Journal of Teaching Language Skills (JTLS)
38(4), Winter 2020, pp. 123-164- ISSN: 2008-8191
DOI: 10.22099/jtls.2020.36293.2780

Exploring the Factors Iranian EFL Institute Teachers Consider in Grading Using Personal Construct Theory

Majid Nowruzi [*] **Majid Amerian** [**]

Abstract

Although grades are the most ubiquitous currency of educational measurement around the globe, their meaning, particularly in understudied settings as in Iran, is still shrouded in mystery. The purpose of this study was to investigate EFL teachers' class grades by focusing on the less verbalized aspects of grading to see what a grade means. Five Iranian English language teachers working part-time in a private EFL institute were interviewed using the repertory grid interviewing technique, Kelly's (1955) unique data collection instrument used extensively in personal construct theory (PCT). The results of the content analysis revealed that of the 92 elicited constructs, over 70% were categorized as non-academic, pointing to a heavy reliance on such criteria for grading, and consequently leading to the invalidity of assigned grades. Further, the results of principal component analysis (PCA) of each teacher's elicited constructs endorsed hodgepodge grading by demonstrating single main components that accounted for the most variation in teacher grading and that comprised both academic and non-academic factors. However, this phenomenon was interpreted slightly differently when seen from the PCT perspective. Implications of this study for teacher professional development and teacher education programs are discussed.

Keywords: Classroom assessment, Grading, Repertory grid technique, Personal construct theory, Non-academic factors

Received: 03/03/2020 Accepted: 18/06/2020

[*] Ph.D. Candidate, Department of English language, Faculty of Foreign Languages, Arak University, Arak, Iran - Email: m-nowruzi@phd.araku.ac.ir, Corresponding author
[**] Assoicate Professor, Department of English language, Faculty of Foreign Languages, Arak University, Arak, Iran - Email: m-amerian@araku.ac.ir

Assessment is a powerful determinant of legitimate knowledge (Bernstein, 2003; Hay & Penney, 2013). Grades or marks are characterized as the symbols (mostly letters or numbers) that teachers assign to students' works and performances (Brookhart, 2004) or to their overall performances summarized on students' report cards (Brookhart et at., 2016; O'Connor, 2009). Teachers' judgments of their pupils' performances, whether academic or non-academic, are considered as grades as soon as they appear in grade sheets and count toward students' final grades. Simply put, every number or letter or any other symbol identifiable by the stakeholders that is used to represent a student's performance in the context of the classroom and form the basis for making instructional decisions is known as a grade. As long as such subjective judgments result in assigning grades that do not represent the only achievement, the validity of teachers' grades and the decisions made based on such grades are called into question (Allen, 2005). In the Iranian educational setting, it seems that numerical grades, either on a 0-20 scale or on a percentage (0-100) scale, are more commonly used by EFL teachers as evidenced by grades in teachers' grade sheets or those printed on students' report cards.

Grading, as the most predominant and ubiquitous aspect of classroom assessment (CA), has long been surrounded by controversy in research communities. While Newton (2007) defines it as a judgmental and technical process of determining an evaluative mark based on pre-determined performance standards with no decision-making, Sun and Cheng (2013) describe grading as a complex process of decision-making that necessitates teachers to make subjective value judgments concerning student achievement, improvement, and learning, in addition to considering specific grading criteria. This discrepancy over grade meaning makes Linn and Miller (2005) acknowledge that grading is "one of the more frustrating aspects of teaching" (p. 366). Whereas assessment fosters teaching and learning, teachers view it

as a challenging concept due to its mathematical and statistical nature (Shah Ahmadi & Ketabi, 2019).

Numerous empirical studies have shown that teachers use an array of non-cognitive pieces of evidence of achievements such as student effort and ability in determining students' grades in addition to cognitive criteria (Brookhart et al., 2016; Guskey, 2011; McMillan & Nash, 2000; Randall & Engelhard, 2009, 2010; Russell & Austin, 2010; Sun & Cheng, 2013; Svennberg et al., 2014; Yesbeck, 2011). Consequently, the outcome of teachers' grading appears to be "a hodgepodge grade of attitude, effort, and achievement" (Brookhart, 1991, p. 36), leading to confusion when it comes to grade interpretation and use. McMillan (2008) acknowledges that one of the most challenging issues in grading is dealing with such non-achievement factors as effort, improvement, and ability or what he refers to as academic enablers (McMillan, 2001). Additionally, teachers teaching various subject matters consider non-achievement criteria such as attitude and behavior in assigning grades (Brookhart, 1994; Brookhart, 2013; Cizek et al., 1996; Cross & Frary, 1999; Duncan & Noonan, 2007; McMillan, 2003). Likewise, Guskey and Link (2018) verified that, in addition to using different pieces of evidence of student learning in grading, teachers use non-cognitive factors mostly related to student behavior for determining grades across all grade levels. It appears that teachers' grading and assessment practices are influenced by their beliefs, values, and contextual factors, along with formal grading criteria (Chang & Wang, 2007; Davison, 2004).

Brookhart (2004) contends that "the primary purpose of grading for both individual assignments and report cards should be to communicate with students and parents about their achievement of learning goals" (p. 5). Achievement should be the only factor represented by grades. Pre-service and in-service teachers are recommended to base their grading solely on academic achievement (Dyrness & Dyrness, 2008; McMillan, 2008; Merwin, 1989; O'Connor, 2007; Wormeli, 2006), but in practice, grades represent students'

knowledge and skills, coupled with other criteria such as their attitude, class attendance, and motivation (Cox, 2011; Klapp Lekholm & Cliffordson, 2009; Young, 2011). Additionally, teachers' grading also includes less verbalized or internalized factors or what are referred to as gut feelings (Annerstedt & Larsson, 2010; Hay & MacDonald, 2008; Svennberg et al., 2014) that are hard to identify using conventional survey methods. In effect, when grading criteria are internalized, their transparency is called into question, and, as a result, validity, reliability, and fairness of students' grades will be at risk (Annerstedt & Larsson, 2010; Svennberg et al., 2014). Researchers, teachers, and assessment experts have frequently expressed doubts about the accuracy and efficiency of various grading methods (Black & Wiliam, 1998; Guskey & Bailey, 2001; Teaf, 1964). The mixing of factors done by teachers in determining grades confounds the interpretability of grades by introducing construct-irrelevant variance (Brookhart, 1991, 1993). Due to the relationship between grades reported in students' report cards and educational decisions made based on them, variations in teachers' grading adversely influence students' future academic success (Guskey, 2015; Link, 2018).

This study aims to explore the less verbalized factors that five Iranian EFL institute teachers consider when determining students' grades using the repertory grid (RG) interviewing technique, Kelly's (1955) unique data collection tool in PCT. The findings can help clarify the tacit knowledge that teachers bring with them into their grading practices as part of their grading decision-making. To date, not many studies have employed a theoretical lens to look into English language teachers' internalized grading criteria. The majority of the studies concerning teachers' classroom assessment (CA) and grading reviewed have been empirical in nature (Brookhart, 1993; Cheng & Sun, 2015; Cheng & Wang, 2007; McMillan, 2001; McMillan et al., 2002), mostly relying on data obtained from self-report questionnaires or interviews that might, more frequently, be affected by the social desirability phenomenon than repertory grid interviews. Besides, almost no studies have been

conducted on Iranian EFL teachers' classroom grading practices as evidenced by the results of online searches in ERIC and PsycInfo, leaving these teachers' voices mostly unheard. Therefore, it was hoped that the findings of this study could help create an in-depth analysis of grading as an understudied area of research. To address the issues mentioned earlier, this study seeks to answer the following questions:

1. What factors do Iranian English language teachers consider when assigning grades?
2. Do teachers assign a hodgepodge grade? If yes, how is the hodgepodge grading viewed in PCT?

Theoretical Framework

George Kelly, the developer of personal construct theory of human understanding, claims that reality and our perceptions of it are made up of contrasts rather than absolutes (Fransella, Bell, & Bannister, 2004). According to Kelly (1955), human behavior can be understood through a limited number of dichotomous mental constructs that (s)he has personally created on how the world around him/her functions. Two points in this conceptualization of constructs relate to teachers' grading practices: (a) that constructs are created individually in teachers' minds and originate from their experiences with their surroundings, and (b) that constructs shape teacher behavior and action. Kelly (1991) states that "Man looks at his world through transparent patterns or templates which he creates and then attempts to fit over the realities of which the world is composed" (pp. 8-9). Additionally, Kelly (1991) believes that one's constructs can be "explicitly formulated or implicitly acted out, verbally expressed or utterly inarticulate, consistent with other courses of behaviour or inconsistent with them, intellectually reasoned or vegetatively sensed" (p. 7).

The epistemological position underlying Kelly's constructive alternativism is that each person forms his/her theories about making sense of the world and its different phenomena and is his/her own scientist. In doing

so, there is no one single truth out there as opposed to positivists' claims (Jankowicz, 2003). Due to this epistemological discrepancy between positivism and constructivism, differing outcomes are quite probable. According to PCT, behind every single judgment and subsequent decision (either conscious or unconscious) based on that judgment lies one's implicit theory about that event, and the repertory grid technique is the right means of exploring the organization and content of that implicitness. This maxim gains momentum in the present study because various researchers and measurement specialists converge on seeing grading as a judgmental professional decision-making process (Brookhart et al., 2016; Cheng & Sun, 2015; Guskey & Link, 2018). In accordance with the PCT, grading decisions could most likely engage a host of underlying implicit operations done by teachers about the (in)appropriateness of the various criteria they consider in grading.

Constructs, on which the theoretical foundations of PCT are established, are mental signposts that direct our actions and behavior without needing to be externally expressed (Bjorklund, 2008). As with grading, such mental signposts direct teachers' grading practices without needing to be verbalized. Kelly (1955) contends that our constructs are continually being revised, replaced, or subsumed based on our new experiences. He initiated the repertory grid technique and found that constructs facilitate the prediction of the future courses of action. As a result, exploring a teacher's grading constructs could facilitate the prediction of his/her future grading decision-making practices more reliably, which in turn could probably help teachers themselves diagnose what is right or wrong with the grades they assign, leading to grading validity improvements.

The repertory grid technique is one way to make tacit internal knowledge explicit (Bjorklund, 2008). In effect, grids are used "for arriving at straightforward descriptions of how a person views the world, or some smaller part of it, in his or her own terms" (Jankowicz, 2003, p. 8). Specifically, the benefit of using the RG technique is to help participants express their

internalized constructs by making comparisons and showing similarities and differences when giving a description or definition is difficult for them (Bjorklund, 2008). Claims about the reliability and validity of the RG technique are hard to make because the method has no standard form and can be applied to various areas of inquiry. Additionally, it is used in this study and other similar studies (e.g., Svennberg et al., 2014, 2016) as a technique rather than a method, and subsequently, its validity lies in the uses to which it is put. Concerning the validity of repertory grids, Fransella et al. (2004) contend that grid validity should be conceived "in a very different way from that in which we talk about the validity of, say, a questionnaire" (p. 144). However, as they later argue, this technique has proven to be highly valid for behavior prediction and group differentiation.

Another advantage of the RG interviewing is that through using this technique, the risks of directing the interview by the interviewer and his/her questions are minimized because once the interviewees know the procedure, they will systematically compare elements and generate and rate constructs with the least intervention from the interviewer. However, for the RG technique to be efficient, it is necessary that teachers possess good familiarity with elements (students in this case) and that elements represent the full range of students at different ability levels in that specific area of inquiry, i.e. the most extensive grade range (Fransella et al., 2004; Kelly, 1955).

Concerning the underpinnings of PCT, it can be stated that looking into the grading phenomenon from the PCT perspective via its unique data collection technique could probably yield a more profound insight into grading than what is expected to come out from regular interviews. This understanding can be achieved by directly eliciting tacit grading constructs, without imposing predetermined frameworks of thinking or performance on teachers' thought processes or contaminating their mental manifestations as might be the case with survey studies that employ questionnaires for data collection.

Literature Review

Literature has revealed significant evidence concerning the factors that teachers use to determine grades (e.g., Brookhart, 1991, 1993; Cheng & Sun, 2015; Cizek, et al., 1996; Cross & Frary, 1999; Duncan & Noonan, 2007; Guskey & Link, 2018). Although most of such research has been practical in nature, valuable grading-related findings have been reported to date as to what influences teachers' grading decision-making the most substantially. Throughout the last quarter of a century, numerous studies have centered on the meanings the grades communicate to stakeholders or the validity of teachers' classroom assessment and grading practices (e.g., Randall & Engelhard, 2010; Sun & Cheng, 2013), as the most important concept in educational measurement. Teachers were found to use nonacademic factors or what McMillan (2001, p. 25) referred to as "academic enabling behaviors or traits" extensively when assigning grades; consequently, putting the validity and the interpretation of teachers' grades at serious risk. Although academic achievement was regarded as the key determinant of teachers' grades, the amalgamation of nonacademic criteria appeared to be concerning. Numerous studies also pointed to the existence of great variability in teachers' use of such academic and nonacademic factors, even when grading policies were prescribed that called for grading consistency. A review of a number of these studies may help set the stage for the implementation of the present study.

As a preliminary study of the validity of teachers' grades, Brookhart (1993) studied 84 teachers for the meanings and value judgments they perceived from grades. The results revealed that teachers relied heavily on effort as a grading criterion and students whose performance was below average but made an effort received a passing score, while average and above average students got their own grades. Also, it was found that teachers tried to be fair in grading and stated that grades were seen as a kind of payment for student work, assigned for completed work rather than for academic achievement, an indication that effort was a significant constituent in

assigning a grade. It was also pointed out that because teachers should consider student motivation, personality-related issues, and the consequences of grades assigned, grading based merely on achievement is a rarity.

Later, in an effort to explore the purposes of grading, Guskey (1996) categorized sources of evidence for grading into three kinds: (a) *product factors* of achievement that show what students currently know and can do, also labeled achievement factors in other studies, (b) *process factors* such as homework completion or participation in class that support and enhance learning, and (c) *progress factors* that highlight student improvement in approaching learning goals. This classification could be regarded as a preliminary systematic formulation of grading criteria endorsed by Brookhart's (1993) study.

Few years later, McMillan and Nash (2000) proposed a model of teachers' grading decision-making consisting of internal and external factors by interviewing 24 teachers. The most salient internal factors included teachers' philosophy of teaching and learning along with their beliefs and values, whereas the most significant external factors consisted of nationwide standards of learning, high-stakes tests, grading policies, and parental pressure for more accountability. Their model has been supported by other studies (e.g., Cheng, Rogers, & Hu, 2004; Cheng, Rogers, & Wang, 2008; Cheng & Wang, 2007) that investigated English language teachers' grading practices in tertiary schools in China, Canada, and Hong Kong. The results of these studies, consistent with what McMillan and Nash (2000) had proposed, indicated that teachers were influenced by their own beliefs about assessment, their purposes for assessment, their teaching experiences and assessment literacy, and other contextual factors specific to each setting such as class size and the weight attached to high-stakes testing. They concluded that these factors, along with classroom realities, greatly influenced students' grades. For instance, most Canadian and Chinese teachers in Cheng and Wang's (2007) study commented that they prepared their grading criteria using a wide variety of

sources such as their colleagues' views. The researchers suggested that such differences may pertain to teachers' personal beliefs and values about what counts the most in the specific setting where grading occurs. Their understanding was regarded as an endorsement of Brookhart's (1993) findings in claiming that solely-achievement-based grading was quite rare.

In a later study, Randall and Engelhard (2010) explored the meaning of grades with a focus on borderline cases. They collected data from 516 public school teachers using a 53-item survey and found that teachers tried to stick to formal grading criteria prescribed by schools. Still, there were occasions, as for borderline cases, when teachers heavily relied on other factors including effort, student conduct, and motivation in determining grades. What is noteworthy to mention here is that in the absence of such formal grading criteria, teachers may assign more weight to non-academic factors for grading.

Building on earlier efforts to unpack the meaning of a grade, Cheng and Sun (2015) studied the assessment and grading practices of 350 Chinese secondary English language teachers. The results showed that teachers used both achievement and non-achievement factors in grading, placing more emphasis on the latter, which included effort, study habits, and homework. The results of factor analysis revealed that three underlying components existed in factors used for grading consisting of (a) *norm/objective-referenced factors* including other teachers' grade distributions, achieved learning objectives, incomplete assignments, school policy, performance compared with students from previous years, and participation and attention, (b) *effort factors* consisting of homework, effort, improvement, work habits, and also disruptive behavior, and (c) *performance factors* including academic and non-academic performance, performance compared with other students, and academic ability. Their categorization of grading criteria closely resembles what Guskey (1996) did in generating product, process, and progress factors, with the difference mostly relating to a name game.

Other quantitative studies were carried out that yielded nearly identical results concerning what a grade included (e.g., Brookhart et al., 2016; Guskey, 2011; Yesbeck, 2011). However, Svennberg et al. (2014, 2016) in two separate qualitative studies with similar designs but slightly different purposes investigated less verbalized internalized factors used by seven Swedish physical education (PE) teachers (four in the first study and three in the second) with varying years of teaching experience. They conducted interviews with teachers using Kelly's repertory grid technique. Of the 86 and 125 constructs elicited from teachers in the first and second studies respectively, four themes were created, including (a) *skills and knowledge*, (b) *motivation*, (c) *confidence*, and (d) *social skills* (interaction with others). These criteria were not used consistently by teachers when grading, and teachers sometimes encountered difficulty giving weight to such standards based on their relative importance. Teachers used the elicited criteria to promote student learning and classroom management as well as to encourage decent classroom behavior. The researchers contended that such standards should reflect the realities of classrooms and other restrictive conditions. The results showed that no matter whether precise knowledge requirements were set or not, teachers continued to include both knowledge-related and non-knowledge-related factors such as values and norms in their grading practices in physical education courses, a result replicated in Randall and Engelhard's (2010) study. One explanation for the inclusion of motivation in grading was that knowledge alone could not motivate people to take action, and that motivation appears to be an essential prerequisite to do so. The findings of this study corroborated the use of non-academic factors in grading reported in other studies (e.g., Brookhart, 1993; Cheng & Sun, 2015; Randall & Engelhard, 2010) and showed that such factors were significant contributors to grading, even though learning objectives had been specified in advance.

Concerning the literature on teachers' grading practices, it can be argued that the majority of reviewed studies on teachers' grading were survey studies

mainly using questionnaires for data collection. The disadvantage of using questionnaires is twofold. Firstly, they present respondents with some factors that they might have been unaware of before setting out to complete the surveys, and by doing so may impose predetermined factors on teachers' pool of actual grading criteria (their constructs in PCT terms) and, as a result, extract responses that have been nonexistent in reality. Secondly, they may fall short of capturing the full range of factors that teachers consider when grading by limiting respondents to the items included in the questionnaire and failing to capture individual differences in grading, resulting in undesirable homogenization of teachers' grading practices across contexts due to the imposition that originates from the questionnaire itself. In other words, respondents may see items they have not thought about before, and they may not see items they have in mind, probably resulting in the production of a reduced version of reality at its best.

Still, another limitation in the literature on teachers' grading practices is that, in comparison with repertory grid interviews, surveys may be influenced more frequently by phenomena such as social desirability that subsequently threaten the internal validity of research studies. Additionally, few studies have employed a theoretical lens such as Personal Construct Theory to study teachers' classroom grading practices to date, except for Svennberg et al. (2014, 2016) who studied Swedish PE teachers' less verbalized grading criteria using the repertory grid interviewing technique. To address the preceding issues, this study aims at exploring the factors that Iranian English language teachers consider when determining grades by using Personal Construct Theory (PCT) as its theoretical foundation.

Method

Participants

This study was conducted in the Iran Language Institute (ILI) as one of the oldest national English language institutes in Iran with approximately the

largest population of EFL teachers (over 2500 teachers) and learners (more than 1,200,000 foreign language learners) with 290 language centers in over 131 cities across the country[1]. Additionally, the choice of this specific site for the present study pertains to other specific organizational factors that may help minimize the effects of confounding variables including (a) hiring the most competent teachers through clear teacher selection and recruitment procedures, (b) holding more regular teacher training programs, (c) administering various pre-service and in-service training and professional development courses, and (d) constant monitoring of language teaching and learning quality utilizing strict class observation, specification of curricular aims and objectives, and administration of nationally standardized tests.

Five teachers (three males and two females) were selected for this study. Purposive sampling was carried out based on the maximal variation sampling criterion concerning gender and teaching experience, with both novice (fewer than two years' teaching experience) and experienced teachers (more than ten years of teaching experience). The selected teachers were informed about the purpose and procedure of the study after having obtained their participation consent. They were also assured that their participation in this study was voluntary and that they were entitled to withdraw from the study at any time. Additionally, pseudonyms were used to refer to teachers in order to keep their identities confidential.

Noah is 37, with a minimum of 15 years' teaching experience. He is a Ph.D. student of Teaching English as a Foreign Language (TEFL) and has a BA and an MA in English language translation and TEFL, respectively. He is ranked a senior teacher by the institute and is permitted to teach all proficiency levels (from Basic to Advanced). Chris, 41, has been teaching English for 20 years now (by fall 2019) but holds a Ph.D. in power engineering and has a corresponding BA and MA in power engineering, too. He teaches English

[1] This information was obtained from http://www.ili.ir/en/aboutus/history

part-time and is ranked a senior teacher of English, teaching across the full range of proficiency levels. His first language is Turkish, but speaks both Persian and English fluently, as well. Simon, 31, has been teaching English for two years now and is ranked a novice teacher by the institute, meaning he is only allowed to teach students at the first three proficiency levels (Basic, Elementary, and Pre-intermediate). He has a BA and an MA in English language literature. The next participant is Julia, 41, who is a senior EFL teacher with 15 years of teaching experience. She has a BA in English language translation and an MA in English language literature. She has permission to teach across the full range of proficiency levels, too. Lastly, Mary, 25, is a novice English language teacher with only two years' teaching experience. She is comparable to Simon concerning career status, and the same as him, she is only authorized to teach up to pre-intermediate level by fall 2019. She has a BA and an MA in English literature. All teachers, except for Noah, admitted that their assessment literacy is limited to the few language testing courses they formally studied at the university and also to the very general guidelines for classroom assessment presented in teachers' manuals specified by the institute where they work. Besides, all participants except Chris speak Persian as their first language, and all come from the same city with almost similar socioeconomic, religious, and ethnic backgrounds.

Repertory Grid (RG) Data Collection

The RG interview technique was employed in the present study to help teachers verbalize the internalized criteria they use in assigning grades. As discussed earlier, Fransella et al. (2004) claim that this method is ideal in helping individuals express their internalized perceptions and tacit knowledge about the specific subject matter with which they are most familiar. Teachers may usually find it difficult to overtly explain the factors they use in their grading in regular interviews. Still, they are expected to do so more effortlessly when, in the repertory grid interviewing, they are asked to

compare and contrast elements (students) across the full spectrum of grades (high, average, and low grades) in their classes and then to express the differences in relation to the grading criteria used.

The one-on-one repertory grid interviews were conducted in private rooms and were simultaneously tape-recorded after receiving each interviewee's prior permission for doing so. Each interview lasted around 70 to 90 minutes. This technique consists of three steps. In the first step, called the *element generation step*, the interviewees (the teachers) were asked to select nine students, including three learners with high, average, and low grades from the grade sheets belonging to one of their classes with whose students they had the most familiarity. The teachers were told beforehand that it was important that they know students well enough prior to their selection as elements. The nominated students functioned as elements for the subsequent triading and systematic comparison once their names were written on separate notes.

The second step of the interview, referred to as *construct elicitation*, consisted of generating constructs by randomly selecting three elements from the pool of nine available learners (triading) from only one single class while keeping the student gender and level constant and then requiring the teacher to choose the two that are similar in one way and different from the third concerning the grades assigned. The main purpose of the RG data collection technique is to extract as many factors (both achievement and non-achievement) from the interviewees as possible. As a result, they are free to select their own triads and generate the constructs that make the most sense to them, regardless of whether such constructs are academic or non-academic. The similarities between the two elements expressed by the teacher constituted one pole of a construct, whereas the difference between these two with the third element in the same triad constituted the other pole of the generated construct. The elicited constructs were subsequently written in separate rows of the construct elicitation form (Appendix), and the interviewee was then requested to rate these three elements on a scale of 1 to

6 based on the degree with which the specific element could be identified with the elicited construct (1 represented the *least effect* for the intended construct while 6 represented *the most effect*). The teacher rated all the other elements in the same fashion for each of the generated constructs in turn. Different combinations of triads (choices of three elements) were randomly presented to the teacher for construct elicitation until the interviewee could generate no further constructs. The generated constructs represent the criteria that teachers attached more weight to when assigning grades. The total number of generated constructs from five separate repertory grid interviews with 45 elements (nine elements per each interview) was 92.

Teachers' grading practices, as reviewed in the literature here and elsewhere, may be restricted to one or a number of subject matters such as English, history, social sciences, etc., and in this case English. However, they are not restricted to any single skill or ability and are treated holistically simply because this is the nature of the teachers' grading that is under study. As endorsed by other research (Brookhart, 1991, 1993; McMillan, 2001, 2003), teachers assign a hodgepodge grade of effort and ability coupled with achievement. This has been found to be the case in various subjects and grade levels.

RG Data Analysis

The data analysis consisted of two distinct phases. In Phase I, *the content analysis phase*, the generated constructs were coded and categorized separately by two raters in order to identify the underlying themes that the constructs fit into best. Two identical copies of the generated constructs were handed to the raters who were requested to categorize the constructs under the headings that most clearly represented the intended construct(s). Afterward, the two copies were collected and compared with each other to find the similarities and possible contrasts in coding. Subsequently, the raters including the researcher and a faculty member discussed possible discrepancies between their coding so that the most-agreed-upon themes were generated in the end. The categories included a) academic enablers (effort, ability, participation, improvement, attention, and work habits), b) cognitive

factors (self-confidence, motivation, assertiveness, and creativity), c) student behavior, d) homework, and e) teacher-specific factors. For instance, if the interviewee stated that how the student behaved in class mattered to him/her when assigning a grade, or more specifically mentioned student impoliteness as an influential factor, the intended construct was categorized under *student behavior*.

The inter- and intra-coder reliability estimates were also reported. The intra-coder reliability estimates (two separate coding attempts with a two-week interval reported as matching percentages) for each of the raters 1 and 2 were 96.7% and 87.5%, respectively. The inter-coder reliability was 85%. Subsequently, member-checking was carried out as an effort to establish the validity of the elicited constructs by discussing the generated themes and their corresponding categories with each interviewee to obtain their probable (dis)approval of the appropriateness of the thematic map created.

In phase II, *the PrinGrid analysis phase*, each interviewee's generated constructs, and their corresponding Likert-scale ratings were analyzed using Rep Plus software version 1.1, a program specifically developed for analyzing repertory grid data. The resultant PrinGrid maps, the graphic equivalents of the data produced through Principal Component Analysis (PCA), were used to explore each teacher's grading criteria (presented as constructs earlier) concerning the ratings they assigned to each element when considering the generated constructs. The positioning of the elements (students) on the PrinGrid maps where the constructs for each interviewee are presented and centered around the underlying factors (represented as axes) would help us know how teachers employ the generated constructs in assigning grades in reality by uncovering the factors that explain the most variance in grading. The tables of elicited constructs and the corresponding PrinGrid maps are presented in the following section.

Results

The total number of grading-relevant constructs elicited from the five English language teachers using the semi-structured repertory grid

interviewing technique was 92. Table 1 presents the frequency of the constructs generated by each interviewee. As is evident, approximately the same number of constructs was elicited from each interviewee except for Simon with the least number of constructs (15 constructs or 16.3 % of the total).

Table 1.
Interviewees' Elicited Constructs Summary

Interviewees	Number of constructs n(%)
Chris	19 (20.7)
Simon	15 (16.3)
Julia	19 (20.7)
Mary	20 (21.7)
Noah	19 (20.7)
Total	92 (100.0)

Concerning the emphasis of the present study on the theme development underlying grading practices using the PCT, Table 2 presents the output of the coding process by displaying the major and minor themes and their categories along with some of their corresponding sample constructs. As can be seen, the two major themes are labeled as (I) *academic factors*, and (II) *non-academic factors*, with the latter theme consisting of subthemes labeled (a) *academic enablers*, (b) *cognitive factors*, (c) *student behavior*, (d) *homework*, and (e) *teacher-specific factors*. The subtheme labeled *academic enablers* consists of six underlying categories that stem from the literature reviewed earlier, namely as (1) *effort*, (2) *ability*, (3) *improvement*, (4) *work habits*, (5) *attention*, and (6) *participation*. Additionally, the *cognitive factors* subtheme encompasses four underlying categories as (1) *self-confidence*, (2) *motivation*, (3) *assertiveness*, and (4) *creativity*. The use of the combination of Roman numerals, small letters, and Arabic numerals is just for the ease of visualizing the relationships among the elicited themes and categories.

Table 2.

Themes, Categories, and Sample Constructs from the Interviewees After Coding

Major themes	Minor themes	Categories	Sample constructs
Academic factors		Speaking	High fluency in speaking / low fluency in speaking (Chris)
		Listening	Good listening comprehension / poor listening comprehension (Simon)
		Vocabulary	Extensive vocabulary knowledge / poor vocabulary knowledge (Chris)
		Grammar	Highly accurate grammar / inaccurate use of grammar (Simon)
		Pronunciation	Accurate pronunciation of words / poor pronunciation of words (Noah)
Non-academic factors	**Academic enablers**	Effort	High effort for learning / low effort for learning (Simon)
		Ability	Having a native-like accent / having a poor Persian accent (Noah)
		Improvement	Making good academic progress / making poor or not making any academic progress (Mary)
		Work Habits	Having a good handwriting / having poor and illegible handwriting (Noah)
		Attention	Highly attentive in class / poor attention in class (Julia)
		Participation	High class participation / low class participation (Julia)
	Cognitive factors	Self-confidence	Having high self-confidence / low self-confidence (Mary)
		Motivation	High motivation for learning / poor motivation in class (Noah)
		Assertiveness	Highly assertive / having low assertiveness (Julia)
		Creativity	Expressing creative views in discussions / having little if any creativity in discussions (Simon)

	Student behavior	Polite behavior / impolite behavior (Simon)
	Homework	Neat and tidy homework / untidy homework (Simon)
	Teacher-specific factors	Having a positive impact on peers / having no impact on peers (Mary)

Ability, as an academic enabler, is exemplified by constructs such as speaking English with a native-like as opposed to a Persian-like accent. It was disqualified from being recognized as an academic criterion because none of the teachers stated that they had taught either of the American or British accents in class previously as they were not required by their syllabi to do so. Nevertheless, four teachers (Chris, Julia, Mary, and Noah) acknowledged that the native-likeness of a language learner's accent positively influenced their grading. For instance, Noah stated that he thought highly of those students who spoke pure American or British accent in their first-class encounters with him by saying, "I am quite sure they are abler than other students as they have put some extra effort into picking up the details of the language they're learning". Similarly, Mary, in her sixth elicited construct, referred to accent as a sub-dimension of ability by saying, "Students, in my class, who have better accents have proven to be better in other language abilities such as communicating with others or answering teachers' questions".

As shown in Table 2, the *work habits* category is exemplified by constructs such as the quality of a learner's handwriting (legibility vs. poorness or illegibility of one's handwriting), as referred to by Noah. During the interviews, three teachers pointed out that learners' handwriting influenced the way they graded students' written assignments. For example, Julia noted that the tidiness of learners' homework in both their workbooks and notebooks mattered to her and, subsequently, affected the grades assigned for their written work partially. This and similar comments made by other teachers led to the elicitation of the construct *neat homework vs. untidy*

homework. *Participation,* or the degree of a student's involvement in different class activities endorsed by the teacher, was another academic enabler that influenced teachers' grading. According to Julia, who mentioned *participation* in the beginning of her interview, higher levels of participation in class should be accompanied with higher grades simply because participation results in better and more efficient mastery of skills such as speaking and listening along with its role in motivating others to take the initiative in class. This view was corroborated by all other teachers as Noah, Simon, Chris, and Mary also believed that participation in class activities was regarded as the cornerstone of student language learning. *Effort*, *improvement*, and *attention* are the three remaining academic enablers that refer to the degree of a student's attempts to learn even when such learning does not happen, the extent of a student's progress comparing their performance at the beginning and the end of a term, and the extent a student pays attention to the teacher as perceived by the teacher, respectively.

In this study, *cognitive factors* are conceptualized as language learners' psychological attributes perceived by the teacher to be important in student learning in class. Table 2 illustrates the categories underlying *cognitive factors* together with their corresponding constructs. As referred to by the participants during the interviews, the teachers made alterations to the grades assigned based on their perceptions of, for example, how self-confident a learner was or how creatively (s)he expressed views in activities such as class discussions or group works. Simon stated that producing creative constructions in class discussions or using old words creatively as in sentence-making exercises by the learner really mattered to him as far as grading was concerned and added, "I suppose creativity should be credited and advocated by teachers by assigning relatively higher grades to those who are more creative. This *may* encourage other students to be creative". Further, the *cognitive factors'* theme comprised students' motivational levels and assertiveness as perceived by the teacher in the classroom. For example, Chris spoke about how students who

appeared to be more motivated than others and asked him for advice about how best to learn ultimately earned better grades and test scores. He briefly commented, "Motivation means success. [It] means everything about learning".

For the other three minor themes underlying non-academic factors, in vivo coding scheme was used. For instance, the subtheme labeled *student behavior* comprised of constructs directly related to student conduct toward their teacher and peers. To all the teachers interviewed, the manner a student treated them in class or even outside the class and the degree of respect the teachers perceived students showed had a substantial impact on the grades the teachers assigned. Student politeness for Simon and student sitting posture in class for Julia were significant factors contributing to variations in grades, particularly when borderline cases were concerned. The other two minor themes of *homework* and *teacher-specific factors* pertained to the degree and the quality of homework done ([in]complete or [un]tidy homework) and grading-relevant idiosyncrasies for each teacher respectively. Examples of the teacher-specific factors include learners' effects on their peers, comparisons of their current performance with that of the previous semesters with the same teacher, or what Brookhart (2009) termed self-referenced grading or even loudness and clarity of a student's voice when asking questions or expressing ideas.

Table 3 presents the frequency of constructs across the generated themes and categories for each interviewee. The rationale behind quantifying these qualitative data at this point is not to immerse ourselves in numerical data, but to create a more comprehensive picture of the distribution of non-academic factors. Of the 92 constructs in total, 72 (78.3%) were coded as non-academic factors, while only 20 (21.7%) belonged to the academic factors' theme. In other words, the non-academic factors outnumber the academic factors by nearly four times. The percentage difference observed is replicated for almost all individual interviewees. That is, the percentage of non-academic to

academic factors for Chris is 73.3% to 26.3%, for Simon 73.3% to 26.7%, for Julia 78.9% to 21.1%, for Mary 85% to 15%, and for Noah 78.9% to 21.1%, respectively.

Table 3.

Descriptive Statistics for the Elicited Themes and Categories

			Interviewees					
Major Themes	**Minor themes**	**Categories**	**Chris n (%)**	**Simon n(%)**	**Julia n(%)**	**Mary n(%)**	**Noah n(%)**	**Total**
Academic factors			5 (26.3)	4 (26.7)	4 (21.1)	3 (15)	4 (21.1)	20 (21.7)
Non-academic factors			14 (73.7)	11 (73.3)	15 (78.9)	17 (85)	15 (78.9)	72 (78.3)
	Academic enablers		4 (21)	4 (26.6)	6 (31.6)	6 (30)	6 (31.6)	26 (28.3)
		Effort	2 (10.5)	3 (20.0)	0 (0.0)	0 (0.0)	0 (0.0)	5 (5.4)
		Ability	1 (5.3)	0 (0.0)	1 (5.3)	2 (10.0)	1 (5.3)	5 (5.4)
		Improvement	0 (0.0)	0 (0.0)	1 (5.3)	1 (5.0)	0 (0.0)	2 (2.2)
		Work Habits	0 (0.0)	0 (0.0)	0 (0.0)	1 (5.0)	2 (10.6)	3 (3.3)
		Attention	0 (0.0)	0 (0.0)	1 (5.3)	1 (5.0)	0 (0.0)	2 (2.2)
		Participation	1 (5.3)	1 (6.6)	3 (15.7)	1 (5.0)	2 (10.6)	9 (9.8)
	Cognitive factors		3 (15.8)	1 (6.7)	2 (10.5)	4 (20)	3 (15.8)	13 (14.2)
		Self-confidence	1 (5.3)	0 (0.0)	0 (0.0)	1 (5.0)	1 (5.3)	3 (3.3)
		Motivation	1 (5.3)	0 (0.0)	0 (0.0)	1 (5.0)	1 (5.3)	3 (3.3)
		Assertiveness	0 (0.0)	0 (0.0)	1 (5.3)	1 (5.0)	1 (5.3)	3 (3.3)
		Creativity	1 (5.3)	1 (6.6)	1 (5.3)	1 (5.0)	0 (0.0)	4 (4.3)
	Student behavior		3	5	4	4	4	20

		(15.8)	(33.3)	(21.1)	(20)	(21.1)	(21.7)
	Homework	1 (5.3)	1 (6.7)	2 (10.5)	1 (5)	1 (5.2)	6 (6.5)
	Teacher-specific factors	3 (15.8)	0 (0.0)	1 (5.2)	2 (10)	1 (5.2)	7 (7.6)
Total		19 (100.0)	15 (100.0)	19 (100.0)	20 (100.0)	19 (100.0)	92 (100.0)

The most frequently-referenced non-academic factors are academic enablers with 26 (28.3%) constructs, student behavior with 20 (21.7%) constructs and cognitive factors with 13 (14.2%) constructs in total. Teacher-specific factors with seven (7.6%) constructs and homework with six (6.5%) of the total number of elicited constructs rank four and five in this listing. As far as the categories are concerned, participation, as an academic enabler, was the most frequently-referenced of all the six categories underlying this theme, with nine (10%) constructs in total. Of the 26 constructs belonging to the academic enablers, *participation*, *effort*, and *ability* included more constructs than others, with participation-related constructs almost double the number of constructs in each of the other two categories.

Student behavior theme, with 20 (27.8%) constructs out of the total 72 non-academic constructs, was the most prominent non-achievement subtheme following academic enablers. Based on the data presented in Table 3, academic enablers, student behavior, and cognitive factors are the most popular minor themes of non-academic theme contributing significantly to variations in grading. Homework and teacher-specific subthemes do not seem to be significant actors here, as judged by the construct allocation percentages.

Figure 1 shows a graphical presentation of the factors used in teachers' grading decision-making in this study. The two-sided arrows point to the trade-offs between major themes, minor themes, and their underlying categories. Although grade-relevant factors are broken down into their constituents from top to bottom, grading decision making appears to be a bottom-up process where teachers observe performances, behaviors, or

attributes, and then follow the arrows upward in their mental hierarchy of grading factors to arrive at decisions for assigning fair grades.

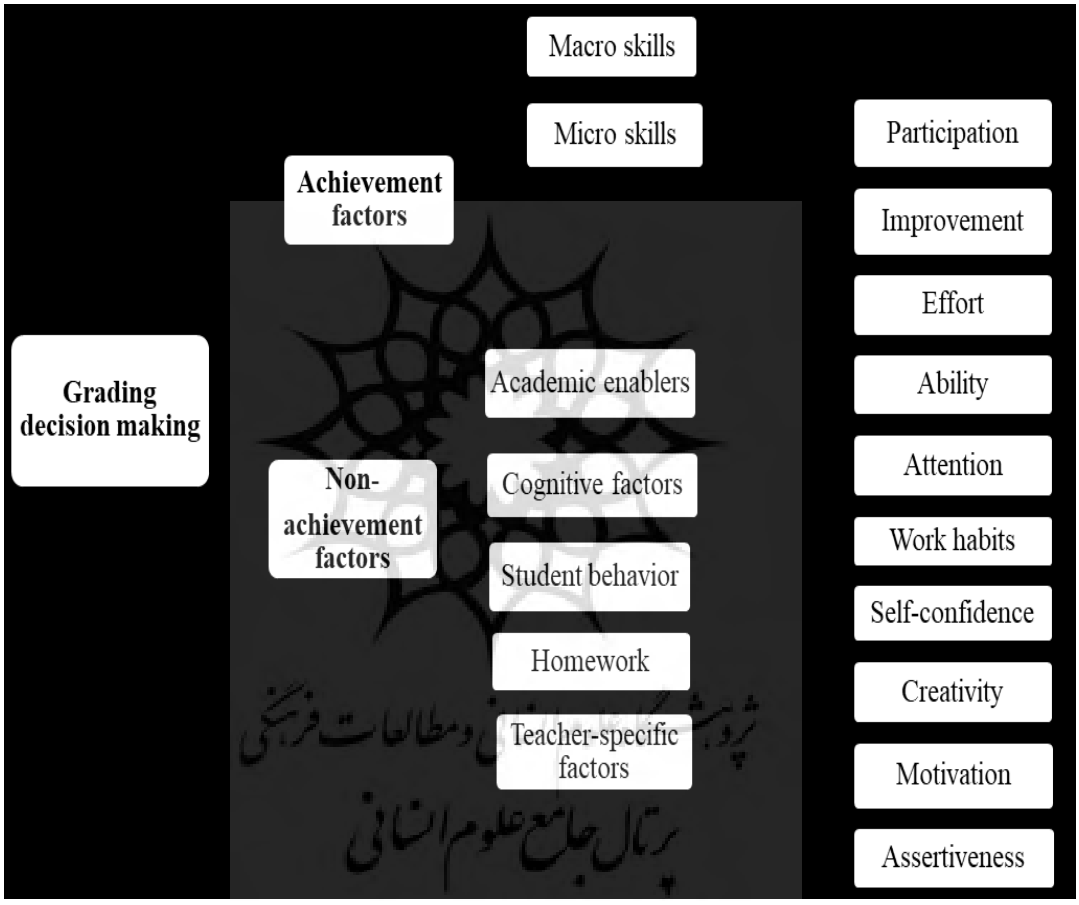

Figure 1.

EFL Teachers' Grading Decision Making Components

Figure 2 shows the PrinGrid map of the elicited constructs used in Noah's grading. This figure is part of the Principal Component Analysis (PCA) to explore the factors underlying teachers' grading decision-making as part of the repertory grid data analysis. Of the five components (factors) generated,

only the first two components that represent the most substantial amount of variation in grading are shown here as X and Y axes, representing 85.7% and 6.4% of the variance in each component, respectively, and the other three factors represent only 8.0% of such variation. In this figure, constructs are represented as lines crossing the center and elements (the nine learners used for construct elicitation) are shown with labels such as HG (high graders), AG (average graders), and LG (low graders) on the diagram with numbers referring to their identity. What is important here is the weight given to the first component with a clustering of the majority of constructs, including an array of both academic and non-academic factors such as attention, handwriting, pronunciation, and homework around this dominant component. The second component (factor) explains only 6.4% of the variance in teachers' grading, which is considered negligible.

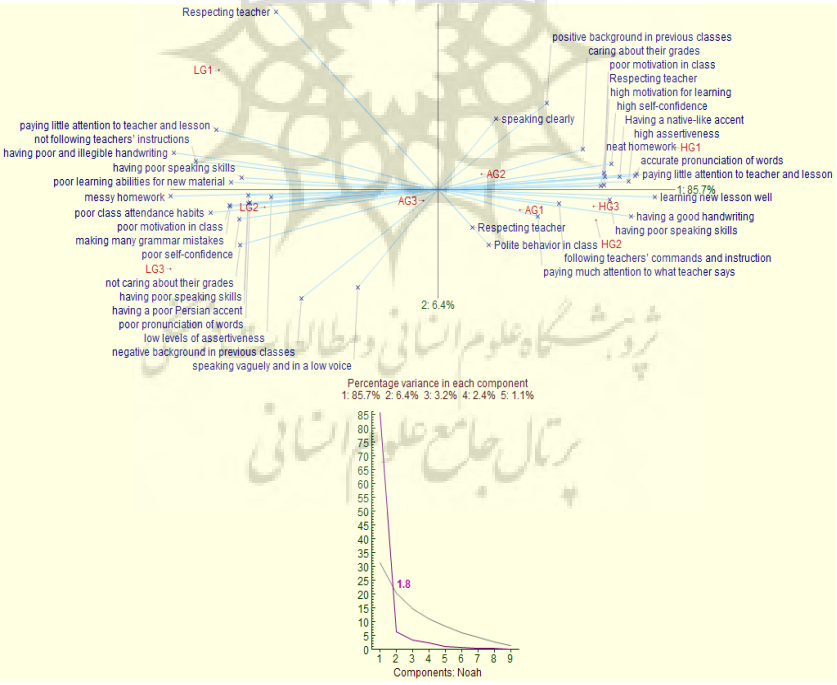

Figure 2. *Noah's PrinGrid Map of the Elicited Constructs Used in Grading EFL Learners*

Discussion

This study aimed to explore the factors that five Iranian EFL teachers use to determine students' grades using Kelly's (1955) Personal Construct Theory. The findings of this study verify that teachers use both academic and non-academic factors in grading, but attach more weight to the latter, contrary to measurement experts' recommendations. Further, the results of PCA show that, in the absence of clearly-specified grading criteria, teachers assign a hodgepodge grade by combining achievement and non-achievement factors. Similarly, the problems that stimulated this study to pertain to validity concerns of teachers' grading decision-making that, in turn, raise doubts about what it is that a reported grade conveys to the stakeholders involved. In order to address these issues, the first research question was formulated as follows:

RQ1: What factors do Iranian English language teachers consider when assigning grades?

To answer this question, repertory grid interviews with the EFL teachers were conducted, and overall, 92 constructs were elicited. After coding the elicited constructs, two major themes (*academic* and *non-academic* factors), five minor non-achievement themes, and their ten underlying categories were obtained, as shown in Figure 1. Following the calculation of the percentages of constructs that belonged to each theme or subtheme, it was found that the majority of the elicited constructs, that is 72 (78.3%) constructs, belonged to the non-academic factors' theme, pointing to the pervasiveness of such factors in teachers' grading. Likewise, the extensive use of non-academic factors for determining grades has been reported in other studies on grading (e.g., Brookhart, 1993; Cheng & Sun, 2015; Cizek et al., 1996; Frary et al., 1993; Gullickson, 1985; Lawrenz & Orton, 1989; McMillan et al., 2002; Randall & Engelhard, 2010; and Stiggins et al., 1989). Although teachers' heavy reliance on non-achievement criteria was evidenced when assigning grades, discrepancies were observed on the extent that these factors were threaded into the final grade. McMillan (2001), in his study of secondary teachers'

grading criteria, asserts that although academic enablers such as participation, effort, and improvement are essential for many teachers, academic achievement is the most significant component in grading, a finding that contrasts with those in the present study due to the observed dominance of non-academic factors. In effect, Brookhart's (1993) assertion concerning the scarcity of grades based on achievement alone is endorsed in the present study. The rationale behind teachers' decisions to use non-academic factors extensively may be teachers' trying to be fair in their grading by taking into account all that a student brings to class and not just what he/she academically achieves. Such perceptions further substantiate the concept of seeing a grade as payment for student work, proposed by Brookhart (1993). Additionally, Svennberg et al. (2014, 2016) assert that teachers continue to extensively consider non-academic factors, probably to keep students motivated even when precise achievement requirements are specified in advance. Similarly, Faravani and Atai (2015) believe that teachers may prefer to adapt course materials to students' multiple-intelligences, which hopefully may contribute to creating more appealing learning environments where students are encouraged to be risk takers or to employ various task-specific reasoning skills.

The dominance and the prevalence of non-achievement factors were evident in each interviewee's constructs. This finding appears to be in sharp contrast with measurement experts' recommendations to teachers and educators to base their grading solely on achievement, as Brookhart (2004) contends that grading aims to let students and parents know to what extent the learning goals have been achieved. In this study, however, it was found that teachers relied heavily on non-achievement criteria such as student behavior, class participation, cognitive factors, and academic enablers at the expense of marginalizing achievement criteria. The model presented here corroborates McMillan and Nash's (2000) grading decision-making model in sorting all these criteria under internal factors which consist of teachers' philosophy of

teaching and learning and their beliefs and values. The point of departure, however, is that in the absence of dominant external factors such as pressures for accountability and transparent grading policies in this study's setting, internal factors have gained more momentum. Besides, in their model, McMillan and Nash (2000) did not clearly specify what was meant by teachers' teaching-learning philosophy and their beliefs and values except for generalizing them under internal factors. In contrast, the present model sheds more light on what constitutes these internal factors by tapping into the less verbalized criteria such as cognitive factors that influence grading, but have mainly remained understudied. Guskey's (1996) *process factors* could similarly be interpreted to include cognitive factors and academic enablers, though they seem to be generalized under processes that support and promote learning. Further, Cheng and Sun (2015), in their categorization of grading criteria, did not appear to be concerned about cognitive factors for the simple reason that such factors as teachers' perceptions of student creativity, motivation, and self-confidence in performing various class tasks were not mentioned.

Classroom realities and teachers' own beliefs and values concerning assessment in a specific context are essential elements that influence teachers' grades (Cheng, Rogers, & Wang, 2008; Cheng & Wang, 2007). The current findings are also consistent with these results and those of Randall and Engelhard's (2010) study in pointing to a heavy reliance on non-achievement factors such as behavior and motivation. The difference, however, lies in the fact that in the latter study, teachers resorted to such factors only for assigning grades to borderline cases, whereas in the present study the reliance on non-academic factors is more far-reaching than those anticipated by previous studies, and the weight attached to non-academic factors extends into teachers' day-to-day grading practices. In Brookhart's (1993) study, teachers' frequent consideration of effort as a grading criterion was solely limited to students with below-average performances, and McMillan and Nash (2000)

referred to the saliency of teachers' philosophy of teaching and learning and their beliefs and values as a corollary to the achievement criteria while in the current study, non-achievement factors are dominant and play a far more significant role in grading. While it seems that various grading models seek to capture the essence of a teacher's grade, it makes sense here to assert that there might probably be no one single best grading model that fits every teacher's grading practices everywhere.

The fact that teachers in this study used a combination of academic and non-academic criteria for grading was not hard to predict because numerous studies have also come to the same conclusion concerning grading (e.g., Brookhart et al., 2016; Guskey, 2011; Randall & Engelhard, 2009, 2010; Russell & Austin, 2010; Sun & Cheng, 2013; Svennberg et al., 2014; Yesbeck, 2011). What seems worth mentioning in the present study, though, is the very high percentage of non-academic factors represented by constructs compared to that of the academic factors across the interviewees. In other words, the rate of non-achievement factors is above 70% while this percentage for achievement factors is below 26% for all interviewees as shown in Table 3, meaning that only one-fourth of the factors used for determining grades by the teachers in this study was labeled achievement factors and the remaining were non-achievement. The findings of this study corroborate those found by other studies as far as the typology of grading criteria is concerned. Additionally, teachers seemed to vary from each other in the weight they attached to non-academic grading criteria, a finding endorsed by other studies (e.g., Adrian, 2012; Cross & Frary, 1999; Duncan & Noonan, 2007; McMillan & Lawson, 2001; McMillan, Myran, & Workman, 2002; Randall & Engelhard, 2009, 2010).

One possible explanation behind the predominance of non-academic grading criteria in the present study could be the use of repertory grid technique to disclose less verbalized or implicit criteria that, according to Kelly (1955), could be "formulated or implicitly acted out" (p. 9). In line with

this assertion, the prevalence of non-academic criteria may relate to the epistemology of constructivism, which disapproves of the existence of only one single truth out there and recognizes the multiplicity of worldviews when various phenomena such as grading are concerned. The subtle point concerning the observed discrepancy in the extent of non-academic grading criteria between this and other survey descriptive studies may be attributable to this epistemological difference. There might also exist other causes that add to the complexity of the issue at hand such as poor assessment literacy, lack of uniform grading policies, and the impact of sociocultural forces, among others, whose study lies well beyond the scope of the present study.

By scrutinizing the data presented in Table 3, one will also realize that despite the similarities among teachers in their proportionate use of academic and non-academic factors discussed earlier, variations exist when it comes to teachers' preferences for the subthemes and categories underlying the major non-academic factors' theme, a result endorsed by Svennberg et al. (2014, 2016). While cognitive factors such as self-confidence or motivation make up only one (6.7%) of the total of 15 constructs for Simon, they make up four (20%) of the sum of Mary's 20 constructs elicited. Similar trends were also observed for subthemes such as student behavior for Chris and Simon. It might be concluded that teachers' grading practices undergo both systematicity and variability as percentage trends were nearly identical for all interviewees, meanwhile, teachers vary from one another in their preferences for the criteria selected for grading. What seems evident, however, is that extensive use of non-academic grading criteria and observed individual differences among teachers in their preference for such standards underscore the need to pay special attention to teachers' assessment literacy via teacher education programs. Back to the first research question, it can now be concluded with more confidence that non-academic factors constitute an indispensable part of the grades reported by teachers and in doing so the interpretability of grades

assigned and the validity of their use may be seriously at stake, raising more questions about what a grade means.

The second research question concerns hodgepodge grading discussed extensively in the literature and is formulated as follows:

RQ2: Do the teachers assign a hodgepodge grade? If yes, how is the hodgepodge grading viewed from the Personal Construct Theory perspective?

To answer this question, we should investigate the PrinGrid maps generated for each interviewee. As was shown in Noah's PrinGrid in Figure 2, of the five components extracted, only the first one accounted for nearly 86% of the variance in this teacher's grading with the other four remaining components accounting for only 14% of the grading variance in total. This finding is approximately the same in all the other four generated PrinGrid maps, highlighting the existence of one major component accounting for the most variance in grading. The issue that teachers use an amalgamation of academic and non-academic factors in the grades they assign as evidenced by the proximity of different criteria to the single main factors (X-axes) in all PrinGrids also verified in Figure 2, could be interpreted as hodgepodge grading corroborated in other studies (e.g., Brookhart, 1991, 1993; Brookhart et al., 2016; Guskey, 2011; McMillan, 2001; Randall & Engelhard, 2010). However, an interesting point here is that although teachers more likely assign a hodgepodge grade confirmed by the existence of single main components and also use non-academic factors extensively as corroborated in the content analysis phase, they seem to be doing so consistently. This finding is corroborated by the fact that the nine elements, including the three high-, the three average-, and the three low-graders are clustered close to each other in each of the five generated PrinGrids. In other words, while the construct validity of teachers' grading may suffer greatly due to the inclusion of non-achievement factors as seemingly construct-irrelevant that raise concerns about grade interpretation and use among stakeholders, the grading reliability might not be as worrying.

Looking at the reliability through the PCT lens, it can be argued that a hodgepodge grade represents the inter-relatedness of teachers' internal mental constructs. No single construct could operate in isolation, without drawing on the other constructs that are used to describe a unique phenomenon. In the teacher's mental world of constructs, a host of factors come into play when an event such as grade assignment is encountered. That may be why finding instances of achievement-based-only grading is a rarity (Brookhart, 1993). This explanation may account for the observed variability in grading. The consistency with which teachers grade students could also be viewed from the PCT perspective. One probable reason could be that teachers in specific contexts also operate under the influence of similar sociocultural factors that constitute their construct systems, leading to the formation of nearly identical viewpoints, worldviews, and interpretations of the phenomena they encounter in their surroundings. This shared sociocultural force might bring about uniformity, particularly in environments where interactions among stakeholders for settling disputes are extensive, as in schools and also relationships between teachers and learners influence teachers' evaluative decisions (Shah Ahmadi & Ketabi, 2019). As far as grading is concerned, this might result in an unconscious consistency when grading. In sum, it can be stated that although the hodgepodge grading is endorsed in this and other similar studies, this might not be interpreted as chaos since both systematicity and variability appear to be the mechanisms that function when a grade is awarded.

Conclusion

The findings of the current study are congruent with those of other studies (e.g., Brookhart, 1993; Cheng & Sun, 2015; McMillan & Nash, 2000; Randall & Engelhard, 2009; Svennberg et al., 2014, 2016; Yesbeck, 2011) in that teachers include non-academic evidence of achievements such as participation, student behavior, and cognitive factors like perceived levels of

learner motivation or self-confidence in their grading practices. The point of departure from previous studies, however, appears to be the degree of reliance on such non-academic factors. The teachers in this study attached heavier weight to non-academic criteria compared to those who were inquired in other survey descriptive studies, which could be attributed to the theoretical foundations used here or, more specifically, to the unique RG data collection technique employed. By probing deeper into one's mental constructs, the repertory grid technique provides researchers with better opportunities to uncover less verbalized or internalized aspects of phenomena such as grading. This feature might account for the elicitation of a broader range of non-academic factors than those obtained from survey studies, an indication that teachers might probably rely more extensively on non-academic criteria for grading than expected, particularly when uniform grading criteria are at worst missing, or at best not transparent.

Heavy reliance on non-academic criteria may give rise to hodgepodge grading, as was the case in the present study where the PrinGrids revealed that teacher grading was mainly centered on one single major component that accounted for the largest variation in grading and that consisted of both academic and non-academic factors, though this amalgamation did not seem to be haphazard. In other words, the grades that teachers assigned were not random collections of various pieces of evidence of achievement but showed systematicity that most likely originated from teachers' internal beliefs and values and their teaching-learning philosophy. Investigating teachers' socioeconomic status and their belief systems could shed more light on why a grade includes what it includes.

Limitations of the Study

Some issues limit the implications and inferences that can be drawn from this study. Firstly, interviews, like other forms of data collection such as surveys, are subject to the social desirability phenomenon that threatens the

internal validity of a study. In this study, however, the repertory grid interviewing technique was used with the hope of reducing the impacts of social desirability by minimizing the role and the felt presence of the interviewer throughout the interviews and letting the interviewee take the lead and navigate the data elicitation process him/herself. To what extent the researchers have been successful in doing so remains open to discussion, though. A second limitation pertains to the difference between teachers' actual and reported grading practices due to the unfeasibility for the researchers to investigate teachers' grade sheets or to observe their classes. Subsequently, this limitation offers future researchers the opportunity to study teachers' grading practices by using more direct approaches such as class observation or artifact analysis using teachers' grade sheets to obtain more reliable information on their grading practices. Additionally, the limited number of participants or the elicited constructs is another limitation of this study. It is suggested that future studies be carried out with more participants (20 or 30 participants) to elicit a larger number of constructs so that the results are warranted with higher reliability and validity.

Implications

This study's findings have implications for teacher professional development and teacher education programs. Informing English language teachers on what constitutes an assigned grade and pointing to the prevalence of non-academic factors in their grading practices, contrary to measurement experts' recommendations, can help promote teachers' assessment and grading literacy, that in the age of educational accountability, appears to be an absolute necessity. This understanding could encourage teachers to reconsider and reconceptualize their own grading practices by being more critical of their grading and, consequently, be better aligned with assessment experts' recommendations. This practice, in turn, can help ease tensions between

school administrators and teachers, on the one hand, and parents and students on the other hand.

Additionally, teacher education programs are among those that can benefit from the findings of this study by educating their pre-service teachers in and sensitizing them to their grading and requiring them to more critically consider what needs to be included in a grade, which can lead to more systematicity in teachers' grade giving practices. All this could spark interest in institutionalizing more contemporary approaches to grading, such as standards-based grading (SBG) and promoting accountability movements in Iran's educational system.

References

Adrian, C. A. (2012). *Implementing standards-based grading: Elementary teachers' beliefs, practices and concerns*. (Doctoral dissertation). Retrieved from ProQuest (1032540669)

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

Annerstedt, C., & Larsson, S. (2010). 'I have my own picture of what the demands are…': Grading in Swedish PEH problems of validity, comparability and fairness. *European Physical Education Review, 16*(2), 97-115. doi:10.1177/1356336X10381299

Bernstein, B. (2003). *Class, codes and control. (Vol. 3) Towards a theory of educational transmission.* London: Routledge & Kegan Paul.

Bjorklund, L. E. (2008). The repertory grid technique, Making tacit knowledge explicit: Assessing creative work and problem-solving skills. In H. Middleton (Ed.), *Researching Technology Education: Methods and Techniques* (pp. 46-69). Rotterdam: Sense Publishers.

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

Brookhart, S. M. (1991). 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*, 123-142. doi:10.1111/j.1745-3984.1993.tb01070.x

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

Brookhart, S. M. (2004). *Grading.* Upper Saddle River, New Jersey: Pearson Education.

Brookhart, S. M. (2009). *Grading* (2nd ed.). New York: Merrill.

Brookhart, S. M. (2013). The use of teacher judgment for summative assessment in the USA. *Assessment In Education: Principles, Policy & Practice, 20*(1), 69-90. doi:10.1080/0969594X.2012.703170

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4), 803-848. doi:10.3102/0034654316672069

Cheng, L., & Sun, Y. (2015). Teachers' grading decision making: Multiple influencing factors and methods. *Language Assessment Quarterly, 12*(2), 213-233. doi:10.1080/15434303.2015.1010726

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

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

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., Fitzgeral, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3*(2), 159-179.

Cox, K. B. (2011). Putting classroom grading on the table, a reform in progress. *American Secondary Education, 40*(1), 67-87.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

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.

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

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

Faravani, A., & Atai, M. (2015). Multiple intelligences, dialogic-based portfolio assessment, and the enhancement of higher-order thinking. *Journal of Teaching Language Skills*, 33(4), 19-44. doi: 10.22099/jtls.2015.3015

Fransella, F., Bell, R., & Bannister, D. (2004). *A manual for repertory grid technique* (2nd ed.). Chichester: Wiley.

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 & Practice, 12*(3), 23-30. doi:10.1111/j.1745-3992.1993.tb00539.x

Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research, 79*(2), 96-100.

Guskey, T. R. (1996). Reporting on student learning: Lessons from the past – Prescriptions for the future. In T. R. Guskey (Ed.), *Communicating student learning. 1996 Yearbook of the association for supervision and curriculum development* (pp. 13-24). Alexandria, VA: Association for Supervision and Curriculum Development.

Guskey, T. R. (2011). Stability and change in high school grades. *NSAAP Bulletin, 95*, 85-98.

Guskey, T. R. (2015). *On your mark.* Bloomington, IN: Solution Tree Press.

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

Guskey, T. R., & Link, L. J. (2018). Exploring the factors teachers consider in determining students' grades. *Assessment in Education: Principles, Policy & Practice, 26*(3), 303-320. doi:10.1080/0969594X.2018.1555515

Hay, P. J., & Macdonald, D. (2008). (Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject. *Assessment in Education, 15*, 153-168. doi:10.1080/09695940802164184

Hay, P., & Penney, D. (2013). *Assessment in physical education: A sociocultural perspective.* London: Routledge.

Jankowicz, D. (2003). *The easy guide to repertory grids.* Chichester: John Wiley & Sons.

Kelly, G. A. (1955/1991). *The psychology of personal constructs* (2nd ed.). London: Routledge.

Klapp Lekholm, A., & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation, 15*(1), 1-23. doi:10.1080/13803610802470425

Lawrenz, F., & Orton, R. E. (1989). A comparison of critical thinking related teaching practices of seventh and eighth-grade science and mathematics teachers. *School Science and Mathematics, 89*(5), 361-372.

Link, L. J. (2018). Teachers' perceptions of grading practices: How pre-service training makes a difference. *Journal of Research in Education, 28*(1), 62-91.

Linn, R., & Miller, M. (2005). *Measurement and assessment in teaching.* Upper Saddle River, NJ: Pearson Prentice Hall.

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.

McMillan, J. H., & Lawson, S. R. (2001). *Secondary science teachers' classroom assessment and grading practices.* Richmond, VA: Metropolitan Educational Research Consortium. Retrieved from ERIC database.

McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making. *National Council on Measurement in Education.* New Orleans: LA.

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

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

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.* Boston, MA: Pearson.

O'Connor, K. (2009). *How to grade for learning: Linking grades to standards* (3rd ed.). Glenview, IL: Pearson Professional Development.

Randall, J., & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1), 1-18.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*, 1372-1380. doi:10.1016/j.tate.2010.03.008

Russell , J. A., & Austin, J. R. (2010). Assessment practices of secondary music teachers. *Journal of Research in Music Education, 58*, 37-54.

Shah Ahmadi, M., & Ketabi, S. (2019). Features of language assessment literacy in Iranian English language teachers' perceptions and practices. *Journal of Teaching Language Skills*, 38(1), 191-223. doi: 10.22099/jtls.2020.34843.2739

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practices, 8*(2), 5-14.

Sun, Y., & Cheng, L. (2013). Teachers' grading practices: Meaning and values assigned. *Assessment in Education: Principles, Policy & Practice, 21*(3), 326-343. doi:10.1080/0969594X.2013.768207

Svennberg, L., Meckbach, J., & Redelius, K. (2014). Exploring PE teachers' 'gut feelings': An attempt to verbalize and discuss teachers' internalized grading criteria. *European Physical Education Review, 20*(2), 199-214. doi:10.1177/1356336X13517437

Svennberg, L., Meckbach, J., & Redelius, K. (2016). Swedish PE teachers struggle with assessment in a criterion-referenced grading system. *Education and Society, 23*(4), 381-393. doi:10.1080/13573322.2016.1200025

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

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 (Doctoral dissertation).* Retrieved from ProQuest.

Young, S. (2011). A survey of student assessment practice in physical education: Recommendations for grading. *Strategies: A Journal for Physical and Sport Educators, 24*(6), 24-26. doi:10.1080/08924562.2011.10590959

Appendix: Construct Elicitation Form