Graduate Student SYMPOSIUM

Selected Papers*

Vol. 5

2009- 2010

Queen's University

Faculty of Education

Marcea Ingersoll
Editor

Rebecca Luce-Kapler
Managing Editor

GIUDING IN LANGUAGE ASSESSMENT: A LITERATURE REVIEW

Youyi Sun

ABSTRACT

This paper reviews empirical studies in the field of language assessment on grading or scoring. Most of the reviewed studies have attempted to examine systematic effects on scoring of variables associated with teachers/raters for the purpose of score consistency and reliability. Following the measurement paradigm, these studies have generally taken a positivist approach and used primarily quantitative methodology. More recently, researchers have begun to pay increasing attention to teachers' grading practices in different contexts to explore the validity of classroom-based assessment. This line of research interprets grading or scoring as a professional decision-making process, focusing on understanding teacher-raters ' grading or scoring practices in relation to broader educational, social and cultural contexts. Studies in this line of research have followed an interpretivism paradigm and employed primarily qualitative research methodology. This paper analyzes these two themes of research on grading/scoring in relation to two trends in language assessment: language performance assessment and assessment for learning, and discusses their significant implications for future research.

INTRODUCTION

Grading or scoring, as it is often called in most large-scale testing research literature, is a value-laden decision-making process. Teachers' beliefs, particularly their values and beliefs about teaching and learning, have much influence on the grading process. A host of factors related to rater characteristics and the interactions between these factors and other facets of testing such as the rating scale and the task have effects on the scoring process. On the other hand, grades or scores may be used for different purposes by different stake-holders - students, parents, school administrators, teachers themselves, and other score users _ and may have various effects on them. Therefore, grades or scores assigned by teachers/raters should convey to stake-holders accurate, consistent, interpretable and appropriate information about the student's achievement or performance. Therefore reliability and validity are two critical issues in grading/scoring.

The history of grading dates back to the 1640s in early American universities such as Harvard and Yale, where examinations were used mainly for the purpose of degree awarding (Brookhart, 2004). Early researchers and educators (e.g., Dobbin & Smith, 1960; Finkelstein 1913) were primarily concerned with reliability of grading focusing their attention on pursuing commonly accepted reliable grading systems in the educational setting. Then with increasing dominions of grading in education, concerns and doubts about the adequacy and effectiveness of grading systems were expressed by teachers, educational administrators and researchers. In 1963, the Conference on College Grading Systems was held in Pennsylvania. Summarizing the conference discussion, Teaf (1964, p.88) concluded that "grades will not be abandoned—but their dominions are challenged." Later researchers such as Black and Wiliam (1998) echoed this conclusion, pointing out the overemphasized grading function of assessment at the expense of its learning ftmction in education.

More recently, school assessment reform, which is characterized by two approaches—the rise of large-scale assessment and changes in teachers' classroom assessment practices—has generated interest in research on teachers' grading practices (e.g., Brookhart, 2003; McMillan, 2003). This research has pointed out various challenges teachers have to face in grading to accommodate the classroom realities, internal policies, and large-scale testing. There is also the need for new theoretical development to accommodate assessment reform in school settings. For example, the growing interest in incorporating assessment in the classroom for the purpose of enhancing student learning, or assessment for learning, has made consequential aspect of grading a major concern. This has brought the validity issue to the fore in current research on grading.

Grading or scoring is even a more critical topic in language assessment due to the increasingly wide use of large scale highstakes language testing for various purposes, the complex nature of language and language use as well as the inherent subjectivity of the grading process in language assessment. Reliability of language performance assessment, which entails scoring by human

raters, has received substantial research attention in the past 15 years (e.g., Brown, 1995; McNamara, 1996). The last decade has also seen a significant shift in thinking about the role of classroom-based language assessment (Brindley, 2007). Research on assessment practices of teachers in different contexts such as first language (L]), second language (L2), international, national, secondary and postsecondary schools has begun to appear in language assessment literature.

The purpose of this study is to review empirical studies in the field of language assessment on grading or scoring published in major language-, education- and assessment-related journals from 1998 to present, While most of the reviewed studies are attempts to examine systematic effects on scoring of variables associated with teachers/raters for the purpose of score consistency and reliability, researchers have begun to pay increasing attention to teachers' grading practices in different contexts to explore the validity of classroom-based assessment. These two major research themes have followed different philosophical paradigms and have generally adopted different research methodologies. Discussions of these two themes in relation to two trends in language assessment—language performance assessment and assessment for learning—have significant implications for future research.

METHODOLOGY

Two substantial review articles serve as a baseline for this review: Black and Wiliam (1998) and McMillan & Workman (1998). Therefore, with a few exceptions, all of the articles to be covered in this review were published during or after 1998. Although I make no claim to exhaustion or completeness in this review, the themes I present emerged from strategies for systematic synthesis of primary research domains (Cooper & Hedges, 1994). The current review includes empirical research studies on grading and scoring in major language, education and assessment journals published in English between 1998 and 2009. The literature search was conducted by several means. One approach was to search by the combinations of scoring OR grading AND language assessment in the Google Scholar and ERIC databases; this is an inefficient approach because of a lack of terms used in a uniform way which define the field of grading in language assessment. This was supplemented with a second approach, the 'snowball' approach of following up the reference lists of articles found. Finally, the contents of all issues from 1998

to the present of the major journals in language, education and assessment were scanned. The final corpus included in the review contained 28 articles.

EFFECTS OF RATER CHARACTERISTICS

Grading is a complex decision-making process that involves different facets. Researchers have investigated the effects of various relevant factors on the grading or scoring process in language assessment. For example, in response to recent appearances of new assignment, test and scoring formats, researchers have examined the effects of factors such as typographic features (Hartley, Trueman, Betts, & Brodie, 2006); handwriting and print (Klein & Taub, 2005); and marking mediums e.g., paper-based marking and onscreen marking (Coniam, 2009). In oral language assessment research, Nakatsuhara (2008) examined the variability of interviewer behavior, its influence on the candidate's performance and raters' consequent perceptions of the candidate's ability on analytical rating scales. The results of this study clearly exemplified a possible relationship between the characteristics of interviewer behavior and ratings of particular components of language ability.

Rater effect has been a long-standing discussion in language assessment. Researchers are interested in investigating grade/ score inconsistency and biases that are attributable to variables associated with teachers/raters. For example, Fitzpatrick, Ercikan, Yen & Ferrara (1998) investigated the consistency of scores obtained from raters who had evaluated the same student work in different years in the Maryland School Performance Assessment Program and found that the groups of raters used in different years differed in severity, and particularly, the raters in the language arts areas were the least consistent. Shores and Weseley (2007) investigated effect of educators' political biases on their grading of student essays. Their experiment showed that essays that matched educators' self-reported political views received higher holistic grades than those that did not. This same relationship was also found when educators used a rubric, indicating that a rubric was not an effective tool in preventing grader bias.

However, mixed results were obtained by two studies investigating faculty's grading of students' writing. Roberts and Cimasko (2008) investigated the response of social science and engineering science faculty to a naturally occurring sample ofL2 writing, and found that there was a tendency across faculty to edit

semantic gaps as opposed to grammatical items. In a largescale study, Stern and Solomon (2006) analyzed faculty comments from 598 graded papers written for hundreds of courses from 30 different departments in Southern Illinois University. Results of this study indicated that most comments were technical corrections that addressed spelling, grammar, word choice, and missing words. Macro- and mid-level comments that addressed paper organization and quality of the ideas contained were surprisingly absent. Stern and Solomon (2006) noted that the lack of these larger idea and argument centered comments may prevent students from improving the quality of the larger issues in writing and refocus them on the smaller, albeit important, technical issues of writing.

Schoonen (2005) used a more sophisticated approach — structural equation modeling (SEM) —to estimate the variance components in the writing scores. In this study, eighty-nine grade 6 students wrote four essays, each of which was scored by five raters using two scoring methods (i.e., holistically and analytically) for two traits (i.e., content and organization, and language use). Analyses of these ratings showed that the generalizability of writing scores and the effects of raters and topics were very much dependent on the way the essays were scored and the trait that was scored. The overall conclusion of this study was that writing tasks contribute more to the score variance than raters do, suggesting the necessity of taking into consideration the interactions between rater characteristics and other facets when interpreting rater effects in grading.

Two variables associated with raters that have received most attention in research on grading of students' writing are raters' language background and experience. Hyland and Anan (2006) compared evaluations of the same piece of 150-word writing of a

Japanese EFL student by three groups of participants, a Japanese teacher group (JT), a group of native English speaking nonteachers (NES), and a group of native English speaking teachers (NST). The participants were asked to identify and correct errors in the writing. Drawing on descriptive data collected from an error identification and correction task and a questionnaire on the participants' error perceptions, the study showed that in spite of extraordinary agreement on an holistic evaluation (5 out of 10) among the three groups of participants, variability between them existed in terms of quantities and types of errors they identified as well as their error perceptions. It was also found that native English speakers tended to be more lenient in grading errors than non-native speakers, and non-native English speaking teachers tended to employ infringement of rules in judging error gravity far more than native English speakers, with the NES group stressing unintelligibility in their judgments. This exploratory study also points to the fact that the grading practice closely relates to the teaching experience of the rater.

Similar results were obtained by Porte (1999), who investigated teachers' reactions to the writing of non-native students. Participants in this study consisted of 14 native- speaker (NS) and 16 non-native speaker (NNS) university professors. They were required to indicate how many points they would deduct, on a holistic scoring guide, for each of the twenty erroneous sentences on a questionnaire. Results of t-tests showed significant differences between the NS and the NNS groups in terms of error toleration and the perceived gravity of specific errors. However, teachers in this study generally agreed in their judgments.

The relationship between teachers' evaluations of L2 writing and their years of teaching L2 writing was also evidenced by Shi, Wang and Wen's (2003) study. Forty-six English teachers (23 native English speakers and 23 non-native English speakers) from twenty-three tertiary institutions in Mainland China holistically evaluated ten essays written by Sinophone English majors and justified their scores for each essay with qualitative comments. Results showed that the most experienced writing teachers gave significantly lower scores than did the less or the least experienced writing teachers for four of the ten essays. Analyses of the qualitative comments on these four essays suggested that the

experienced teachers made either more negative or fewer positive comments on aspects such as general organization, language fluency, ideas and general language. However, one limitation of this study is that most of the least experienced teachers were non-native English speakers while the most experienced teachers were native speakers; differences in evaluation groups might result not only from their diverse teaching experiences but also from their differing Ll backgrounds, since raters' Ll background is also a variable that potentially affects the way raters assess L2 writing as evidenced by Hyland and Anan's (2006) and Porte's (1999) studies.

While the studies referred to above focused on classroom teachers' grading of students' writing from the classroom setting, Royal-Dawson and Baird (2009) addressed the question whether teaching experience was a necessary selection criterion for rating extended English questions in a large scale assessment context — England's Year 9 English National Curriculum Test. They compared scoring accuracy of teachers, trainee teachers and graduates of the same subject with experienced raters. In this study, fifty-seven raters with different backgrounds were trained in the normal manner and scored the same 97 students' work. Hierarchical linear models were set up to investigate whether there were significant mean differences in accuracy between groups. By comparing the scoring accuracy of graduates with a degree in English, teacher trainees, experienced teachers and experienced raters, the researchers concluded that teaching experience was not a necessary selection criterion.

RATERS' INTERPRETATIONS AND USE OF SCORING CRITERIA

Scoring criteria play a crucial role in the scoring process. A number of studies have investigated from a rater cognition perspective how scoring criteria channel raters' attention to different aspects of language performance and how different groups of raters perceive and use scoring criteria in different ways. For example, Xi (2007) summarized the advantages and disadvantages of using holistic and analytic rubrics for rating in oral tests, noting that holistic scoring promises efficiency in scoring and is likely to impose a lesser cognitive load on raters but raters may weight the components in the holistic rubrics differentially depending on their background and experience and

their perceptions of how a particular weakness or strength impacts the overall communicative quality in a particular assessment context. Meanwhile, in holistic scoring, there are often few explicit rules that raters can utilize when making a global judgment. In contrast, analytic scoring makes possible a systematic way for raters to weight different dimensions of the rubrics but may pose higher cognitive load on raters and produce potential rating inconsistencles.

Lumley (2002) examined the process by which raters of texts written by English as a second language (ESL) learners made their scoring decisions using an analytic rating scale in a large scale assessment context in Australia—the Special Test of English Proficiency. By analyzing raters' think-aloud protocols describing the rating process as they rated the texts, Lumley was able to demonstrate the sequence of rating, the interpretations the raters made of the scoring categories in the analytic rating scale, and the difficulties raters faced in rating. It was shown that on the one hand, raters tried to remain close to the rating scale; but on the other hand, they were heavily influenced by the complex intuitive impression of the text obtained when they first read it. This set up a tension between the rules and the intuitive impression, and raters developed various strategies to help them cope with problematic aspects of the rating process. Findings of this study also showed that raters sometimes applied the contents of the scale in quite different ways, giving different emphases to the various components of the scale descriptors. Therefore, Lumley concluded that it is the rater that is at the center of the scoring process.

The diversity of raters' interpretations of rating criteria is also revealed by Eckes (2008), who investigated raters' interpretation of scoring criteria of the writing section of the Test of German as a Foreign Language (Test Deutsch als Fremdsprache, TestDaF). Employing many-facet Rasch analysis and a twomode clustering technique, Eckes was able to show that raters differed significantly in their views on the importance of the various criteria and that raters were far from dividing their attention evenly among the set of criteria. Six rater types were identified, each characterized by a distinct scoring profile. Moreover, findings of this study revealed that rater background variables partially accounted for the scoring profile differences.

A number of studies have investigated raters' decisionmaking behavior in the grading process using more sophisticated research methods. For example, Decarlo (2005) presented an approach to essay grading based on signal detection theory (SDT), which provided a theory of psychological processes underlying the raters' behavior. Latent class SDT was applied to essay grading in this study, and was compared with item response theory (IRT). Findings of this study revealed that validity coefficients were about equal in magnitude across SDT and IRT models.

Myford and Wolfe (2009) proposed a framework for monitoring rater performance over time, presenting several statistical indices to identify raters whose standards drifted. They analyzed rating data from the 2002 Advanced Placement English Literature and Composition Examination, employing a multifaceted Rasch approach to determine whether raters exhibited evidence of two types of differential rater functioning over time (i.e., changes in levels of accuracy or scale category use). Results showed that some raters showed statistically significant changes in their levels of accuracy as the scoring progressed, while other raters displayed evidence of differential scale category use over time.

Based on three coordinated, exploratory studies, Cumming, Kantor, and Powers (2002) developed a framework to describe the decisions that experienced writing assessors made when evaluating English as a second/foreign language (EFL) written compositions. Their findings revealed that raters attended more extensively to rhetoric and ideas (compared to language) in compositions they scored high than in compositions they scored low. They also found that ESL/EFL raters attended more extensively to language than to rhetoric and ideas overall, whereas the English-mother-tongue raters balanced more evenly their attention to these main features of the written compositions. Meanwhile, most participants in these studies perceived that their previous experiences of rating compositions and teaching English had influenced their criteria and their processes of rating the compositions.

While the majority of the studies referred to above have mainly focused on the effects on the rater's decision-making process of various characteristics associated with raters as well as the interactions between rater characteristics and other relevant factors of the micro context of grading or scoring, other studies on teachers' grading practices have been situated in the macro social, cultural, economic and political contexts. For example, Hunter, Mayenga and Gambell (2006) analyzed from an anthropological perspective pan Canadian data from a 2002 English teacher questionnaire (N— 4070) about self-reported assessment practices in terms of tool choice and use by secondary teachers of different experience and qualification levels. Four underlying variables were identified in the teachers' choice of assessment tools: whether affective traits such as attendance, effort, motivation or participation were factors; whether self-assessment and peer evaluation were considered; whether portfolios or examples of student work were variables in grading practices; and whether multiple choice or short response tasks were chosen. In terms of tool use, the three salient variables were: the nature of the feedback cycle with students; whether homework contributed to grades; and whether homework served in large group instruction. A number of significant differences by career stage and credential level were revealed in assessment instrument choice and use in the Canadian context.

In the United States, Zoeckler (2007) conducted a case study in a high school to examine how English teachers arrived at a fair grade while weighting both achievement and nonachievement factors and the role of teachers' expectations. In this study a theoretical framework was used which considers grading processes in terms of truth, worthwhileness, trust, and intellectual and moral attentiveness. Results indicated that grading was influenced by grading systems, perceptions of effort, and concern for moral development and that these teachers struggled with issues of fairness, but were confident that their grades communicated the messages they hoped to send.

Two studies compared teachers' grading practices in different contexts. Cheng and colleagues (2004, 2007) conducted a 3year study comparing ESL/EFL teacher assessment practices in three

different tertiary institutional contexts: Canada, Hong Kong and China. Their comparative survey findings (reported in Cheng, Rogers & Hu, 2004) demonstrated the range of procedures that teachers reported using when making decisions about their students' language abilities and the complex and multifaceted roles assessment played in these settings. Their interview findings (reported in Cheng & Wang, 2007) demonstrated how teachers made day-to-day decisions in these three contexts, thus adding to our understanding of ESL/EFL classroom assessment practices. Results of their study showed both commonality and diversity among the teachers across the three contexts. Cheng and Wang identified a range of contextual factors that may help to account for the differences that emerged in their study across these settings. For example, they analyzed teachers' preferences for different methods of marking in terms of 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.

Cheng and Wang's (2007) study provides substantial documentation and comparison of teacher grading practices in different settings, showing how teachers' assessment practices were influenced by their belief systems and the external context. However, their investigations were made on the basis of teachers' self -reported data. Noting limitations of this research methodology, the researchers suggested the need for future studies of assessment practices to include both teachers' perceptions and their actions in the classroom, including observations of teaching as well as examination of curriculum documents and teaching materials related to assessment and evaluation.

In a similar vein, Davison (2004) compared ESL teacher assessment practices in Australian and Hong Kong secondary schools, focusing on underlying constructs and criteria teachers used to grade students' written arguments in these two contexts. Adopting a sociocultural approach, the study highlighted how teacher's decision-making process was shaped and constrained by the different assessment cultures of Australia and Hong Kong. The methodology chosen for this study was primarily qualitative and interpretive: questionnaires, verbal protocols, individual and group interviews and self-reports were used to explore 12 teachers' assessment beliefs, attitudes and practices in each of

these two contexts. A number of contrasts were found in these aspects. Davison analyzed these contrasts in relation to the social, cultural and educational differences between these two contexts, and concluded that traditional notions of validity may need to be reconceptualized in high-stakes teacher-based assessment, with professional judgment, interaction and trust given much higher priority in the assessment process.

Two other studies, both from Australia, highlighted the conflict and tension between teachers' classroom assessment practices and the state or provincial accountability system. Arkoudis and O'Loughlin (2004) reported on a collaborative study involving ESL teachers in an Australian English Language Centre grading the students' written work using the ESL companion, a companion volume to the English Curriculum standards framework in Victoria, Australia, as an assessment tool. This study focused on conflicts that emerged in the teachers' grading of the students' work. The teachers were required to demonstrate the students' progress in a consistent way using the ESL companion, but they did not have a shared perspective on what they were expecting students to be able to do at each of the different levels in the ESL companion because the assessment criteria were too general and vague and did not link with the teachers' experiential understandings of their students' progress. In order to align the assessment standards with reference to their own pedagogical practice, the teachers reworked the ESL companion, modifying it based on their own pedagogic understandings about ESL teaching and learning. But the teachers' re-working of the descriptors was not endorsed by the Department of Education on the grounds that it did not have a sufficiently firm theoretical basis. Arkoudis and O'Loughlin (2004) analyzed these conflicts using the positioning theory, in which power is a key feature and conversations are considered as complex interactions between partners with different rights and responsibilities. Findings of their study showed that when teachers made their grading decisions they found themselves at the confluence of different assessment cultures and faced with significant dilemmas in their assessment practices.

In the other study, Cooksey, Freebody and Wyatt-Smith (2007) highlighted the tension between "system" validity (the accountability of teachers' assessments to the parameters of state

or national educational systems) and "site" validity (the accountability of teacher assessments within their localized classroom and community contexts) and the complexities inherent in the teacher's responsibility to make professional judgment accountable to stakeholders with varying interests and needs. They took a "snapshot" of how teachers had resolved such tensions and complexities by analyzing teachers' judgments of students' written texts, documenting how 20 primary school teachers from the Brisbane (Queensland, Australia) metropolitan area used evidence in ways that depended both on their knowledge of the students and on the assessment framework they needed to use. They analyzed teachers' judgments by contrasting the structures of assessments made using teachers' normal classroom judgment processes with those made using an external set of "benchmark" standards. Their conclusions were that current understandings of teacher judgment processes generally fail to account for the complexity and dynamism of this routine classroom activity. Cooksey et al. (2007) also suggested using judgment analysis, combined with think-aloud protocols to understand the complexities associated with the operation of judgment in educational assessment.

Feinberg and Shapiro (2009) also noted the importance of understanding external influences when considering teacher reports of student performance, but they took a very different perspective from that of Davison (2004) and Cooksey et al. (2007). In Feinberg and Shapiro's study 74 teachers were asked to predict average- and low-performing students' reading performance using a rating scale and actual curriculum-based measures of oral reading fluency. Teachers' judgments of students' performance on curriculum-based measures and a standardized achievement measure of reading were compared. Results of this study showed that correlations between predicted teacher scores and actual student performance in general ranged from moderate to strong, but a how-close analysis revealed that the score difference was larger on the part of the low-performing students. Feinberg and Shapiro concluded that teachers had lower levels of accuracy for lower achieving students.

An earlier study that took social factors into consideration was conducted by Wyatt-Smith (1999). The study focused on how teachers read student writing in the context of criteria-based

assessment in secondary schools in Queensland, Australia. The researcher used the term 'reading' to refer to teachers' responding to students' writing and assessing it for grading purposes. She was interested in seeking to interpret what individual teachers had to say about their experiences as reader-assessors in terms of their own conceptions while also taking social interactions between teacher and student into account. Based on the analysis of data on three case-study teachers, Wyatt-Smith proposed a databased model of how teacher readings of student writing occur. This model includes four units: (a) individual teachers' philosophies of teaching writing coupled with their repertoire of language skills, reading strategies and prior readings, etc.; (b) attitudes to and purposes of reading in conjunction with teacher beliefs about status or position in the school, and other beliefs about classroom social practices; (c) available knowledge 'files'; and (d) the teacher's attempts at actively reconstituting conceptions of quality.

Two LINES OF GRADING RESEARCH AND THEIR IMPLICATIONS

The studies referred to in this review show the continued efforts in the past decade to address various issues associated with grading or scoring in language assessment, particularly the issues of reliability and validity. These studies have generally followed two lines of research. One line follows a research agenda set out by language performance assessment models (e.g., McNamara, 1996; Skehan, 1998), focusing on examining the systematic effects on scores of a host of factors that are deemed relevant to the scoring/grading process. Two-thirds of the reviewed studies have followed this line. Following the measurement paradigm, most of these studies have taken a positivist approach and the methodologies used in them are primarily quantitative. This review shows that while rater effect remains a major topic in grading or scoring research (e.g., Shi et al., 2003; Hyland & Anan, 2006; Royal-Dawson & Baird, 2009), researchers have also focused interest on exploring the interactions between rater characteristics and other facets of language performance assessment in the grading or scoring process such as the rating scale (Eckes, 2008) and the task (Schoonen, 2005). There have also been calls for further investigation of raters' cognitive behavior in the grading or scoring process (Cumming et al., 2002). More sophisticated research methods have been employed in these

studies such as SEM (Schoonen, 2005) and multifaceted Rasch approach (Myford & Wolfe, 2009). Findings of these studies have significant implications for assessing the quality of large-scale rater-mediated language testing, rater monitoring, and rater training and for improving the reliability of language performance assessment through a combination of rater training, better specification of scoring criteria and better task design.

The other line of research interprets grading or scoring as a professional decision-making process, focusing on understanding teacher-raters' grading or scoring practices in relation to broader educational, social and cultural contexts. Studies in this line of research have followed an interpretivism paradigm and taken different theoretical approaches such as anthropology (Hunter et al., 2006), positioning theory (Arkoudis & O'Loughlin, 2004) and sociocultural theory (Davison, 2004), and employed primarily qualitative research methods, e.g., case study (Wyatt-Smith, 1999; Zoeckler, 2007); interview (Cheng et al., 2007); and/or mixed-methods (Cooksey et al., 2007). Findings of these studies suggest that grading is never context-free; nor is it a purely technical activity. Teachers' beliefs, knowledge, theories, assumptions, and attitudes, as well as the social, cultural and educational contexts all play significant roles in shaping teachers' grading decisions. In the end, the grade assigned by the teacher is characterized by interactions between these various factors rather than a direct representation of the student's language proficiency. Therefore, understanding interactions between these factors should be a necessary constituent of validation research.

FUTURE RESEARCH

The two research lines revealed by the studies in this literature review are relevant to two trends in educational assessment in general and in language assessment in particular. First, the past 15 years has seen the widespread use of performance assessment as a measure of language proficiency particularly in terms of writing and speaking. From the perspective of grading or scoring, performance assessment is a double-edged sword. Professional judgments of skilled human raters can provide richly informed interpretations of the value and worth of students' language use abilities. On the other hand, however, since ratings in performance

assessment are necessarily subjective, the lack of reliability achievable in assessment is a major concern. This literature review shows continued efforts in the last decade in language assessment research to examine systematic effects on scoring of variables other than the test taker's language proficiency that are considered potentially relevant to the scoring process. The research in this area is far from complete. In the future, researchers will continue their efforts to identify, conceptualize and model variables associated with rater variability. Rater behavior will receive increasing attention and will be explored from different approaches using more sophisticated research methods.

In light of the measurement paradigm, the key issue associated with scoring in performance assessment is score consistency, which is often addressed by means of rater training at the technical level because grading is considered exclusively at the judgment level as purely a standards-referenced technical process (Newton, 2007). However, as suggested by Lumley's (2002) study, this kind of training may just result in raters' surface-level processing of rating criteria. One solution to this problem is to inform rater training by further research on the rating practice in relation to raters' belief and value systems. Research in this area will also shed new light on examining the validity of language performance assessment, which has not been sufficiently addressed up to the present day.

Second, the last decade has seen a significant shift in thinking about the role of classroom assessment in both language learning programs (e.g., Brindley, 2007) and in educational settings in general (e.g., Andrade, 2009). The notion of Assessment for Learning (AfL) has gained increasing popularity and research has offered ample evidence in support of Afl as a promising pedagogic approach (e.g., Hume & Coll, 2009; Kirton, Hallam, Peffers, Robertson, & Stobart, 2007; Marshall & Drummond, 2006). However, there is also a wealth of research evidence that teachers' everyday assessment practices are beset with problems and shortcomings, or what Black and Wiliam (1998) called in their seminal review of classroom assessment "a poverty of practice." Among these problems is the grading practice that does not provide accurate and useful information about learners' progress and achievement and how their work can be improved. Studies have also reported results showing a contradiction between

teachers' grading practices and the recommended practices that measurement specialists have offered in pre-service and inservice teacher education. For example, teachers tend to consider a hodgepodge of factors when assigning grades (Cross & Frary, 1996) in spite of the suggestion that grades should be based on students' academic achievement without including confounding factors such as effort and work habits (Merwin, 1989).

A critical challenge in promoting AfL is the lack of effective models for teacher professional development (PD) on assessment (Lee & Wiliam, 2005). The majority of current research on teacher PD recommends a shift away from top-down models that are disconnected from teacher practice and calls for models of PD that focus on teachers as active learners (Atherton, 2009) while promoting contextualized learning (Sparks & Hirsh, 1997) and reflective practice (Osterman & Kottkamp, 2004). In light of the current teacher PD models, a best approach to solving the discrepancy between the recommended and the actual grading practices is to understand how teachers make their grading decisions in particular social, cultural and educational contexts. Davison's (2004) and Cheng and Wang's (2007) studies referred to in this review provided substantial documentation and comparison of teachers' grading practices in different settings in relation to teachers' belief systems and external contexts. However, the factors that determine teachers' grading practices and the assessment methods they use to assign grades are still unknown. Future research is needed in language assessment to investigate teachers' grading practices within the framework of validity.

Over history, grading has long been "a necessary evil" or "a black box" in education. Studies in the past decade in the field of language assessment, particularly those conducted from the sociocultural perspective indicate that we might not see it this way. Instead, we might see it as a kaleidoscope, or a dynamic landscape of sense and relationship, where teachers play a particular role as participating members in particular sociocultural contexts. To explore this landscape, researchers need to be equipped with what Bourdieu and Wacquant (1992) called "doublefocused analytical lens"; that is, an analysis of the regularities of the field and an analysis of agents' internalizations

of these regularities through their reflexive deliberations and actions.

REFERENCES

Andrade, H. (2009). This issue. Theory into Practice, 48(1), 1-3.

Arkoudis, S., & O'Loughlin, K. (2004). Tensions between validity and outcomes: Teacher assessment of written work of recently arrived immigrant ESL students. Language Testing, 21 , 284-304.

Atherton, J. S. (2009). Learning and teaching: Reflection and reflective practice. Retrieved April 10, 2010 from http:// www.learningandteaching.info/learning/reflecti.htm.

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

Bourdieu, P. , & Wacquant, L. (1992). An invitation to reflexive sociology. Cambridge: Polity Press.

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

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. Educational Measurement: Issues and Practice, 22(4), 5-25.

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

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. Language Testing, 12(1), 1-15.

Cheng, L., Rogers, T & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: purposes, methods, and procedures. Language Testing, 21, 360-389.

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

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. Educational Research and Evaluation: An International Journal on Theory and Practice, 15(3), 243-263.

Cooksey, R., Freebody, P., & Wyatt-smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate

students' writing. Educational Research and Evaluation, 13(5), 401-434.

Cooper, H. M. & Hedges, L. V. (1994). The handbook of research synthesis. New York: Russell Sage Foundation Publications.

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.

Cumming, A, Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. The Modern Language Journal, 86(1), 67-96.

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

Decarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. Journal of Educational Measurement, 42, 53-76.

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. Language Testing, 25, 155 - 185.

Feinberg, A. B. & Shapiro, E.S. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. The Journal ofEducation-al Research, 102, 453-462.

Finkelstein, I. E. (1913). The marking system in theory and practice. Baltimore: Warwick & York, Inc.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. Applied Measurement in Education, 11(2), 195 _208.

Hartley, J., Tnteman, M., Betts, L., & Brodie, L. (2006). What price presentation? The effects of typographic variables on essay grades. Assessment & Evaluation in Higher Education, 31(5), 523-534.

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.
Hunter, D., Mayenga, C., & Gambell, T. (2006). Classroom assessment tools and uses: Canadian English teachers' practices for writing. Assessing Writing, 11(1), 42-65.
Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. System, 34(4), 509-519.
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.
Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. Assessing Writing, 10(2), 134-148.
Lee, C. , & Wiliam, D. (2005). Studying changes in the practice of two teachers developing assessment for learning. Teacher Development, 9(2), 265-283.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19, 246-276.
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. (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. & Workman, D. J. (1998). Classroom assessment and grading practices: A review of the literature. Richmond, VA: Metropolitan Educational Research Consortium.
McNamara, T. F. (1996). Measuring second language performance. London: Longman.
Menvin, J. C. (1989). Evaluation. In M.C. Reynolds, (Ed.), Knowledge base for the beginning teacher (pp. 185-192). Oxford: Pergamon Press.
Myford, C. M. & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. Journal ofEducational Measurement, 46(4), 371-389.

Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. ELTJourna1, 62(3), 266-275.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. Assessment in Education: Principles, policy & practice, 14, 149-170.

Osterman, K. F. & Kottkamp, R. B. (2004). Reflective practice for educators (2nd ed.). Thousand Oaks, CA: Corwin Press.

Porte, G. (1999). Where to draw the red line: Error toleration of native and non-native EFL faculty. Foreign Language AnnaIS, 32 426—434.

Roberts, F. & Cimasko, T (2008). Evaluating ESL: Making sense of university professors' responses to second language writing. Journal of Second Language Writing, 17(3), 125143.

Royal-Dawson, L. & Baird, J.A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? Educational Measurement: Issues and Practice, 28(2), 2-8.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. Language Testing, 22, 1-30.

Shi, L., Wang, W., & wen, Q. (2003). Teaching experience and evaluation of second-language students' writing. The Canadian Journal ofApplied Linguistics, 6, 219-236.

Shores, M. , & Weseley, A. (2007). When the A is for agreement: Factors that affect educators' evaluations of student essays. Action in Teacher Education, 29(3), 4-11.

Skehan, P. (1998). A cognitive approach to language learning. Oxford: Oxford University Press.

Sparks, D. & Hirsh, S. (1997). A new vision for staff development. Oxford, OH: National Staff Development Council.

Stern, L. , & Solomon, A. (2006). Effective faculty feedback: The road less traveled. Assessing Writing, Il (l), 22-41.

Teaf, H. M. (1964). Grades: Their dominion is challenged. The Journal ofHigher Education, 35(2), 87-88

Wyatt-Smith, C. (1999). Reading for assessment: How teachers ascribe meaning and value to student writing. Assessment in Education: Principles, Policy & Practice, 6(2), 195—223.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL@ Academic Speaking Test (TAST) for operational use. Language Testing, 24, 251-286.

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