Language Assessment Quarterly, 12: 213–233, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2015.1010726

Routledge
Taylor & Francis Group

Teachers' Grading Decision Making: Multiple Influencing Factors and Methods

Liying Cheng and Youyi Sun
Queen's University, Kingston, Ontario, Canada

This study investigated Chinese secondary school English language teachers' grading decision making, focusing on the factors they considered and types of assessment they used for grading. A questionnaire was issued to 350 secondary school English language teachers in China. Descriptive analyses of the questionnaire data showed that these teachers of English considered achievement and non-achievement factors in grading, placing greater weight on non-achievement factors, such as effort, homework, and study habits, and that they used multiple types of assessment, including performance and project-based assessment, teacher self-developed assessment, as well as paper and pencil tests for grading. MANOVA results suggested that both internal and external factors, such as the grade level teachers teach, the assessment training they have received, and their class size affect different aspects of their grading decision making. Multiple regression results further showed a significant relationship between the factors teachers considered and the types of assessment they used for grading. This study contributes to the understanding of the classroom English language teachers' grading decision making in general and especially in the Chinese context and has significant implications for teacher education.

INTRODUCTION

The history of grading dates back to the 1640s when early American universities used examinations mainly for the purpose of awarding degrees (Brookhart, 2004). Historically, concerns and doubts about the adequacy and effectiveness of various grading systems have been expressed by teachers, educational administrators, and researchers (Black & Wiliam, 1998; Guskey & Bailey, 2001; Teaf, 1964). These stakeholders (e.g., Dobbin & Smith, 1960; Finkelstein, 1913) were primarily concerned with the reliability of teachers' grading, focusing their attention on pursuing commonly accepted reliable grading systems in the educational setting. Recently, with the growing popularity of the notion of assessment *for* learning (e.g., Hume & Coll, 2009; Kirton, Hallam, Peffers, Robertson, & Stobart, 2007; Marshall & Drummond, 2006) and with increasing interest in research on classroom assessment (Andrade, 2009), researchers investigating teachers' grading decision making have shown that teachers often consider an array of achievement and non-achievement factors in making grading decisions (Guskey, 2011; Randall & Engelhard,

Correspondence should be sent to Liying Cheng, Queen's University, Education, 511 Union Street, Duncan McArthur Hall, Kingston, Ontario K7M 5R7, Canada. E-mail: liying.cheng@queensu.ca

2009; Yesbeck, 2011). These factors may directly conflict with each other and thus complicate the ability to interpret grades assigned by teachers (Brookhart, 1991; Cross & Frary, 1996).

Research on classroom assessment of teachers of English as a foreign or second language has emerged only recently (Brindley, 2007; Rea-Dickins, 2004). However, relatively few studies conducted so far have focused on these teachers' grading practices. Cheng and Wang (2007) and Davison (2004) provided documentation and comparison of teachers' grading in Hong Kong, Australia, Canada, and mainland China. These studies have shown the complexity of teachers' assessment practices, which are influenced by their beliefs and external contexts. However, the factors that determine the grades that teachers assign and the assessment methods they use to assign grades are still unknown.

Investigating teachers' decision-making practices for grading has particularly significant implications in the Chinese context considering the long, important, yet socially accepted role implications in the Chinese context considering the long, important, yet socially accepted role that examinations have played in China (Cheng, 2010). Recently, there have been an increasing number of discussions on system-wide evaluation and scoring reforms in large-scale high-stakes testing in China (Guo, 2007; Liu, 2007), and studies of different grading procedures, such as percentage grading and letter grading (Liu, 2005), as well as studies of standards-based grading (Bian & Shan, 2006). However, empirical studies that investigate English language teachers' grading decision making within the Chinese context are very limited.

TEACHERS' GRADING PRACTICES

Newton (2007) defines grading simply as the process of assigning an evaluative mark and proposes that grading should be considered exclusively at the judgment level in a purely standards-referenced technical process. This suggests that teachers' grading could be improved through professional training. However, a wealth of research evidence has shown that classroom teachers do not always follow recommended grading practices. For example, in pre-service and in-service teacher education, the recommended practice for grading is to base students' grades exclusively on their academic achievement (Dyrness & Dyrness, 2008; McMillan, 2008; Merwin, 1989; O'Connor, 2007; Wormeli, 2006). In practice, however, teachers tend to consider an array of factors, such as effort, work habits, and achievement when assigning grades (Guskey, 2011; Randall & Engelhard, 2009; Yesbeck, 2011). Brookhart (1991) noted that teachers who have received assessment training and who have knowledge of measurement theory and principles of grading still struggle with the issue of grading on achievement alone. The discrepancy between teachers' grading practices and the recommended grading practices reflects the dual roles of teachers in the classroom assessment setting: judges and coaches (Bishop, 1992; Wilson, 1996). As judges, they should base students' grades exclusively on achievement, giving priority to fairness, justice, and objectiveness of their judgment decision making. As coaches, they appeal to what is beneficial for students' development, building in considerations of many non-achievement factors, such as effort, encouragement, and improvement through grading. Bishop (1992) argued that teachers cannot act as judges and coaches at the same time and suggested that teachers should give up the judging role to external assessment and focus on developing mentoring relationships with their students to fully function as coaches, mentors, or advocates for students in their own classroom assessment.

However, from the perspective of validity, the inclusion of factors other than academic achievement in grading may present construct-irrelevant variance, thus threatening the construct validity and interpretability of the grades that teachers assign. Brookhart (1991, 1993) used Messick's (1989) validity framework to understand the discrepancy between the recommended and teachers' actual grading practices. She pointed out that in grading, teachers' concerns over the many uses and the consequences of grade use sometimes outweigh their consideration of grade interpretation. "Teachers' decisions about what to put into grades work backward from the use: Given report cards, teachers must decide what to put into the grades that go on them" (Brookhart, 1993, p. 125). This suggests that teachers' grading should be considered as a decision-making process rather than a mere technical process in which teachers use the standards to assign grades as judgments about students' achievement, as proposed by Newton (2007).

McMillan and Nash (2000) proposed a model of teachers' classroom assessment and grading decision making. Their model highlights the influences of both the internal factors (e.g., teachers' beliefs and values about learning) and the external factors (e.g., pressures of state accountability testing and parental influence) on teachers' grading decision making. These factors and other classroom realities influence teachers' use of different types of assessment, such as tests, quizzes, and performance assessments. In a more recent case study, Zoeckler (2007) examined how English language teachers in an American high school attempted to arrive at a fair grade while weighting both achievement and non-achievement factors and the role of teachers' expectations. Results of this study indicated that grading was influenced by grading systems, teachers' perceptions of effort, and their concerns for moral development.

Cheng and Wang (2007) investigated English language teachers' grading practices in three different tertiary institutional contexts: Canada, Hong Kong, and mainland China. They identified a range of contextual factors that may help to account for the differences in teachers' grading practices across these settings. For example, they analyzed teachers' preferences for different methods of grading in the beliefs these teachers held about the orientation of their assessment. They also noted that practicality, particularly class size, was another factor that had influence on teachers' assessment practice.

A considerable number of studies have investigated the relationships between teachers' grading practices and the teacher characteristics, such as gender, years of experiences (Cizek, Fitzgerald, & Rachor, 1995), the grade level at which teachers teach (Randall & Engelhard, 2009), and the subject matter they teach (McMillan, 2001). These studies have reported mixed results (Duncan & Noonan, 2007), showing that teachers' grading practices were highly variable, and unpredictable from these characteristics. One variable that has also received relatively more research attention is the amount of assessment training teachers receive. Brookhart (1993) found that **"**measurement instruction makes a difference in how teachers think about the meaning of grades but not in the amount or kind of thinking they do about the value implications and social consequences of grades" (pp. 137–138). Cheng, Rogers, and Wang (2008) argued that while training may have an influence on teachers' choice of assessment methods, "overall, teaching contexts played a bigger role compared with teacher characteristics" (p. 25). Within the Chinese context, large class size is a distinguishing feature of the secondary educational context. Moreover, there are other contextual factors of junior and senior secondary education in China that may also influence the assessment training that junior and senior secondary school teachers receive. Thus, the current study focuses on three factors that influence teachers' grading decision making: training

in assessment, grade level, and class size. Teachers' grading decision making was investigated for the factors they considered as well as the assessment methods they used for grading.

To sum up, previous research on classroom teachers' grading practices has shown the complexity of grading. It is a complex evaluation process that requires teachers to make professional decisions to accommodate their classroom realities and the impact of external factors. Teachers make their assessment and grading decisions based on their knowledge, beliefs, expectations, and values about teaching and learning. Teachers' beliefs and values often conflict with the external impact and demands. This tension presents challenges to classroom teachers. They need to, on the one hand, convey clear and interpretable information through grades to different stakeholders about students' achievement, and on the other hand, consider the functions of grades such as motivating and enhancing students' learning, as well as the undesirable consequences of grade uses. To help teachers address these challenges, researchers (e.g., Brookhart, 2003; McMillan, 2003; Moss, 2003; Shepard, 2000) have called for reconceptualization of measurement theory to make it more relevant to the assessment context of classroom teachers. For this, it is necessary to understand how teachers make their grading decisions in particular social, cultural, and educational contexts.

The current study aims to investigate the effects of three teacher- and teaching-related variables (grade level, assessment training, and class size) on grading decision making of Chinese secondary school English language teachers in the factors they consider and the types of assessment they use for grading. Secondary school teachers were chosen as the focus of the study, considering the increasing concerns over grading at the secondary school level leading to university entrance in China (Bian & Shan, 2006; Liu, 2005). Specifically, this study addresses the following three research questions:

1. To what extent do the Chinese secondary school English language teachers consider various factors and use various types of assessment for grading?
2. How are the different aspects of these teachers' grading decision making (i.e., the factors they consider and the assessment types they use) affected by variables such as grade level, assessment training, and class size?
3. What are the relationships between the factors that teachers consider and the assessment types that they use for grading?

METHOD

Context of the Study

Secondary education in China includes junior secondary (初中) and senior secondary (高中) schools. Students study in junior secondary schools for three years (Junior High 1 to 3) and take a competitive entrance exam if they want to continue their formal education in senior secondary schools for another three years (Senior High 1 to 3). Senior secondary school graduates need to take the highly competitive university entrance examination for higher education in colleges or universities. In the province where this study was conducted, the senior secondary school entrance test and the university entrance examination are administered by education departments at the municipal level and the provincial level, respectively. In both test batteries, English is one

of the three compulsory and most weighted subjects; the other two are Chinese and mathematics. English inspectors, who are staff members in the education departments at provincial or municipal levels, are responsible for assisting teachers and students in their teaching, learning, and preparation for the exams. They are also involved in teacher training at different levels.

Participants

The participants of this study were 350 secondary school English language teachers from 20 schools in 5 school districts of a northern city in China (76.1% females; 23.9% males). This city was chosen as the research site because the university entrance examination has particularly high stakes and is extremely competitive for the students there due to the huge test-taking population. This, in turn, has had an impact on teachers' grading and thus makes a good case to study teachers' grading practice as this relates to the external, large-scale high-stakes testing.

Most of the participants were between 26 and 40 years old (26–30 = 23.9%; 31–35 = 31.6%; 36–40 = 21.8%). They were teaching in junior secondary ($N = 188$) and senior secondary ($N = 162$) schools, with their teaching experiences ranging from 1 year to 35 years ($M = 12.6$; $SD = 7.1$, range = 34). For level of education, 55.3% hold certificate/diploma in teaching English as a foreign language, 38% hold a BA or BEd degree. Among the participants, 33.7% had completed a full course on language assessment, and 32.2% had completed a course in which language assessment was one of the topics, and 24.4% had no training in language assessment in the course of their academic studies or in-service training. Their mean workload was 9.2 hours per week ($SD = 2.8$, range = 16.25), with a mean class size of 54.5 students ($SD = 10.1$, range = 85).

Instrument

The instrument used to collect data in this study was a questionnaire designed in English with key terms translated into Chinese to ensure the participants' accurate understanding of the questions. The questionnaire was designed on the basis of previous research on teachers' assessment and grading practices (Brookhart, 1993; Frary, Cross, & Weber, 1993; McMillan, Myran, & Workman, 2002). It included four sections. The first section aimed to measure the extent to which teachers consider different factors in assigning grades (17 items) on a 5-point scale ranging from 1 as *Never* to 5 as *Always*. Items in this section included factors that teachers consider in giving grades, such as student performance, effort, and improvement. For example, the first question was "I consider disruptive student behavior in class when assigning final grades."

The second section was designed to measure the extent to which teachers use different types of assessment to determine grades (10 items), again using a 5-point scale with 1 as *Never* and 5 as *Always*. Items in this section included assessment types that teachers use in giving grades, such as examinations, performance assessment, and teacher self-developed assessment. The third section consisted of three scenarios and respective open-ended questions exploring the meaning and values associated with grades assigned by teachers. Because of the space limit, this portion of the questionnaire is not reported in this article. The last section of the questionnaire included eight items gathering demographic data of the participants (gender and age) and background information related to teaching (educational qualification, teaching experience, grade level taught, workload, class size, and training in assessment). The questionnaire was piloted with 15 secondary school English language teachers in China, and some minor modifications were made on

the basis of the teachers' feedback and item analyses.[1] The current article reports findings about the factors that teachers considered and types of assessment methods that they used for grading. Cronbach's alpha reliability coefficients for section 1 and 2 measuring these two scale variables were .89 and .76, respectively.

Data Collection

Two English inspectors of the city, one for junior secondary school and the other for senior secondary school, helped to contact the teachers in the city and mailed them the questionnaires as well as the "Letter of Information" and the "Consent Form" of the study for research ethics. Teachers were informed that their participation was entirely voluntary and were assured that their responses would be confidential and that the information they provided would not be used for identification. Participants took 20–30 minutes to complete the questionnaire in their offices. The completed questionnaires were mailed to the two inspectors and were collected by the researchers. The return rate was 89.2%.

Data Analysis

Descriptive statistics were calculated for the variables on the questionnaire, and outliers were detected. For example, for the variable of workload, values less than 3.75 hours were regarded as outliers because the minimum workload for the local secondary school English teachers was 5 teaching sessions per week (45 min per session). The proportion of cases with missing data across all the variables measured in the first and second sections was small, the maximum being 4.29%. A missing data analysis was conducted by using Little's MCAR test (Little, 1988), and the results indicated that the data were missing completely at random (MCAR). Given the small proportion of missing data and the MCAR missing data pattern, it is relatively "safe" to use the typical method of listwise deletion (SPSS Inc., 2009) in analyzing the two variables (factors considered in grading and types of assessment methods used in determining grades). Therefore, bias associated with missing data was not considered a major concern in the follow-up analysis.

Four analytical procedures were then used. First, descriptive statistics were computed for the items on the first two sections of the questionnaire to summarize the extent to which teachers considered various factors in assigning grades and used various types of assessment to determine grades.

Second, principal component analysis (PCA) was conducted first for the items relating to factors teachers considered and then for the terms relating to types of assessment they used in grading. The purpose of the PCA is to explore without a theory to see what patterns emerge in the data (Brown, 2009); therefore, exploratory factor analysis was not conducted in the current study. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy for the factors teachers considered in grading was .89. The KMO for the assessment methods teachers used was .78. Bartlett's test of sphericity was significant ($p < .001$) for both variables concerning factors teachers considered and variables concerning assessment methods they used in grading. These statistics indicate

[1] The initial questionnaire included 19 items in the first section and 11 items in the second section. Five teachers of English from China were first asked to review the items for clarity and completeness. Appropriate revisions were made, and two items from the first section and one item from the second section were deleted.

that the variables are correlated enough for a PCA to be appropriate. Varimax was used as the method for rotation in the PCAs to minimize the number of variables with high loadings on a component, thereby enhancing the interpretability of the component. To guide components selection, two standards were set: (a) they obtained eigenvalues of 1 or above and (b) they were located at or above a point of inflexion (or "elbow") on a scree plot. Component scores were calculated by using the Anderson-Rubin method for each case, because this takes into account the respective weight of each item to the component. These scores were used in subsequent analyses to represent the values of the components.

Third, two sets of MANOVA analyses were conducted by using component scores from the PCAs as dependent variables to examine the effects of grade level teachers taught, assessment training they received, and their class size on their grading practices. The first set of MANOVA analyses was conducted for the factors teachers considered and the second set for the assessment methods they used in grading. Where necessary, post hoc analyses were conducted by using Bonferroni. Ideally, a 3-way MANOVA ought to be conducted in this case, yet this procedure was not conducted, given the small sample size and considering that this would create extremely unequal cell sizes, resulting in large deviations from the assumptions of MANOVA and making interpretations more difficult. Thus, three separate two by two 2-way MANOVAs (grade level by assessment training, class size by grade level, assessment training by class size) were conducted in each set.

For these analyses, the participants were grouped in grade level (junior secondary and senior secondary school teachers), training in assessment (teachers with assessment training [i.e., teachers who had completed a full course on language assessment], and teachers without assessment training), and class size (small classes with less than 55 students and large classes that have more than 56 students based on the mean class size). The frequency distribution chart of the class size variable showed that the distribution was bimodal, with the upper group (large class size) being well over 60 and lower group (small class size) well below 55. To examine the homogeneity of variance across the three grouping factors (assessment training, class size, and grade level), 12 Levene's tests (6 for the factors teachers considered and 6 for the assessment types they used) were conducted and variance ratios were calculated. Levene's test was not found to be statistically significant for any of the dependent variables. The assumption of univariate normality was checked in turn by means of skewness and kurtosis values as well as histograms and plots. The results indicated that the assumption of univariate normality was satisfied. The MANOVAs were followed up with discriminant function analyses (DFAs). The purpose of the DFAs was to explain the differences between junior and senior secondary teachers; teachers with small and large class sizes; and teachers with and without training in assessment in the factors they considered and types of assessment they used for grading.

Finally, three multiple-regression analyses were conducted to investigate how the three types of assessment used in teachers' grading could be predicted by the factors they considered. The assumption of linearity was checked by examining the residual plots (plots of the standardized residuals as a function of standardized predicted values).

SPSS 17.0 was used for data analysis in this study. A critical value of $\alpha = 0.05$ was set for these analyses.

RESULTS

This section first reports results of analyses on the factors teachers considered in making their grading decisions and on the types of assessment teachers used for grading. It then presents findings on the effects of teacher- and teaching-related variables (grade level, assessment training, and class size) on teachers' grading practices as well as findings on the relationships between factors teachers considered and assessment methods they used for grading.

Factors That Teachers Considered in Grading

The extent to which teachers considered different factors in making grading decisions was examined by the respective mean scores of the items in the first section of the questionnaire. Table 1 presents the means and standard deviations of these items. These descriptive statistics indicated that various factors contributed to the grades that teachers assigned. The means ranged from 2.83 to 4.27 on a 5-point Likert-type scale. There were six items with mean scores over 4.00. *Effort* was the factor that was most often considered by this group of teachers of English. The item with the lowest mean score was grade distribution of other teachers. The mean scores of the remaining items ranged from 3.12 to 3.97. The means of the two items concerning achievement (academic performance and specific learning objectives mastered) were 3.61 and 3.52, respectively. The standard deviations for all 17 items were fairly large, showing considerable variation in the extent to which this group of teachers considered various factors in their grading practices.

TABLE 1
Descriptive Statistics for Factors Teachers Considered in Grading

Factors used in Determining Grades	N	M	SD
Student effort	344	4.27	.92
Nontest indicators for borderline cases	346	4.27	.91
Improvement of academic performance	346	4.25	.99
Quality of completed homework	346	4.24	.93
Study habits	337	4.12	1.05
Completion of homework	346	4.11	1.01
Disruptive student behaviour	348	3.97	1.11
The degree to which student pay attention, participate in class or both	346	3.89	1.02
Academic ability level	345	3.89	1.00
Academic performance	345	3.61	.92
Specific learning objectives mastered	346	3.52	1.12
Performance compared with other students	346	3.48	1.11
Formal or informal school policy	344	3.34	1.18
Inclusion of zero for incomplete assignments	339	3.29	1.02
Extra credit for non-academic performance	337	3.23	1.16
Performance compared with other students from previous years	343	3.12	1.28
Grade distribution of other teachers	343	2.83	1.31

TABLE 2
Component Loadings for Factors Considered in Grading

	Component		
Item	*1*	*2*	*3*
Grade distribution of other teachers	**.713**	.178	.119
Specific learning objectives mastered	**.697**	.155	.139
Inclusion of zero for incomplete assignments	**.670**	.099	.370
Formal or informal school policy	**.642**	.069	.216
Performance compared with other students from previous years	**.602**	.067	.241
The degree to which student pay attention, participate in class or both	**.572**	.327	.172
Completion of homework	.218	**.722**	.322
Student effort	.079	**.704**	.303
Quality of completed homework	.156	**.689**	.435
Study habits	.216	**.604**	.448
Improvement of academic performance	.089	**.567**	−.082
Disruptive student behavior	.438	**.545**	−.055
Academic performance	.313	.004	**.733**
Academic ability level	.034	.330	**.672**
Extra credit for non-academic performance	.325	.150	**.594**
Performance compared with other students	.290	.186	**.552**
Nontest indicators for borderline cases	.227	.424	.430

The PCA resulted in three components. The component loadings of different items are summarized in Table 2. These loadings represent correlations between components and observed variables. Common themes among highly loading items on a component can help identify the construct that the component represents. Six items loaded highly on the first component. These items included grade distribution of other teachers, mastered learning objectives, incomplete assignments, school policy, performance compared with students from previous years, and attention and participation. This component was, therefore, labeled *norm/objective-referenced factor* (i.e., teachers considered the learning goal and other students' grades as references when determining the grades they assigned). Six items loaded highly on the second component. They primarily concerned homework, effort, improvement, work habits, and disruptive behavior. This component was then defined as *effort factor.* The items that loaded highly on this factor show relatively higher means and lower standard deviations (see Table 1). Finally, the four items that loaded highly on component 3 seemed to relate to performance: academic and nonacademic performance, academic ability, and performance compared with other students. Thus, this component was referred to as *performance factor*.

Types of Assessment That Teachers Used for Grading

Table 3 shows the means and standard deviations for the different types of assessment methods that teachers used for grading. These descriptive statistics indicate that teachers used various types of assessment methods in their grading practices to different degrees and with some variability.

TABLE 3
Descriptive Statistics for Assessment Methods Teachers Used in Grading

	N	M	SD
Major examinations	348	4.28	.94
Quizzes	346	3.78	1.08
Objective assessment[3]	345	3.71	1.16
Assessment designed by yourself	345	3.60	1.09
Projects by individuals	348	3.59	1.10
Performance assessment	347	3.52	.99
Projects by teams	348	3.51	1.20
Oral presentation	347	3.33	1.23
Assessment provided by publishers	335	3.08	1.13
Essay-type questions	347	2.97	1.30

TABLE 4
Component Loadings for Types of Assessment Methods Used in Grading

	Component		
Item	*1*	*2*	*3*
Projects by teams	**.827**	.093	.042
Essay-type questions	**.786**	.178	−.103
Assessment provided by publishers	**.671**	.145	.023
Projects by individuals	**.647**	−.009	.303
Performance assessment	**.533**	.420	.128
Assessment designed by yourself	.177	**.765**	.118
Objective assessment	.016	**.741**	.264
Oral presentation	.187	**.712**	−.164
Major examination	.005	.007	**.802**
Quizzes	.146	.177	**.788**

The means of the 10 measured assessment methods ranged from 2.97 to 4.28. Major examination ($M = 4.28$, $SD = .94$) was the type of assessment that was most frequently used by this group of teachers, and the teachers showed the lowest variation in using this type of assessment. Essay-type questions ($M = 2.97$, $SD = 1.30$) were the type of assessment with the lowest mean score and showed the highest variation, suggesting that the teachers reported using this assessment method the least. The mean scores of the remaining assessment methods ranged from 3.08 to 3.78 with *SD* from .99 to 1.23.

The PCA produced three components of assessment types used by teachers to grade their students. The factor loadings of different items are summarized in Table 4. Five items loaded highly on the first component. These items primarily concerned *performance and project-based assessments*. The three items loading highly on the second component seemed to relate to *teacher*

[3] Objective assessment refers to classroom testing that teachers use in their classroom instruction primarily in the form of multiple choice, whereas major examinations refer to any tests at the school level and beyond usually in the mid or final term.

self-developed assessment. The two items that loaded highly on component 3, major examination and quiz, were the assessment types with the highest mean scores. They were primarily *paper and pencil tests* often used for summative purpose.

Effects of Grade Level, Assessment Training, and Class Size on Factors Considered in Grading

Three 2-way MANOVAs were run to test the effects of grade level teachers taught, assessment training they received, and the sizes of their classes, as independent variables, on the factor components they considered in grading. Factor scores on the three components (i.e., the *norm/objective-referenced factor, the effort factor,* and the *performance factor*) were used as dependent variables in these analyses. Partial η^2 is reported to show the proportion of the variance in the dependent variables that is attributable to the factors in question.

The grade level by assessment training MANOVA produced significant interaction effect ($F(3, 166) = 7.59$, $p < .001$, partial $\eta^2 = .12$), as well as significant main effect of training ($F(3, 166) = 3.01$, $p = .03$, partial $\eta^2 = .05$). The follow-up ANOVAs yielded significant interaction effects for the *norm/objective-referenced factor,* $F(1, 168) = 14.50$, $p < .001$, partial $\eta^2 = .08$ and the *effort factor,* $F(1, 168) = 4.50$, $p = .035$, partial $\eta^2 = .03$. The interaction effects of assessment training and grade level on teachers' considerations of these two factors are illustrated by Figure 1 and Figure 2, respectively. These figures show that the effect of assessment

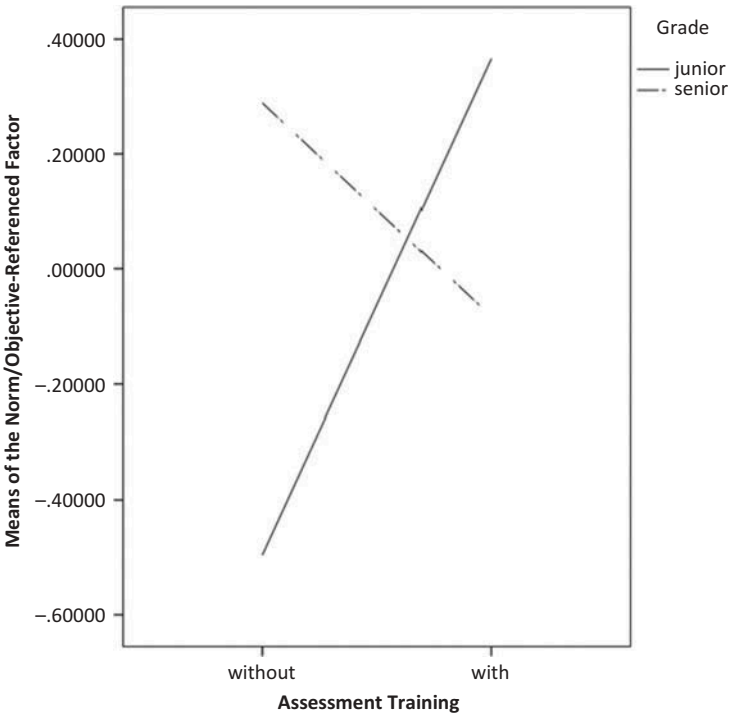

FIGURE 1 Interaction effect of training and grade on the norm/objective-referenced factor.

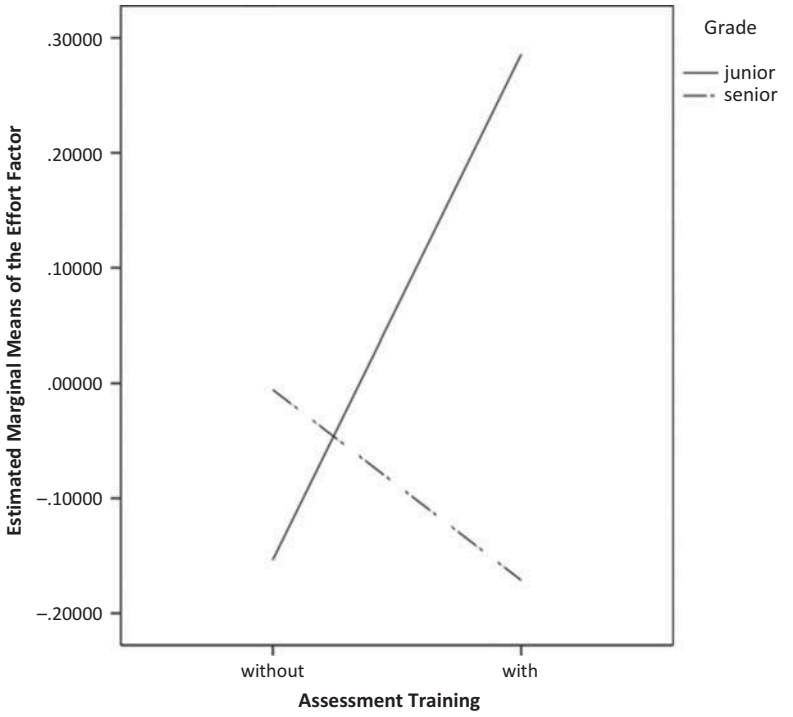

FIGURE 2 Interaction effect of training and grade on the effort factor.

training is prevalent for both junior and senior secondary school teachers in their considerations of the *norm/objective-referenced* and the *effort* factors in grading. However, the direction of the effect is different for these two groups of teachers. For senior secondary school teachers, their considerations of the two factors declined after the assessment training, whereas for junior secondary school teachers, their considerations seemed to be enhanced with assessment training.

The class size by grade MANOVA did not yield any significant interactions or main effects. The assessment training by class size MANOVA produced significant main effects of training on the *norm/objective-referenced* and the *effort* factors, as well as main effect of class size on the *effort* factors. However, because the interactions were disordinal, results for these main effects are not reported here.

The follow-up DFA revealed a discriminant function that significantly differentiated the two assessment training groups ($R^2 = .06$, $\Lambda = .94$, χ^2 (3) $= 11.04$, $p = .012$). The correlations between predictors and the discriminant function showed that the *performance* factor loaded highly on the function (r $= .-74$), and the *norm/objective-referenced* factor and the *effort* factor loaded moderately on the function (r $= .53$ and r $= .40$, respectively).

Effects of Grade Level, Training Assessment, and Class Size on Types of Assessment Used for Grading

Three 2-way MANOVAs were run to test the effects of grade level teachers taught, assessment training they received, and the sizes of their classes as independent variables on the types of assessment they used in grading. Factor scores on the three components of assessment types (i.e., performance and project-based assessment, *teacher self-developed assessment,* and *paper and pencil test*) were used as dependent variables in these analyses.

The grade level by assessment training MANOVA produced significant interaction effect ($F(3, 188) = 25.56$, $p < .001$, partial $\eta^2 = .29$) and main effect for grade ($F(3, 188) = 13.49$, $p < .001$, partial $\eta^2 = .18$). Subsequent ANOVAs showed significant interaction effects for both the *performance and project-based* assessment, $F(1, 190) = 18.72$, $p < .001$, partial $\eta^2 = .09$, and *teacher self-developed* assessment $F(1, 190) = 50.95$, $p < .001$, partial $\eta^2 = .21$. Figure 3 and Figure 4 illustrate the interaction effects of grade and training on the two assessment methods respectively. Figure 3 shows that the direction of the effect of assessment training was different for junior and senior secondary school teachers. Junior secondary school teachers' use of the *performance and project-based* assessment declined with training in assessment while senior

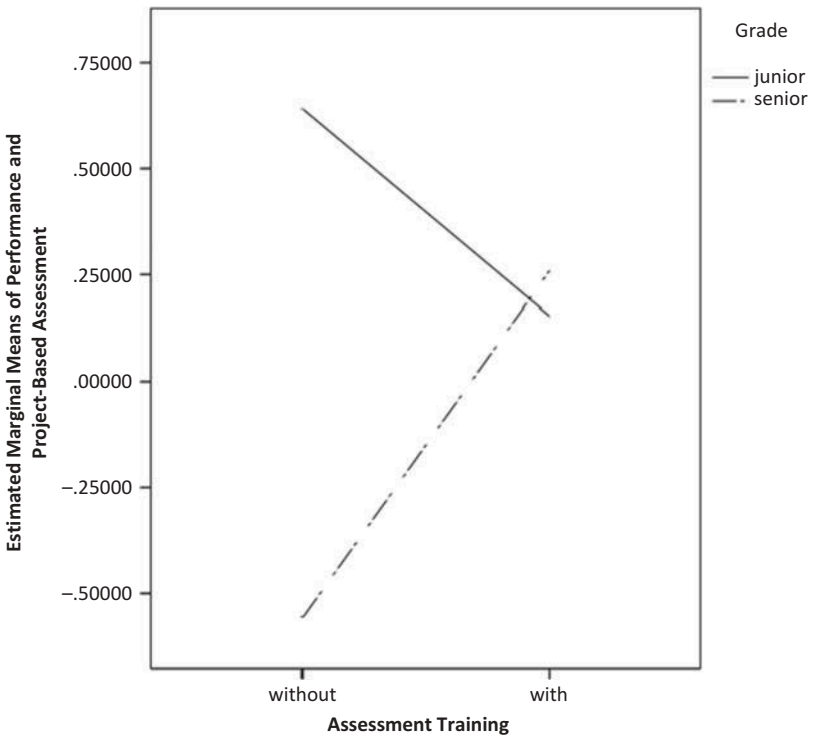

FIGURE 3 Interaction effect of training and grade on performance and project-based assessment.

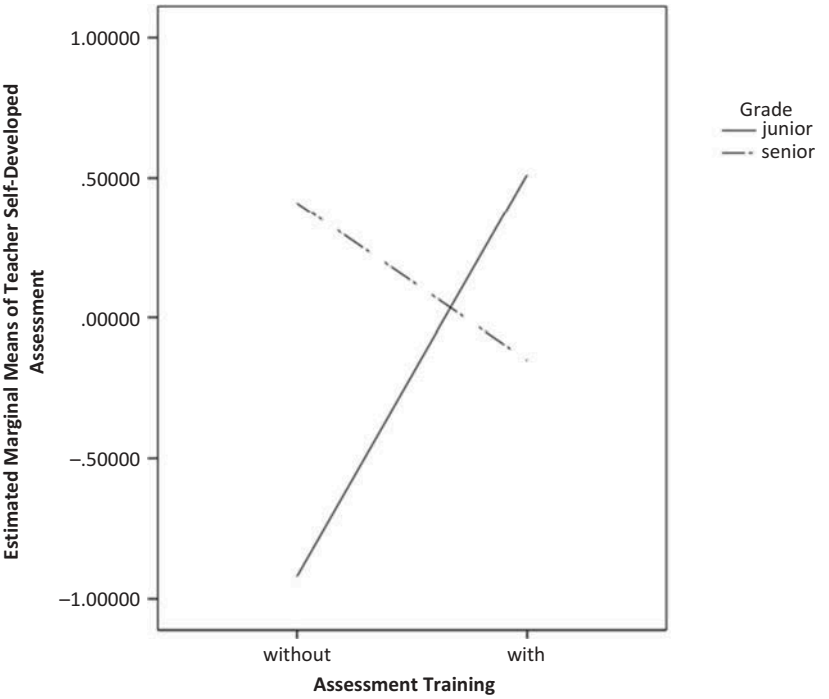

FIGURE 4 Interaction effect of training and grade on teacher self-developed assessment.

secondary teachers' use of this assessment method increased with assessment training. Figure 4 indicates that assessment training also influenced junior and senior secondary school teachers in different directions in their uses of *teacher self-developed* assessment. Junior secondary school teachers with assessment training used this type of assessment more often than those without, whereas senior secondary school teachers with assessment training were less likely to use it than those without.

The class size by grade level MANOVA yielded no significant interaction effect ($F(3, 197) = .62$, $p = .602$, partial $\eta^2 = .01$). Significant main effect of class size was produced ($F(3, 197) = 4.76$, $p = .003$, partial $\eta^2 = .07$). Subsequent ANOVAs yielded significant main effects of class size on both the *paper and pencil* test and the *teacher self-developed* assessment. Teachers with small classes were more likely to use the *teacher self-developed* assessment, $F(1, 199) = 7.44$, $p = .007$, partial $\eta^2 = .04$, whereas teachers with large classes were more likely to use the *paper and pencil* test $F(1, 199) = 4.87$, $p = .03$, partial $\eta^2 = .02$.

The assessment training by class size MANOVA produced significant main effect of class size ($F(1, 116) = 4.51$, $p = .005$, partial $\eta^2 = .11$). Subsequent analyses yielded significant main effect of class size on the *teacher self-developed* assessment. Teachers teaching small classes were found to be more likely to use this type of assessment than teachers with large classes $F(1, 116) = 7.24$, $p = .008$, partial $\eta^2 = .06$.

TABLE 5
Discrimination Function Analyses of Assessment Types Used by Teachers in Grading

Predicted Group Membership	*Function*				*Canonical Variate Correlation (r)*		
	R^2	Λ	χ^2	p	*PPBA**	*TSDA***	*PPT****
Training (without and with)	.04	.96	11.10	.011	−.24	.90	−.35
Grade (junior and senior)	.04	.96	9.88	.020	.02	.19	.97
Classsize (small and large)	.06	.94	10.16	.017	−.24	−.41	.78

*PPBA = performance and project-based assessment.
**TSDA = teacher self-developed assessment.
***PPT = paper and pencil test.

Results of the follow-up DFAs are presented in Table 5. For each of the dependent variables (grade, training, class size), a discriminant function was identified that significantly differentiated the two groups within them. However, the effect sizes of these functions were small, with R^2 ranging from .04 to .06. Λ=.94, χ^2 (3) = 11.04, p = .012). The DFAs results revealed that the predictor variables correlated highly with each of the functions. For example, *teacher self-developed* assessment loaded highly on the function discriminating the groups with and without assessment training (r = .90), and the function discriminating the groups with small and large class sizes (r = .78). *Paper and pencil* test loaded highly on the function discriminating the junior and senior secondary groups (r = .97) and loaded moderately on the function discriminating the groups with small and large class sizes (r = −.41) as well as the function discriminating the groups with and without assessment training (r = −.35).

Relationships Between Factors Considered and Methods Used in Grading

To examine the relationship between the factors that teachers considered in determining grades and the types of assessment they used, multiple-regression analyses were conducted. In these analyses, component scores on the three components of the factors considered in grading (i.e., *norm/objective-referenced factor, effort factor,* and *performance factor*) were used as independent variables; factor scores on the three components of the assessment types used in grading (i.e., performance and project-based assessment, teacher self-developed assessment, and paper and *pencil test*) were used as dependent variables. Results of these analyses are presented in Table 6. This table indicates the extent to which teachers used the three types of assessment in determining grades could be predicted by the three factors they considered in grading. In general, these factors could well predict their uses of *performance and project-based* assessment (R^2 = .26, $p < .001$) and *teacher self-developed* assessment ($R^2 = .32$, $p < .001$). The standardized beta (ß) values were all measured in standard deviation units, so they were directly comparable. These values indicate the number of standard deviations that the dependent variable will change with one standard deviation change in the independent variable and thus provide an insight into the importance of an independent variable in the model. Specifically, the *performance* factor was a powerful predictor of the *performance and project-based* assessment (ß = .47, $p < .001$) and the *norm/objective referenced* factor was a strong predictor of the *teacher self-developed* assessment

TABLE 6
Multiple Regression Results on Use of Assessment Types and Factors Considered in Grading

Variable	*ß*	T	P	*95% CI*	R^2
(Constant)		−.59	.557	[−.14, .08]	
Referential factor	.11	2.04	.042	[.00, .22]	
Effort factor	.16	3.07	.002	[.06, .27]	
Performance factor	.47	8.76	.000	[.37, .59]	.26*
Dependent variable: performance/project-based assessment					
$N = 269$					
(Constant)		.63	.530	[−.07, .13]	
Referential factor	.51	10.07	.000	[.41, .60]	
Effort factor	.24	4.73	.000	[.14, .34]	
Performance factor	.05	1.07	.284	[−.05, .16]	.32*
Dependent variable: teacher self-developed assessment					
$N = 269$					
(Constant)		.15	.879	[−.10, .12]	
Referential factor	.13	2.24	.026	[.02, .24]	
Effort factor	.18	2.97	.003	[.06, .28]	
Performance factor	.16	2.74	.007	[.04, .27]	.08*
Dependent variable: paper and pencil test					
$N = 269$					

CI = confidence interval.
*$p < .05$.

(ß = .51, $p < .001$). However, the effect size of the model using the three factors considered by teachers in grading to predict their use of *paper and pencil* assessment was relatively small ($R^2 = .08$, $p < .001$), indicating that only 8% of the variation in teachers' use of this assessment type could be accounted for by their considerations of the *norm/objective-referenced* factor, *the effort* factor, and the *performance* factor.

DISCUSSIONS

The findings of the current study show the complexity with which teachers make grading decisions in the Chinese context. First, non-achievement factors are the primary consideration of the Chinese teachers. This finding is consistent with previous studies within the North American context (e.g., Cross & Frary, 1996). The Chinese secondary school English teachers gave particular weighting to effort in their grading decisions. This is not surprising within the Chinese context, where learning is believed to depend on effort rather than ability (Wang, 2008).

Furthermore, the results from the MANOVAs indicate that there is a significant difference between teachers with and without training in assessment in their considerations in grading. However, the influence of assessment training on teachers' considerations should be considered in relation to the grade level at which they teach. Assessment training for junior secondary school teachers and for senior secondary school teachers in the context of this study may have

different foci, or junior secondary and senior secondary school teachers may perceive the assessment training in different ways. For example, assessment training influences junior and senior secondary teachers in different directions in their use of the *teacher self-developed* assessment and the *performance and project-based* assessment. One possible explanation is the significant influence of the extremely high-stakes university entrance examination on the training of senior secondary school teachers and their classroom assessment. Students' performances on the university entrance examination are often compared across cities within the province and are used as a reward and/or punishment. An important focus of the training at the city level is to help teachers align their classroom assessment with this external examination, which has high stakes for students, teachers, parents, and education administrators (Cheng, 2008; Cheng & Qi, 2006; Qi, 2005). This type of assessment training often provides the senior secondary school teachers with mock test papers prepared by the municipal education. Teachers are encouraged to use these test preparation materials in their classroom assessment. In a similar study, some Chinese students even have identified the criterion of being a good teacher as one who helps students pass examinations (Shi, 2006).

The use of a *paper and pencil* test was also found to be a significant factor that differentiates junior and senior secondary school teachers in this study. This difference again reflects the influences of the external large-scale high-stakes testing on teachers' grading decision making in the types of assessment they use for grading. Senior secondary school graduates in the context of this study take the university entrance examinations to enter institutions of higher education. English is one of the three compulsory and most weighted subjects in this test battery. The English test is a standardized test based on the senior secondary school English curriculum standards. The selection function of the test and the high stakes associated with the test influence classroom teaching and assessment to a great extent (Cheng & Qi, 2006; Qi, 2005). Gaining higher scores is the goal for both teachers and students across the three senior secondary grade levels because entering a better university is the goal for all the senior school students. Thus, senior secondary school teachers tend to align their classroom assessment with the high-stakes university entrance examination and give priority to paper and pencil tests that are primarily used for summative purpose. This examination produces a significant undesirable washback effect on teaching (i.e., "teaching-to-the-test") (Tang & Biggs, 1996, p. 163).

In contrast, although junior secondary school graduates also have to take a senior secondary school entrance test, this test is not as competitive as the university entrance examination in the city where the data were collected. In addition to going to senior secondary schools, these students can choose to go either to the job market or to technical and vocational institutions. As a result, the external test may exert less influence on the classroom English teaching and assessment in junior secondary schools.

For the types of assessment methods, it is not surprising that major examinations were most frequently used by this group of teachers, because the influence of major examinations is dominant in the Chinese assessment practices (Cheng, 2008). On the other hand, using an essay-type question was the assessment method with the lowest mean score, yet showed the highest degree of variation. In other words, this method was used by some teachers, but not others.

Add to the complex nature of grading, the types of assessments could significantly differentiate junior from senior secondary teachers, teachers with from those without assessment training, and teachers with small from those with large class sizes. Generally, the effect sizes of assessment training, grade level, and class size on the teachers' uses of different types of assessment are

larger than their effects on teachers' factor considerations. These findings indicate that assessment training, grade level, and class size have more influence on the assessment types that teachers used than on the factors they considered in grading.

The significant effects of class size on the use of the *teacher self-developed* assessment and *paper and pencil* test suggest that small class size makes it more possible for teachers to develop assessment themselves, and *paper and pencil* tests are a more convenient type of assessment for teachers to use in large-sized classrooms. Therefore, class size does matter when it comes to the reality of classroom assessment in the methods that teachers choose to use with their students, considering that the mean class size in this Chinese context is 54.5 students.

The results from the multiple regression suggested that teachers' grading could be considered a decision-making process, in which the types of assessment methods teachers used could be predicted by the factors that they considered for grading. Therefore, studies on these factors, which are primarily essential aspects of teachers' classroom assessment practices, have important implications for teacher assessment training. While it is important to enhance teachers' ability to use various types of assessment methods, it is even more important to explore the value judgment in their grading decision making. These results also show the complexity of teachers' grading practices. This complexity lies in the fact that this process is not only influenced by teacher-related factors such as training in assessment but also constrained by factors related to the teachers' classroom realities (teaching-related factors), such as class size and grade level they teach, as well as possible factors related to the broader educational and societal context. Statistically, the low variability of the use of paper and pencil tests may explain the small effect size of the model using the three factors considered by teachers in grading to predict their use of this type of assessment. However, to analyze the dominance of the paper and pencil test used for summative purpose may require a full understanding of the teaching context.

CONCLUSIONS

It is interesting but not surprising that the findings of this study are consistent with those from previously mentioned research conducted in other parts of the world (e.g., Guskey, 2011; Randall & Engelhard, 2009; Yesbeck, 2011). The focus of non-achievement factor (e.g., effort) is clearly evident in the current context. Previous literature has pointed out that the inclusion of non-achievement factors in grading potentially introduces construct-irrelevant variance and jeopardizes the interpretability of the grades assigned. To address this issue, it is perhaps important to turn our attention to understanding teachers' values about teaching and learning as well as their considerations of the consequential aspect of grading within a specific instructional and societal context (Brookhart, 1991, 1993; Cheng *et al*., 2008). A more thorough understanding from the sociocultural perspective of the values teachers hold, in particular, social, cultural, and educational contexts, is extremely important.

Assessment methods used by these teachers, however, largely reflect the collection of objective achievement data. This indicates a complex and rather contradictory result. Considering the predicted model of the factors teachers considered to the methods they chose, the result could be due to the nature of the survey on the restricted number of assessment methods. Future studies with experimental design could provide better explanation of the teachers' grading. Indeed, teachers' grading decision making is a complex process that is influenced by various contextual factors.

Furthermore, the impact of assessment training was different for junior and senior secondary teachers. To better understand teachers' grading decision making, further research is needed to investigate the content and nature of assessment training for junior and senior secondary school English language teachers in the Chinese context. Future research also needs to focus on the analysis of the local educational and societal contexts, as suggested by the study of Cheng *et al*. (2008).

The present study highlights the importance of the influences of contextual factors on teachers' grading practices. First, learning English has been a national priority in China (Lam, 2005), and scores on English examinations are important for each higher level of education (Cheng, 2008). In this context, large-scale high-stakes testing has significant impact on teachers' classroom assessment, particularly in the methods they use for grading. Second, large-scale high-stakes testing also influences the types of assessment training the teachers receive. It is essential to take into account this influence to better understand the effect of assessment training on teachers' grading decision making. Third, because the consequences of grading on students' leaning are teachers' primary concerns,[2] the factors they consider in making grading decisions reflect their beliefs and assumptions about learning, which are in turn influenced by the social, cultural, and educational contexts.

Methodologically, there is a need for further qualitative open-ended data collection about the reasons why these teachers make their grading decisions as they do and also for further data from the students they teach to triangulate the data from the teachers. Only by obtaining data from both teachers and students can the validity of the grading practices be better understood. Data analysis of this study would be further enhanced if all three independent variables in the MANOVAs were modeled to break up the analysis into three 2 by 2 matrices. Regardless of this limitation, the findings of the study contribute to an understanding of teachers' grading decision making and have important implications for future research regarding English language teacher education in the Chinese context and elsewhere in the world.

REFERENCES

Andrade, H. (2009). This issue. *Theory Into Practice*, *48*(1), 1–3.

Bian, Y., & Shan, H. (2006). Standards-based education: Promoting students' learning by applying the marking rule. [基于标准的教育：利用评分规则促进学生学习]. *Theory and Practice of Education*. [教育理论与实践], *26*(7), 30–33.

Bishop, J. H. (1992). Why U.S. students need incentives to learn. *Educational Leadership*, *49*(6), 15–18.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Brindley, G. (2007). Editorial. *Language Assessment Quarterly*, *4*(1), 1–5.

Brookhart, S. M. (1991). Letter: Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*(1), 35–36.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement*, *30*(2), 123–142.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5–12.

[2] Please see Sun & Cheng (2014) for the finding from analysis of the qualitative data on the questionnaire used in this study.

Brookhart, S. M. (2004). *Grading*. Upper Saddle River, NJ: Pearson-Merrill-Prentice Hall.

Brown, J. D. (2009). Principal components analysis and exploratory factor analysis—Definitions, differences, and choices. *Shiken: JALT Testing & Evaluation SIG Newsletter*, *13*(1), 26–30.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15–38.

Cheng, L. (2010). The history of examinations: Why, how, what and whom to select? In L. Cheng, & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 13–26). New York, NY: Routledge.

Cheng, L., & Qi, L. (2006). Description and examination of the National Matriculation English Test in China. *Language Assessment Quarterly*, *3*(1), 53–70.

Cheng, L., & Wang, X. (2007). Grading, feedback, and reporting in ESL/EFL classrooms. *Language Assessment Quarterly*, *4*(1), 85–107.

Cheng, L., Rogers, T., & Wang, X. (2008). Assessment purposes and procedures in ESL/EFL classrooms. *Assessment & Evaluation in Higher Education*, *33*(1), 9–32.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159–179.

Cross, L., & Frary, R. (1996, April). Hodgepodge grading: Endorsed by students and teachers alike. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, *21*(3), 305–334.

Dobbin, J. E., & Smith, A. Z. (1960). Marks and marking systems. In Harris, C. W. (Ed.), *Encyclopedia of educational research* (3rd ed.) (pp. 783–791). New York, NY: Macmillan Company.

Duncan, C. R., & Noonan. B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research*, *53*(1), 1–21.

Dyrness, R., & Dyrness, A. (2008). Making the grade in middle school. *Kappa Delta Pi Record*, *44*(3), 114–118.

Finkelstein, I. E. (1913). *The marking system in theory and practice. Educational psychology monographs No. 10*. Baltimore: Warwick & York, Inc.

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice*, *12*(3), 23–30.

Guo, Y. (2007). Introspection of the multi-evaluation system reform of National College Entrance Examination [对我国高考多元评价制度改革的反思. *Education and Examinations* [教育与考试], *4*, 24–28.

Guskey, T. (2011). Five obstacles to grading reform. *Educational Leadership*, *69*(3), 17–21.

Guskey, T. R., & Bailey, J. M. (2001). *Developing grading and reporting systems for student learning*. Thousand Oaks, CA: Corwin Press, Inc.

Hume, A., & Coll, R. K. (2009). Assessment of learning, for learning, and as learning: New Zealand case studies. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 269–290.

Kirton, A., Hallam, S., Peffers, J., Robertson, P., & Stobart, G. (2007). Revolution, evolution or a trojan horse? Piloting assessment for learning in some Scottish primary schools. *British Educational Research Journal*, *33*(4), 605–627.

Lam, A. S. L. (2005). *Language education in China: Policy and experience from 1949*. Hong Kong SAR, China: Hong Kong University Press.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.

Liu, Q. (2007). Thoughts on reform in the evaluation system of higher education institution enrollment and entrance examination. [高校招生考试评价体系改革的思路]. *Southeast Academic Research* [东南学术], *4*, 21–5.

Liu, W. (2005). Misconceptions on the reform from hundred-mark system to grading System. [“百分制”向“等级制”变革中的评价理念误区]. *Journal of Jiangsu Institute of Education (Social Science)* [江苏教育学院学报(社会科学版)], *21*(6), 22–23.

Marshall, B., & Drummond, M. J. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education*, *21*(2), 113–149.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20–32.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34–43.

McMillan, J. H. (2008). *Assessment essentials for standards-based education* (2nd ed.). Thousand Oaks, CA: Sage Publications.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203–213.

McMillan, J. H., & Nash, S. (2000). Teachers' classroom assessment and grading decision making. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.

Merwin, J. C. (1989). Evaluation. In M. C. Reynolds (Ed.), *Knowledge base for the beginning teacher* (pp. 185–192). Oxford, UK: Pergamon Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York, NY: MacMillan.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, *22*(4), 13–25.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170.

O'Connor, K. (2007). *A repair kit for grading: 15 fixes for broken grades*. Princeton, NJ: ETS.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stake test. *Language Testing*, *22*(2), 142–173.

Randall, J., & Engelhard, G. (2009). Differences between teachers' grading practices in elementary and middle schools. *Journal of Educational Research*, *102*(3), 175–185.

Rea-Dickins, P. (2004). Editorial: Understanding teachers as agents of assessment. *Language Testing*, *21*(3), 249–258.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.

Shi, L. (2006). The successors to Confucianism or a new generation? A questionnaire study on Chinese students' culture of learning English. *Language, Culture and Curriculum*, *19*(1), 122–147.

SPSS Inc. (2009). *SPSS Missing Values 17.0*. Chicago, IL, SPSS Inc.

Sun, Y., & Cheng, L. (2014). Teachers' grading practices: Meanings and values assigned. *Assessment in Education: Principles, Policy & Practice*, *21*(3), 326–343. doi:10.1080/0969594X.2013.768207

Tang, C., & Biggs, J. (1996). How Hong Kong students cope with assessment. In D. Watkins & J. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 159–182). Hong Kong SAR, China: Comparative Education Research Center.

Teaf, H. M. (1964). Grades: Their dominion is challenged. *The Journal of Higher Education*, *35*(2), 87–88.

Wang, F. (2008). Motivation and English achievement: An exploratory and confirmatory factor analysis of a new measure for Chinese students of English Learning. *North American Journal of Psychology*, *10*(3), 633–646.

Wilson, R. J. (1996). *Assessing students in classrooms and schools*. Toronto, Canada: Allyn & Bacon.

Wormeli, R. (2006). Accountability: Teaching through assessment and feedback, not grading. *American Secondary Education*, *34*(3), 14–27.

Yesbeck, D. M. (2011). Grading practices: Teachers' considerations of academic and non-academic factors (Unpublished doctoral dissertation). Virginia Commonwealth University, Richmond, Virginia.

Zoeckler, L. (2007). Moral aspects of grading: A study of high school English teachers' perceptions. *American Secondary Education*, *35*(2), 83–102.