' classroom judgment, with a regression coefficient of .71. The validity of teacher judgment was also endorsed by Hoge and Coladarci (1989) in their literature review of similar studies, where teachers' estimates of their students' correct responses to test items were 70% accurate on tests like reading and math. Likewise, Leinhardt (1983) pointed to the validity of teacher judgments by studying exact matches or 'hit rates'. He obtained a hit rate of 64% on a reading test. The hit rate referred to the exact match between teachers' judgments of students' correct answers on the test items and students' actual correct answers on those items.

McCandless, Roberts, and Stames (1972) studied the role of variables such as gender, school poverty level, and ethnicity in the relationship between graded and tested achievement for seventh-grade students in 10 schools in subject areas including reading, language, arithmetic, social studies, and science. The results showed that for male participants, for the Caucasian, and for the advantaged students, teachers' grades, IQ, and tested achievement scores did not represent the same underlying constructs. The accuracy of teachers' grades was higher for female participants than for male participants and female participants obtained higher grades than their male counterparts.

Three of the 16 empirical studies reviewed by Hoge and Coladarci (1989) investigated the effect of student gender on the accuracy of teacher judgment against student standardized achievement and found no statistically significant results for this moderator variable (Doherty & Conolly, 1985; Hoge & Butcher, 1984; Sharpley & Edgar, 1986). Dusek and Joseph (1983) asserted that despite the fact that teachers' social-behavioral expectations were different for male participants and female participants, no significant differences were observed in teachers' expectations of male and female academic achievement. This was consistent with other findings in the literature. The resulting correlations between teacher judgment and student achievement for single groups were somehow similar, the median correlation between teacher judgment and criterion was r = 0.64, with a range of 0.28 to 0.86, while the median correlation for within-class correlations was 0.70, ranging from 0.48 to 0.92 (Hoge & Coladarci, 1989).

In another study, Brennan et al. (2001) studied the relationship between grades and achievement test scores of eighth-grade male and female students from different races and ethnicities. The results revealed that teachers' grades were more equitable measures for accountability in measurement than the standardized test scores. The findings of this study validated the way teachers assessed students in class as represented by the grades they assigned. They also investigated the correlation between teacher grades and standardized

test scores as measured by MCAS with respect to student gender and concluded that female participants generally outperformed male participants in both classroom grades and test scores.

Later, Duckworth and Seligman (2006) examined the mediating role of variables such as gender in student performances both inside the classroom and on final examinations. In their study, adolescent female students gained higher grades in their courses compared to their male counterparts, and as far as GPA was concerned, the gender difference was more than twice for female participants than that for male participants. Achievement test scores underestimated female participants' GPAs, while they overestimated male participants' GPAs. They found that male participants performed much better than female participants on IQ tests and the mediating role of IQ scores for female participants' overall GPAs was weaker than that for male participants'.

Guskey (2011) investigated the relationship between teachers' first grades and their final course grades, both on percentage scales, for ninth- to twelfth-grade students in five schools. The other variables of the study included gender, grade level (9-12), student poverty level, and student ethnicity. He concluded that grades were reliable measures of student achievement due to their remarkable stability throughout each term and school year. Small but statistically significant differences were observed for gender, grade level, ethnicity, and poverty level. In other words, female participants earned higher grades and grades increased as students advanced to higher levels. Also, Asians got the highest grades, African American students the lowest, and Caucasians fell somewhere in the middle.

Duckworth, Quinn, and Tsukayama (2012) conducted a study, comparing the GPA and standardized math and reading test scores of 1,364 ninth-graders and 510 eighth-graders using structural equation modeling (SEM) in their research design. The results, however, were nearly identical to those of the earlier studies and revealed that GPAs and standardized math and reading test scores correlated at r = .62 - .66. This finding was later corroborated by Pattison et al.'s (2013) study in which high school GPAs were compared with reading and math test scores and approximately similar results were obtained.

Rauschenberg (2014) studied differential grading where students with similar ability levels studying the same courses earn grades that show inconsistencies on the basis of factors such as teacher grading criteria, student behavior, gender, or teacher stereotypes, rather than their actual performance. Using a three-year database on English and Algebra course grades and test scores and after running multiple-regression statistics, he found that

student characteristics such as gender were stronger predictors than other variables in shaping grades. He also found that female students earned higher grades when factors such as test scores, schools, or districts were held constant. Also, students from low-income families earned lower grades and Black students got lower grades on English than Whites and Asians. He concluded that factors influencing teacher-assigned grades were so strong that they could cause the grades to fluctuate one grade on a 7-point A-F scale. He also observed small inconsistencies with regard to student and teacher race and gender.

Cheng and Sun (2015) studied the grading practices of 350 Chinese English language teachers using descriptive analyses and MANOVAs and found that grade (proficiency) level was one of the factors that influenced their grading decision-making in addition to the class size and teachers' assessment literacy. Their study had general implications for improving teacher grading decision-making in the Chinese context.

In sum, it can be concluded that mediator variables such as student or teacher gender play important parts in the grade-test score relationship and that previous research has been able to generate manifest variability in some aspects and consistency in others. However, the majority of the studies reviewed here and elsewhere were carried out in ESL settings. Studying teachers' grading practices in EFL contexts may help the research community be better equipped to take further steps toward theorizing the field and distancing itself from telling teachers what not to do rather than what to do, as Brookhart (2015) conceded. Concerning the gaps addressed previously, this study aims to seek answers to the following questions:

1. How do teacher grades and achievement test scores relate?
2. Are teacher grades valid measures of student achievement?
3. How do male and female students perform on achievement tests across proficiency levels?
4. How does teacher gender affect the grades assigned to female students in classrooms?

Method

Participants

Table 1 presents the frequency of the sample by gender and proficiency level. The sample consisted of 693 Iranian EFL learners— 310 males (44.7%) and 383 females (55.3%) — whose ages ranged from 14 to 41 years, with a mean age of 20.0 (SD = 2.50) and whose first language was Persian. Female participants outnumbered male participants in all proficiency levels except

in elementary and advanced levels. At the time the data were collected, the learners were all studying English in the Iran Language Institute (ILI). The learners spread across six proficiency levels, as displayed in Table 1. The largest number of EFL learners belonged to the elementary level, with 186 learners (26.8%) of the sample, whereas there were only 16 learners (2.3%) at the advanced level. The rationale behind this variation in the number of learners along proficiency levels may attribute to the increasing difficulty of higher levels and, consequently, the higher failure/dropout rates of learners as they progress toward higher levels.

Table 1. *Frequency of Learners by Gender and Proficiency Level*

Proficiency Level	Gender		Total	
	Male n (%)	Female n (%)	n	%
Basic	76(45.5)	91(54.5)	167	24.1
Elementary	94(50.5)	92(49.5)	186	26.8
Pre-Intermediate	71(39.6)	108(60.4)	179	25.8
Intermediate	43(41.3)	61(58.7)	104	15.0
Upper-Intermediate	18(44.0)	23(56.0)	41	5.9
Advanced	8(50.0)	8(50.0)	16	2.3
Total	310(44.7)	383(55.3)	693	100.0

All class grades were reported to the institute individually by each of the 15 EFL teachers including 7 males (46.6%) and 8 females (53.4%) with an age range of 27 to 43 and a mean age of 31. As the data in Table 2 show, seven teachers (46.7%) had MAs in English literature, four (26.7%) had MAs in applied linguistics or TEFL, two (13.3%) had BAs in English translation, and two teachers (13.3%) had PhDs in non-English majors. Also, most of the teachers, that is nine teachers (60.0%) had between 11 to 20 years of teaching experience, indicating that they were mostly experienced.

Table 2. *Frequency of Teachers by Degree*

Teacher's Qualification	n	%
MA in English literature	7	46.7
MA in applied linguistics or TEFL	4	26.7
BA in Translation	2	13.3

Non-English Major	2	13.3
Total	15	100.0

Note. MA = Master of Arts, BA= Bachelor of Arts

Data Collection and Analysis

The data were originally gathered through Excel spreadsheets containing information about learners' IDs, their gender and proficiency levels, their corresponding teacher gender, and their class grades and achievement test scores. The obtained data consisted of two sets of scores including a) learners' class grades (graded achievement) which were the composite grades assigned by teachers to measure learners' language abilities in each of the four skills of listening, speaking, reading, and writing, and b) learners' achievement test scores (tested achievement) obtained after administering achievement tests that measured the same set of skills and that were made, standardized, and validated nationally by the ILI and adapted to the proficiency levels. The issue of class grades or what is known in the literature as teacher-assigned grades consisted of individual marks assigned to the evidence of learners' performances in the four major skills that were ultimately aggregated and reported as composite percentage grades. The rationale behind comparing these two sets of assessment data pertains to their unique goal of measuring the same underlying construct, i.e., a learner's general English proficiency, a claim similarly endorsed by teachers, students, and educational administrators.

The data were primarily screened for outliers to minimize or eliminate their unfavorable effects on the final statistical outputs. Next, the continuous variables including class grades and achievement test scores were scrutinized in order to investigate the necessary assumptions prior to running statistical operations such as correlations and t-tests. The Kolmogorov-Smirnov tests of normality for both variables were significant at $p < .001$, a result which is unfavorable but is quite common in large samples. The shape of the distribution of scores for both variables was mildly negatively skewed. This, however, should not be considered as the violation of the normality assumption and depends, to a great degree, on the nature of the data. Specifically, the shape of the distribution of final exam scores or class grades where a large number of students compete with each other to earn higher grades is expected to be mildly negatively skewed.

After the primary analyses were conducted to ensure no violations of the normality and linearity assumptions, the data were analyzed in SPSS v. 24 using Pearson product-moment

correlations and independent-samples t-tests to determine the strength and direction of the relationships and the significance of differences between means for various learner groups, respectively. Additionally, the coefficients of determination for correlation coefficients and the effect sizes for t-tests were calculated to examine the amount of the shared variance of the two variables and the magnitude of the observed significant effects, respectively.

The reported Cronbach alpha reliability (α) of the nationally standardized achievement test from which the present assessment data were elicited was α = .87, which was indicative of the high reliability of the test. The content validity of the test was primarily determined by the panel of expert test developers who frequently evaluate the relatedness to and representativeness of test items to the course contents. Additionally, copies of the standardized tests used for data collection in the 2019 winter semester were examined by the researchers to determine the degree of the representativeness of the test items and the suitability of the test contents to measure the constructs under investigation. The researchers who are mostly experienced faculty members unanimously agreed that the degree of correspondence between the content coverage of the test and the teachers' assessment practices in the classroom settings was satisfactory. The tests measured listening comprehension, reading comprehension, writing ability (for intermediate levels and above), vocabulary knowledge, grammar, and fixed spoken expressions through function items.

Both sets of scores were reported on a traditional percentage scale of 0 to 100. The pair of scores for each individual learner along with supplementary data about the learners' gender and proficiency levels and the teachers' gender were obtained, as well. Students' grades are regarded as summary scores of learners' language abilities plus all the other probable non-achievement factors that teachers might consider when grading student work. Non-achievement factors comprise a wide range of criteria such as effort, ability, improvement, and behavior that presumably contaminate the grades assigned by introducing construct-irrelevant variance in grading, contrary to the measurement community's recommendations (Brookhart, 1994; McMillan, 2001).

The final exam scores were one-shot scores obtained from administering standardized multiple-choice format tests of general language proficiency covering areas of listening comprehension, reading comprehension, grammar, and vocabulary knowledge. Both sets of scores were calculated and reported as percentages. The pertinent data were collected from January to March 2019.

Independent-samples t-tests were used to compare the means of class grades and

achievement test scores for male participants and female participants. They were also used to draw comparisons between teachers' grades with respect to the teacher gender and proficiency levels taught. Moreover, correlations were used to inspect the direction and strength of the relationships between grades and achievement test scores of male and female students. The calculated effect sizes (d) were subsequently reported in order to examine the magnitude of any probable statistically significant differences obtained.

Results

In this section, the results of the statistical procedures are presented. Figure 1 illustrates the scatterplot of the correlation between teacher-assigned grades (x-axis) and students' achievement test scores (y-axis). A preliminary inspection of the density of dots and their packed clustering around the fit line and the confidence interval lines (dotted lines) shows that the two variables (sets of scores) are associated positively, though moderately.

As is evident in Figure 1, the 95% confidence interval lines are also illustrated on the scatterplot. These lines are used for inferential statistics purposes. Due to the close proximity of the confidence interval lines to the best fit line (the solid black line), it can be concluded that for 95% of the additional samples that could be obtained from the same population, the best fit line would fall within the dotted lines. Afterward, the Pearson product-moment correlation coefficient was calculated and a medium-strength positive linear correlation $r = .46$, $p < .01$ (two-tailed) was observed. This means that there is a moderate relationship between students' grades and standardized test scores. The fit line direction is positive, meaning that any increase in student grades is associated with increases in achievement test scores. The strength of the relationship was determined on the basis of Cohen's (1988) guidelines for the interpretation of correlation coefficients. The coefficient of determination was calculated as (r^2) = 21.7, meaning that almost 22% of the variance in test scores could be explained by class grades and vice versa.

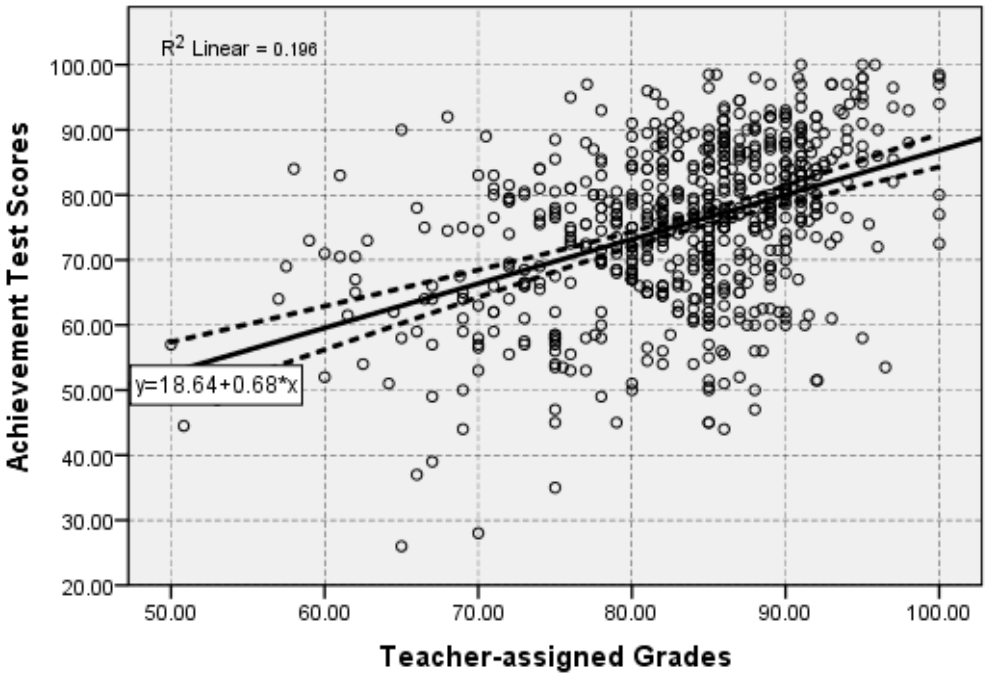

Figure 1. *Scatterplot with Confidence Intervals for the Relationship between Teacher-assigned Grades and Achievement Test Scores.*

Table 3 shows the output of multiple independent-samples t-tests across the proficiency levels. As can be seen, five proficiency levels (basic to upper-intermediate) are identified along with the descriptive statistics for males and females and their corresponding t values. The mean difference values are negative and are diminishing as learners move from basic to pre-intermediate, indicating that female participants scored higher than male participants across these three proficiency levels, but this supremacy tended to subside as learners' proficiency levels progressed. In the basic level, female participants (n = 91, M = 78.48, SD = 14.64) outperformed male participants (n = 76, M = 72.66, SD = 14.56) with a mean difference of 5.82, which was found to be statistically significant where t (165) = -2.56, α = .01. The nearly identical standard deviation values indicate that the scores for male and female students at the basic level were equally dispersed, highlighting the significance of the observed mean difference more strongly. The same is true for the elementary level where female participants (n = 92, M = 77.29, SD = 10.9) scored higher than male participants (n = 93, M = 73.23, SD = 13.47) with a statistically significant mean difference of 4.05 where t (183) = -2.25, α = .02. Although it should be pointed out that achievement test scores of male participants were more widely dispersed here than those of female participants due to the corresponding larger standard deviation values.

When it comes to the pre-intermediate level, the mean difference between male participants' and female participants' achievement test scores is negligible, marginally favoring female participants again (mean difference = -.29) and not statistically significant (α = .86). Male participants (n = 71, M = 73.54, SD = 11.7) and female participants (n = 108, M = 73.84, SD = 11.04) performed equally well as evidenced by the negligible t value of -.17. The same is true for the intermediate level, where the mean difference of .94 is again negligible, but this time positive, giving male participants an edge over female participants. The obtained t (102) equaled .44 and was not statistically significant (α = .65). Further, the approximately similar standard deviation values for male participants and female participants, which are 10.04 and 11.06 respectively, indicate that the scores were equally dispersed, pointing to the fact that male and female learners did equally well here.

However, in an abrupt turn in the trend favoring females over males, male participants (n = 18, M = 82, SD = 10.97) performed better than female participants (n = 23, M = 77.3, SD = 8.87) in the upper-intermediate level where t (39) = 1.51, α = .01. The statistically significant mean difference of 4.69 points to the supremacy of the performance of male participants in comparison to that of female participants in achievement test scores, as presented in Table 3. This finding reversed the trend followed from the first two proficiency levels (Basic and Elementary) where female participants outperformed male participants.

Table 3. *Means, Standard Deviations, and T-test Outputs for the Achievement Test Scores of Male participants and Female participants across Five Proficiency Levels*

Level	Learner gender	n	M	SD	t	df	Mean difference	95% confidence interval		Cohen's d
								lower	upper	
Basic	M	76	72.66	14.56	-2.56*	165	-5.82	-10.3	-1.34	.40
	F	91	78.48	14.64						
Elementary	M	93	73.23	13.47	-2.25*	183	-4.05	-7.61	-.49	.33
	F	92	77.29	10.90						
Pre-Intermediate	M	71	73.54	11.70	-.17	177	-.29	-3.70	3.11	na
	F	108	73.84	11.04						
Intermediate	M	43	76.46	10.04	.44	102	.94	-3.26	5.15	na
	F	61	75.51	11.06						
Upper-Intermediate	M	18	82.00	10.97	1.51*	39	4.69	-1.57	10.96	.47
	F	23	77.30	8.87						

Note. 'M' stands for male and 'F' stands for female. *p < .05.

In line with the fourth research question, female participants' grades assigned by either male or female teachers were examined, as summarized in Table 4. The grades of a subsample of 383 female learners were studied, of whom 80 female participants (20.9%) had male teachers and 303 female participants (79.1%) were taught by female teachers. As is evident in Table 4, the two-class grade means for male and female teachers show variation, with a mean difference of 2.06, which needs to be statistically investigated for significance.

Table 4. *Means and Standard Deviations of the Grades of Female participants Taught by either Male or Female Teachers*

Teacher Gender	n	M	SD	Std. Error Mean
Male	80	85.05	5.44	.608
Female	303	82.99	9.19	.528

The independent-samples t-test was conducted and t (211.4) = 2.55 was statistically significant at p = .012 (two-tailed), indicating that men (n = 80, M = 85.05, SD = 5.44) assigned female learners higher grades than women (n = 303, M = 82.99, SD = 9.19). The magnitude of the observed difference in the means of class grades or the effect size (mean difference = 2.05, 95% CI: .46 to 3.64) was medium (eta squared = 0.30) according to Cohen's (1988) interpretations of effect size values for t-tests (d < .20 = small effect size, d = .20 to .50 = medium effect size, d > .51 = large effect size).

To study female participants' performance with male and female teachers, a bivariate correlation analysis was conducted to see whether the higher grades than female participants earned in men's classes were based on teachers' favoring them over male participants out of bias. The correlation coefficients between class grades and achievement test scores for female participants in classes with male and female teachers are statistically significant at $r_{\text{male teacher}}$ = .68 (n = 80) and $r_{\text{female teacher}}$ = .40 (n = 303), p = .01 (two-tailed), respectively.

Discussion

The first research question was concerned with the relationship between graded and tested achievement. The obtained correlation coefficient, though moderate in strength, shows that teacher-assigned grades were not off assessment track and were positively correlated with standardized test scores. However, caution should be exerted as far as the degree of this

relationship matters. The obtained coefficient approximates in the degree to those reported in other studies (e.g., Duckworth et al., 2012; Helmke & Schrader, 1987; Hoge & Butcher, 1984; Hoge & Coladarci, 1989; Hopkins, George, & Williams, 1985; Pattison et al., 2013; Woodruff & Ziomek, 2004) where the obtained coefficient stood somewhere in the middle of the 0.0 - 1.0 continuum and, more specifically, ranged minimally between .45 to .50, as in Woodruff & Ziomek's (2004) study or maximally between .03 to .90 as in Helmke & Schrader's (1987) study. The lower than expected correlation coefficient in the present study could probably be attributed to not controlling adequately for some confounding variables such as teachers' professional background or teaching experience; as it is reckoned that experienced and novice teachers may not grade students similarly. This speculation, however, runs contrary to the findings of Guskey and Link (2018) who found that "the sources of evidence" that teachers used in determining grades were quite unrelated to their years of teaching experience (p. 313). Regardless of the magnitude of the coefficient, the implication is that grades and test scores are associated. In the absence of distinct psychometric measures for CA, this relatedness offers CA in general and grading, in particular, the gift of validity which will be discussed next.

The second question concerns whether grades can be regarded as valid measures of student achievement. The answer to question one foreshadows this inquiry because it was found that grades and test scores correlated positively. Since standardized tests are most likely expected to have undergone meticulous psychometric scrutiny for their reliability and validity and they are usually, though not necessarily, thought of as reliable and valid measures of achievement, we could conclude that their psychometrically-tested standards are transferable to grades, as the two were found to correlate. This implication is supported by Hoge and Butcher (1984) who argued that student achievement was a stronger predictor of grades. This resulting implication provides a response to the second research question in that grades could be seen as valid measures of achievement especially because the achievement test, used in this study, enjoyed a reliability index of $\alpha = .86$ and was judged to have satisfactory content and construct validity.

It could also be interpreted in the light of criterion-related validity because the achievement test scores played the role of the criterion from which comparisons were drawn. This finding is supported by Hoge and Coladarci's (1989) review where it was pointed out that teacher judgments of students' achievement, reported as grades, were valid measures of such achievement. This finding was further corroborated when Leinhardt (1983) obtained a hit rate of 64% on a rating test and stated that teacher judgments were considered as valid measures of

student achievement. Others have underscored that grades can be thought of as more reliable, more stable, and more equitable measures of student achievement and also for accountability purposes compared to standardized tests (Brennan et al., 2001; Guskey, 2011).

Contrary to the benefits of grades, it was found that they showed something other than only achievement and construct-irrelevant factors such as intelligence were represented by grades as well (Carter, 1952, 1953; McCandless et al., 1972). This finding may pertain to areas where grades and test scores do not point to the same construct(s), as no studies reported absolute or even very strong correlation coefficients between the two variables. However, what matters here is not discrepancies, but commonalities because it is assumed that tests and grades should measure the same constructs overall as they pursue almost the same goals, that is the measurement, either formative or summative in nature.

In order to provide support for the existence of discrepancies between grades and test scores, some underlying factors needed to be taken into account. First, the achievement tests used in the current study were the same for all learners in classes with the same proficiency levels while the teachers' individual differences such as their teaching experience, background, and probably their language assessment literacy were not factored in. This is an important issue because it might cause the obtained correlation coefficient to be underestimated. Consequently, a correlation coefficient of .46 can have a slightly different meaning once such factors are taken into account. As a result, more extensive research is needed to examine if higher correlation coefficients could be obtained if teachers have higher literacy in language assessment or are more experienced.

Concerning the third research question, Carter (1952, 1953) stated that female participants earned higher algebra grades than male participants and McCandless et al. (1972) claimed that not only did female participants generally get higher grades than male participants, but teachers also graded female participants more accurately than male participants. This finding is aligned with the results of Brennan et al.'s (2001) study in which they contended that overall, female participants did better than male participants in both classroom grades and achievement test scores. However, this is only one side of the story because once proficiency levels were taken into account, this general finding happened to be more complicated than it appeared. What Brennan et al.'s (2001) study did not explore was the role of gender and proficiency level in achievement test score variations. Female participants tended to outperform male participants on achievement tests as McCandless et al. (1972) stated in their study, but this was true only in basic and elementary levels, and as learners' mastery of English

enhanced, this trend tended to diminish. In the pre-intermediate level, female participants' mean value on the final exam is only marginally higher than that of male participants and not statistically significant. In intermediate and upper-intermediate levels, it is the male participants who outperform female participants in achievement tests, though these differences are not statistically significant for intermediate level but reach statistical significance for upper-intermediate learners. This variability was also pointed out by Hoge and Butcher (1984) who used gender, IQ, and achievement test scores as predictors in their study. They reported that the accuracy of teacher grades showed variability. The finding that test scores for male participants and female participants vary across proficiency levels seems to be an interesting point in that it may be attributable to the increasing complexity of test items at higher levels which require more intense mental processing. Duckworth and Seligman (2006) stated that male participants performed much better than female participants on IQ tests. However, relating the findings of their study to the better performance of male participants compared to female participants in higher proficiency levels in this study should be treated with more caution.

The pre-intermediate level was the cutting point in this study since the performance of female participants before and after this level contradicted. In other words, in basic and elementary levels, the mean differences favored female participants while after the pre-intermediate level, the trend reversed and male participants had higher achievement test scores means, regardless of whether or not the differences were statistically significant. While basic-level female participants significantly did better than male participants on achievement tests, as also shown in the study by Brennan et al. (2001), in upper-intermediate levels these were the male participants who noticeably outperformed female participants, a piece of evidence inconsistent with what Guskey (2011) reported where female participants got higher grades and kept their distance from male participants as they progressed to higher grade levels. Guskey and Link (2018) showed that grade level played a part in determining students' grades because K-12 teachers at elementary and secondary schools used various grading criteria with varying levels of weight attached to them when determining students' grades. Nevertheless, further research is needed to investigate the effect of proficiency levels on the performance of male participants and female participants in large-scale testing and on teacher grading, while controlling for the possible impacts of confounding variables such as teachers' individual differences.

To study the impact of teacher gender on the accuracy of teachers' grade-giving practices, single-sex classes (only female students) taught by either male or female teachers were of paramount importance because by keeping the student gender variable constant (female participants only), the impact of teacher gender could better be understood. The results of the t-test showed that the observed difference in the means of teacher grades was statistically significant, with a reported medium-strength effect size. This means that male teachers assigned higher grades to female participants than their female counterparts. To restate, female participants did better in classes with male teachers than in classes with the same-sex teachers, provided that the correlations between female participants' class grades and achievement test scores were also higher in classes with male teachers. This condition was met because the correlation coefficient for male teachers was r = .68, and statistically significant at p = 01, while the same coefficient for female teachers was r = .40, p = 01. When the results of t-tests and the correlation statistics are aggregated, it can be concluded that male teachers tend to grade female participants more accurately than do female teachers, a finding similarly endorsed by McCandless et al. (1972). However, Carter (1952, 1953) claimed that these were female teachers who assigned students higher grades than their male counterparts. It should be noted that the teachers in Carter's study most probably taught co-ed classes, while this study, at least in part, examined grading in single-sex education. In addition, the subject matter studied by Carter (1952, 1953) was algebra, not English. This finding that teachers' grading practices vary was also corroborated by Coladarci (1986) who studied variability in grading and achievement tests and found a statistically significant impact for math teachers. In contrast, Guskey (2011) found small but statistically significant differences in variables such as gender. Extensive research is needed to explore the role of teacher gender in the observed variability of students' grades.

Seen from the grading accuracy perspective, one could hypothesize that male teachers grade more accurately than female teachers simply because the correlation between the graded and tested achievement of female participants with male teachers was higher than that with female teachers, a point which requires to be further investigated. However, it would be interesting to study the way female teachers grade male participants compared to what their male counterparts do because the results of such a study could provide more support for the higher accuracy of grading in classes where students and teachers are not the same sex.

Other researchers studying co-ed systems found no statistically significant results for the effect of student gender on the accuracy of teacher judgment (Doherty & Conolly, 1985;

Sharpley & Edgar, 1986). What is noteworthy to mention, as far as these studies are concerned, is that they were carried out in co-ed systems where gender differences are not sensitized. Further research is needed to explore the underlying factors behind these modest differences in grading to see where such discrepancies stem from. For now, what seems reasonable to conclude is that variations exist as far as classroom assessment and grading practices are concerned.

Conclusion

This study was carried out in order to explore the relationship between graded and tested achievement by considering the effects of gender and proficiency levels in an Iranian EFL setting. The most important contribution of this study was to seek validation for teachers' grading practices by finding positive correlations between teachers' grades and students' achievement test scores. This finding showed that teacher grades were not shots in the dark, contrary to the popular belief that the only reliable and valid measures of student achievement are standardized tests. Teacher grades are considered valid evidence of student achievement because they correlate positively but moderately with valid achievement tests. This may strengthen the argument in favor of the validity of teacher judgments and decisions in classrooms.

The roles of gender and proficiency levels, as independent variables, were studied using correlations and t-tests. In line with what was acknowledged in the literature, female participants outperformed male participants in achievement tests. However, when proficiency levels were taken into account, variations in results were observed. Female participants outperformed male participants in achievement tests in lower proficiency levels, but as learners progressed in their proficiency, male participants took over and did significantly better. This finding was endorsed by the relevant literature where it was asserted that female participants did poorer when IQ tests were used (Duckworth & Seligman, 2006). It was also found that male teachers gave female participants more accurate grades than their female counterparts did.

Limitations

Two issues restricted the results of this study. The first one was that grades and test scores were reported as total composite averages on a percentage scale. Therefore, it was not feasible to figure out what percentage of each of the underlying constituents, assessed in classes as

separate skills, contributed to the total grades or test scores. Specifically, each reported grade and test score for each of the learners were aggregates of that learner's performance on distinct language skills. In other words, the student's multi-faceted performance in the class was summarized into one single score that was believed to represent his/her mastery best. This turns into a disadvantage in this study because the concept of averages does not necessarily yield favorable results when variability in performance can generate a more detailed sketch of the realities of learner abilities. Simply put, an average of 70 (on a percentage scale) for a learner who scored 90 in listening and 50 in writing is not going to be as informative as the single scores themselves. This summation of results in favor of reporting averages may probably stand for the moderate correlations obtained in this study. One could safely conclude that associations between sets of scores could be higher if such scores were reported separately. That is if students' grades and test scores could be disintegrated and their constituents could be accounted for a clearer and more comprehensive image of the underlying relationships could be produced.

The second limitation pertains to the restrictions induced by single-sex education and the paucity of situations where female teachers were allowed to grade male students. Consequently, little if any information about female teachers' grading of males was available for analysis. The researchers could not access information concerning female teachers' grading of male learners, which could be informative with respect to examining the role of gender in teacher grading comparatively. The counter-balanced grading in which the likely impacts of gender differences on grades or test scores can be highlighted may be more informative for studying gender-based assessment and grading practices.

Implications and Future Directions

The major implications of this study along with suggestions for further research are discussed in this section. The first and most important implication is that the findings of this study may help EFL teachers have a clearer understanding concerning the validation of their own grading practices in classrooms. The findings are hoped to enhance teachers' realization in comparing grades with externally-validated standardized test scores to the benefit of assigning more valid grades, provided that the two measurement devices are aligned and pursue similar evaluative purposes. This, in turn, may contribute to enhanced grading practices among practicing teachers. In this regard, the researchers suggest that future studies attempt to disintegrate total average grades or scores into their underlying components or subskills and investigate the

relationships between such fine-tuned dimensions of graded and tested achievement. Studying how micro-skills, and not just composite grades, possibly relate can be more enlightening to both teachers and the measurement community by creating lines of dialogs between these two rather than encouraging the traditional producer-consumer relationship.

Additionally, the findings can help inform teachers about their grading practices and make them think more critically when gender differences come into play. The results of this study supported the notion that teachers graded female participants higher than male participants because female participants did better than male participants, but what remains unknown is why female participants and male participants perform differently in classrooms, and why female participants do better. Therefore, it is hereby suggested that the reasons behind the unbalanced performance of either male participants or female participants in graded and tested achievement be explored. The findings of such studies can help tackle teachers' biased grading and stereotyping and provide better opportunities for assessment equity. Such results could also inform teacher education programs about what needs to be done in order to address these discrepancies in a more consistent manner.

References

Allen, J. D. (2005). Grades as Valid Measures of Academic Achievement of Classroom Learning. *The Clearing House, 78*(5), 218-223. DOI:10.3200/TCHS.78.5.218-223

Atkinson, R. C., & Geiser, S. (2009). Reflections on a Century of College Admissions Tests. *Educational Researcher, 38*(9), 665-676. DOI:10.3102/0013189X09351981

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. DOI:10.1080/0969595980050102

Bowers, A. J., Sprott, R., & Taff, S. (2013). Do We Know Who Will Drop Out? A Review of the Predictors of Dropping Out of High School: Precision, Sensitivity, and Specificity. *High School Journal, 96*, 77-100. DOI:10.1353/hsj.2013.0000

Bowers, A. J. (2009). Reconsidering Grades as Data for Decision Making: More Than Just Academic Knowledge. *Journal of Educational Administration, 47*, 609-629. DOI:10.1108/09578230910981080

Bowers, A. J. (2010). Analyzing the Longitudinal K–12 Grading Histories of Entire Cohorts of Students: Grades, Data-Driven Decision Making, Dropping Out, and Hierarchical Cluster Analysis. *Practical Assessment, Research, and Evaluation, 15*(7). DOI:10.7275/r4zq-9c31

Bowers, A. J. (2011). What's in a grade? The Multidimensional Nature of What Teacher Assigned Grades Assess in High School. *Educational Research and Evaluation, 17*, 141-159. DOI:10.1080/13803611.2011.597112

Bowers, A. J., & Sprott, R. (2012). Examining the Multiple Trajectories Associated with Dropping Out of High School: A Growth Mixture Model Analysis. *The Journal of Educational Research, 105*(3), 176-195. DOI:10.1080/00220671.2011.552075

Brennan, R. T., Kim, J. S., & Wenz-Gross, M. (2001). The Relative Equitability of High-Stakes Testing versus Teacher-Assigned Grades: An Analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard Educational Review, 71*(2), 173-216.

Brimi, H. M. (2011). Reliability of Grading High School Work in English. *Practical Assessment, Research, and Evaluation, 16*(17). DOI:10.7275/j531-fz38

Brookhart, S. M. (1994). Teachers' grading: Practice, and Theory. *Applied Measurement in Education, 7*(4), 279-301. DOI:10.1207/s15324818ame0704_2

Brookhart, S. M. (2015). Graded Achievement, tested Achievement, And Validity. *Educational Assessment, 20*(4), 268-296. DOI:10.1080/10627197.2015.1093928

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research, 86*(4), 803-848. DOI:10.3102/0034654316672069

Brophy, J. E. (1983). Research on the Self-fulfilling Prophecy and Teacher Expectations. *Journal of Educational Psychology, 75*(5), 631-661. DOI:10.1037/0022-0663.75.5.631

Carter, R. S. (1952). How Invalid Are Marks Assigned by Teachers? *Journal of Educational Psychology, 43*(4), 218-228. DOI:10.1037/h0061688

Carter, R. S. (1953). Non-intellectual Variables Involved in Teachers' Marks. *Journal of Educational Research, 47*(2), 81-96. DOI:10.1080/00220671.1953.10882084

Cheng, L., & Sun, Y. (2015). Teachers' Grading Decision Making: Multiple Influencing Factors and Methods. *Language Assessment Quarterly, 12*(2), 213-233. DOI:10.1080/15434303.2015.1010726

Cohen, J. W. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coladarci, T. (1986). Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. *Journal of Educational Psychology, 78*(2), 141-146. DOI:10.1037/0022-0663.78.2.141

Crooks, T. J. (1988). The Impact of Classroom Evaluation Practices on Students. *Review of Educational Research, 58*(4), 438-481. DOI:10.3102/00346543058004438

Doherty, J., & Conolly, M. (1985). How Accurately Can Primary School Teachers Predict the Scores of Their Pupils in Standardized Tests of Attainment? A Study of Some Non-cognitive Factors that Influence Specific Judgments. *Educational Studies, 11*(1), 41-60. DOI:10.1080/0305569850110105

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline Gives Female participants the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores. *Journal of Educational Psychology, 98*(1), 198-208. DOI:10.1037/0022-0663.98.1.198

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind Leaves Behind: The Roles of IQ and Self-control in Predicting Standardized Achievement Test Scores and Report Card Grades. *Journal of Educational Psychology, 104*(2), 439-451. DOI:10.1037/a0026280

Dusek, J. B., & Jpseph, G. (1983). The Bases of Teacher Expectancies: A Meta-analysis. *Journal of Educational Psychology, 75*(3), 327-346. DOI:10.1037/0022-0663.75.3.327

Edelbrock, C. (1983). Problems and Issues in using Rating Scales to Assess Child Personality And Psychopathology. *School Psychology Review, 12*(3), 293-299.

Egan, O., & Archer, P. (1985). The Accuracy of Teachers' Ratings of Ability: A Regression Model. *American Educational Research Journal, 22*(1), 25-34. DOI:10.2307/1162985

Elliott, S. N., Gresham, F. M., Freeman, T., & McCloskey, G. (1988). Teacher and Observer Ratings of Children's Social Skills: Validation of the Social Skills Rating Scales. *Journal of Psychoeducational Assessment, 6*(2), 152-161. DOI:10.1177/073428298800600206

Gerber, M. M., & Semmel, M. I. (1984). Teacher as Imperfect Test: Reconceptualizing the Referral Process. *Educational Psychologist, 19*(3), 137-148. DOI:10.1080/00461528409529290

Gresham, F. M. (1981). Social Skills Training with Handicapped Children: A Review. *Review of Educational Research, 51*(1), 139-176. DOI:10.3102/00346543051001139

Guskey, T. R. (2011). Stability and Change in High School Grades. *NASSP Bulletin, 95*(2), 85-98. DOI:10.1177/0192636511409924

Guskey, T. R. (2015). *On Your Mark.* Bloomington: Solution Tree Press.

Guskey, T. R., & Link, L. J. (2018). Exploring the Factors Teachers Consider in Determining Students' Grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303-320. DOI:10.1080/0969594X.2018.1555515

Healy, K. L. (1935). A Study of the Factors Involved in the Rating Of Pupils' Compositions. *The Journal of Experimental Education, 4*(1), 50-53. DOI:10.1080/00220973.1935.11009995

Helmke, A., & Schrader, F. W. (1987). Interactional Effects of instructional Quality and teacher Judgment Accuracy on Achievement. *Teaching and Teacher Education, 3*(2), 91-98. DOI:10.1016/0742-051X(87)90010-2

Hoge, R. D. (1983). Psychometric Properties of Teacher-Judgment Measures of Pupil Aptitudes, Classroom Behaviors, and Achievement Levels. *Journal of Special Education, 17*(4), 401-429. DOI:10.1177/002246698301700404

Hoge, R. D. (1984). The Definition and Measurement of Teacher Expectations: Problems and Prospects. *Canadian Journal of Education, 9*(2), 213-228. DOI:10.2307/1494604

Hoge, R. D., & Butcher, R. (1984). Analysis of Teacher Judgments of Pupil Achievement Levels. *Journal of Educational Psychology, 76*(5), 777-781. DOI:10.1037/0022-0663.76.5.777

Hoge, R. D., & Coladarci, T. (1989). Teacher-based Judgments of Academic Achievement: A Review of the Literature. *Review of Educational Research, 59*(3), 297-313. DOI:10.3102/00346543059003297

Hoge, R. D., & Cudmore, L. (1986). The Use of Teacher-judgment Measures in the Identification of Gifted Pupils. *Teaching and Teacher Education, 2*(2), 181-196. DOI:10.1016/0742-051X(86)90016-8

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The Concurrent Validity of Standardized Achievement Tests by Content Area using Teachers' Ratings as Criteria. *Journal of Educational Measurement, 22*(3), 177-182. DOI:10.1111/j.1745-3984.1985.tb01056.x

Hulten, C. E. (1925). The Personal Element in Teachers' Marks. *Journal of Educational Research, 12*(1), 49-55. DOI:10.1080/00220671.1925.10879575

Klapp Lekholm, A. (2011). Effects of School Characteristics on Grades in Compulsory School. *Scandinavian Journal of Educational Research, 55*, 587-608. DOI:10.1080/00313831.2011.555923

Klapp Lekholm, A., & Cliffordson, C. (2008). Discrepancies Between School Grades and Test Scores at Individual and School Level: Effects of Gender and Family Background. *Educational Research and Evaluation, 14*, 181-199. DOI:10.1080/13803610801956663

Klapp Lekholm, A., & Cliffordson, C. (2009). Effects of Student Characteristics on Grades in Compulsory School. *Educational Research and Evaluation, 15*(1), 1-23. DOI:10.1080/13803610802470425

Lauterbach, C. E. (1928). Some Factors Affecting Teachers' Marks. *Journal of Educational Psychology, 19*(4), 266-271. DOI:10.1037/h0074553

Leinhardt, G. (1983). Novice and Expert Knowledge of Individual Student's Achievement. *Educational Psychologist, 18*(3), 165-179. DOI:10.1080/00461528309529272

Marmeleira, J., Folgado, H., Martinez Guardado, I., & Batalha, N. (2020). Grading in Portuguese Secondary School Physical Education: Assessment Parameters, Gender Differences, and Associations with Academic Achievement. *Physical Education and Sport Pedagogy, 25*(2), 119-136. DOI:10.1080/17408989.2019.1692807

McCandless, B. R., Roberts, A., & Stames, T. (1972). Teachers' Marks, Achievement Test Scores, and Aptitude Relations with respect to Social Class, Race, and Sex. *Journal of Educational Psychology, 63*(2), 153-159. DOI:10.1037/h0032646

McMillan, J. H. (2001). Secondary Teachers' Classroom Assessment and Grading Practices. *Educational Measurement: Issues and Practice, 20*(1), 20-32. DOI:10.1111/j.1745-3992.2001.tb00055.x

McMillan, J. H. (Ed.). (2013). *SAGE Handbook of Research on Classroom Assessment.* Los Angeles, CA: SAGE Publications, Inc.

McMillan, J. H. (2018). *Classroom Assessment: Principles and Practice that Enhance Student Learning and Motivation* (7th ed.). New York, NY: Pearson Education, Inc.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary Teachers' Classroom Assessment and Grading Practices. *Journal of Educational Research, 95*(4), 203-213. DOI:10.1080/00220670209596593

Natriello, G. (1987). The Impact of Evaluation Processes on Students. *Educational Psychologist, 22*(2), 155-175. DOI:10.1207/s15326985ep2202_4

Pattison, E., Grodsky, E., & Muller, C. (2013). Is the sky Falling? Grade Inflation and the Signaling Power of Grades. *Educational Researcher, 42*(5), 259-265. DOI:10.3102/0013189x13481382

Rauschenberg, S. (2014). How Consistent Are Course Grades? An Examination of Differential Grading. *Education Policy Analysis Archives, 22*(92), 1-37. DOI:10.14507/epaa.v22n92.2014

Sawyer, R. (2013). Beyond correlations: Usefulness of High School GPA and Test Scores in Making College Admissions Decisions. *Applied Measurement in Education, 26*(2), 89-112. DOI:10.1080/08957347.2013.765433

Schwab, K., Moseley, B., & Dustin, D. (2018). Grading Grades as a Measure of Student Learning. *SCHOLE: A Journal of Leisure Studies and Recreation Education, 33*(2), 87-95. DOI:10.1080/1937156X.2018.1513276

Sharpley, C. F., & Edgar, E. (1986). Teachers' Ratings vs Standardized Tests: An Empirical Investigation of Agreement Between Two Indices of Achievement. *Psychology in the Schools, 23*(1), 106-111. DOI:10.1002/1520-6807(198601)23:1<106::AID-PITS2310230117>3.0.CO;2-C

Silberstein, N. (1922). The Variability of Teachers' Marks. *English Journal, 11*, 414-424.

Sims, V. M. (1933). Reducing the Variability of Essay Examination Marks through Eliminating Variations in Standards of Grading. *The Journal of Educational Research, 26*, 637-647. DOI:10.1080/00220671.1933.10880358

Starch, D., & Elliott, E. C. (1912). Reliability of the Grading of High-School Work in English. *School Review, 20*, 442-457. DOI:10.1086/435971

Starch, D., & Elliott, E. C. (1913a). Reliability of Grading Work in Mathematics. *School Review, 21*(4), 254-259. DOI:10.1086/436086

Starch, D., & Elliott, E. C. (1913b). Reliability of Grading Work in History. *School Review, 21*(10), 676-681. DOI:10.1086/436185

Thorsen, C., & Cliffordson, C. (2012). Teachers' Grade Assignment and the Predictive Validity of Criterion-Referenced Grades. *Educational Research and Evaluation, 18*(2), 153-172. DOI:10.1080/13803611.2012.659929

Woodruff, D. J., & Ziomek, R. L. (2004). *High School Grade Inflation from 1991 to 2003 (Research Report Series 2004-04).* Iowa City: ACT. DOI:10.1.1.409.9896