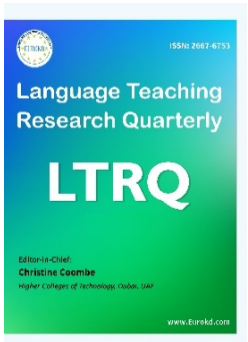

Language Teaching Research Quarterly

2023, Vol. 37, 76–90

Five Golden Rules for Successful Classroom Assessment Based on What We Have Learnt from JD Brown

Aek Phakiti*, Adam Steinhoff

The University of Sydney, Australia

Received *10 January 2023* ***Accepted*** *11 September 2023*

Abstract

Classroom language assessment aims to gather various types of information related to student language learning and use and use it to inform a range of formative-summative decisions, including how to provide feedback to help students improve their learning efficacy, what teaching adjustments are needed, and whether students have demonstrated their learning attainment defined by a given syllabus (Turner, 2012). This article presents five golden rules based on the contributions of JD Brown. These golden rules address a need for a practical and accessible assessment approach that helps language teachers effectively assess students' learning achievement in the language classroom. This article is, therefore, written for language teachers and pre-service TESOL (Teaching English to Speakers of Other Languages) students. The five golden rules presented in this article are: (1) choose the suitable types of language assessment tasks; (2) recognise the two types of assessment interpretations; (3) consider and apply basic classical test theory for analysis of the assessment of learning; (4) develop well-defined assessment criteria and rubrics for judging students' performance; and (5) use assessment of language learning as a source of feedback to improve assessment tasks and student learning.

Keywords: *Classical Test Theory, Classroom Assessment, Feedback, Score Interpretation, Scoring Rubrics, Summative Assessment*

Introduction

JD Brown's significant scholarly contributions to topics and issues, such as classroom assessment, test theory, types of assessment interpretations, and rubrics for performance assessment (e.g., Brown, 2005, 2009, 2012a, 2012c, 2013, 2014a, 2014b, 2022), have influenced the authors' knowledge and practice of language assessment for several years. What makes JD Brown's contributions valuable is the depth of knowledge and insight into the principles and approaches to various testing and assessment methodologies. His ability to communicate complex testing and assessment concepts in an accessible way for his target

* Corresponding author.
E-mail address: aek.phakiti@sydney.edu.au https://doi.org/10.32038/ltrq.2023.37.03

readership stands out, using cohesive explanations of theoretical or technical terms with practical and authentic examples to which his readers can relate.

Since his contributions are considerable, expanding over several decades, it is not possible to reflect on all of his contributions in this short article. Therefore, we formulated five recommendations (golden rules hereafter) informed by JD Brown's work. It is important to clarify that these are not JD Brown's golden rules published elsewhere, but they are synthesised from his contributions to language assessment. In this article, the five golden rules are proposed to be used in relation to assessment of learning in syllabus-based language teaching situations (e.g., English language teaching) rather than situations such as English Medium Instruction (EMI) and Content Language and Integrated Learning (CLIL) in which specific content knowledge or skills (e.g., maths, science, history) are the primary focus. The authors' pedagogical contexts include teaching pre-service teacher training programs (Master of Education in TESOL) and teaching English for academic purposes in Australia. We teach a unit of study named *Language Testing and Assessment* and have applied these golden rules in teaching this unit of study. Our students have found these rules useful for their future language assessment practice, and we would like to share them. Here are some excerpts of our students' reflections.

> *Understanding the motivation behind the different types of assessment has changed my outlook on learning and assessment. It has rekindled a forgotten dream I once had of possibly teaching language one day (Liz).*

> *In the past, when I was an English teacher in China, even though I frequently used language tests and assessments in teaching, I didn't know what they were or what their purposes were. However, after learning these rules, I definitely better understand the relationship between language teaching and assessment (Hui).*

> *I thought that assessment and testing were the same -- which are a means of checking the outcomes of teaching and learners. But actually, each of them is stressed in different aspects. Assessment is more flexible in form and context (Helen).*

> *The discussion on "high-stakes testing" was a real eye-opener. I used to view exams as a "do-or-die" situation, but now, I understand that tests serve various purposes – they can help identify my weaknesses and allow teachers to provide more targeted guidance (Qi).*

Since students' achievement levels (e.g., scores and letter grades) can be permanently recorded in their academic transcripts, and since the assessment of learning affects students' futures (e.g., to repeat the same class, to delay graduation from the program, to have limited opportunities to further education or career advancement), we need to pay careful attention to the practice of summative assessment (discussed further below). This article focuses on the five golden rules that help make summative assessment effective, appropriate, and fair. These rules synergised from JD Brown's publications, and our understanding of language testing and assessment based on our research and teaching experiences are:

1. Choose the suitable types of language assessment tasks.
2. Recognise the two types of assessment interpretations.
3. Consider and apply basic classical test theory for analysis of the assessment of learning.
4. Develop well-defined assessment criteria and rubrics for judging students' performance.
5. Use assessment of language learning as a source of feedback to improve student learning and existing and future assessment tasks.

In articulating each golden rule, we focus on translating its core principles to language assessment practice so that teachers can apply it to their localised contexts.

Background of Classroom Assessment

Language classroom assessment is a process of observing and collecting information or evidence about classroom language learning and performance for decision-making (Bachman & Damböck, 2017; Brown, 2005; Green, 2020; Phakiti & Leung, forthcoming, 2024). At any one time in a given class, information or evidence being collected can be linguistic (e.g., language use accuracy, complexity, relevance, and appropriacy), psychological (e.g., difficulty and challenges, motivation, learning strategies, anxiety, and worries), social (e.g., relationship-building, interactions, cross-cultural communications, social regulation, and social support), and educational (e.g., accountability, selection, diagnostic, and placement).

Two types of assessment are found in the literature: formative and summative (Black & Wiliam, 1998, 2009, 2018; Brown, 2005; Leung & Rea-Dickins, 2007; Rea-Dickins, 2008). *Formative assessment* (also known as *assessment for learning*) collects ongoing and periodic information to inform and improve language teaching and learning relevant to the subject's learning outcomes (e.g., using additional language activities or resources and providing feedback on learning). Some formative assessment can be spontaneous in a given classroom situation (e.g., use of teacher questions to confirm concept learning). Much informal formative assessment is part of everyday teaching and learning activities and is done without recording the results intended to evaluate students' attainment levels. Some formative assessment can be more formal and pre-planned (e.g., use of quizzes to check student learning before the midterm or final examinations; feedback on essay or project drafts before final submission).

Another set of classroom language assessment activities is known as *summative assessment* which is mainly used to inform decisions on students' learning success levels required by the syllabus or course (see Brown, 2005; Popham, 2017). In contrast, summative *assessment* measures students' learning achievement or attainment levels. This type of assessment needs formal and careful record-keeping since decisions about students need to be rigorous and transparent (e.g., achievement tests, assigned coursework, and portfolios). In these examples, formative assessment can be administered under a standardised testing condition (in a test room) or within a specified period (e.g., portfolios, assignments).

Unlike some formative assessment which can be carried out during regular class time and without performance record-keeping, summative assessment requires careful, systematic, and planned administration because of its impact on students and the need for fair and transparent decisions (Brown, 2005). For example, students are formally informed of assessment tasks and tests that will be used to determine their achievement levels when they start the course and when they will need to complete such assessments. Specific dates and times are allocated for test and assessment task completion.

Who Uses Assessment in the Language Classroom?

In considering classroom assessment, it is important to note that teachers are not the only persons engaging in language assessment. Students also use a range of assessments in their learning. We use the term *agencies* to describe teachers and students in the active and purposeful use of assessment. This section describes how teachers and students are agents who independently and collaboratively monitor and evaluate learning processes and achievement levels.

Teachers as Assessment Agencies

Teachers typically embed *informal* assessment during their teaching by monitoring students' behavioural and cognitive engagement with lessons and exercises and by questioning students' comprehension or acquisition of knowledge or skills. When teachers realise that students have difficulties with the language content, they can promptly step in to provide guidance and feedback. Therefore, it can be challenging to separate informal assessment from teaching practice. Teaching-embedded assessments can range from implicit to explicit but can also be planned before teaching. It is also likely that students may not be aware of such teacher assessments during class time.

Teachers also employ *formal* classroom assessments as part of the language syllabus and curriculum. Since formal assessments are explicitly stated, students are aware of such assessments and pay attention to completing them. Such formal assessments include quizzes, assignments, individual or group work, and midterm and final tests that can be aggregated to form final scores for grading students. Formal assessment is used for *accountability*, such as deciding whether students have competently met the learning outcomes. In a classroom context, summative assessment serves a formal accountability purpose (e.g., ensuring that students learn what they are supposed to know). In the language testing and assessment literature (see e.g., Bachman & Damböck, 2017; Brown & Abeywickrama, 2019), formal accountability classroom assessment is regarded as high-stakes in the sense that any resulting decisions can influence students' futures more profoundly (e.g., having to repeat the same subject due to failure, having to delay course completion, having to pay extra tuition fees), as compared with embedded assessments. Informal and formal assessments are interconnected in that they help ensure students' learning and preparedness for assessment activities that determine their future performance or skills (see Black & Wiliam, 1998; Popham, 2017).

In addition to classroom assessment, teachers may be asked to give external standardised assessments to students (e.g., mandatory tests by the academic institute and local and central government agencies). They may be required to help their students prepare to take external standardised examinations or tests successfully (e.g., Year 12 HSC Written Examination in New South Wales, Australia, Gao Kao Examination in China, and English proficiency tests such as TOEFL [Test of English as a Foreign Language] and IELTS [International English Language Testing System]). External assessments can influence the nature of classroom teaching, learning, and assessment activities. It can often be problematic when external assessments must be prepared simultaneously with language curricular activities. Teachers and students tend to shift their classroom attention to preparing for external assessments rather than syllabus-driven formative and summative assessments. Furthermore, mandatory external standardised assessment can impact teachers' employment or promotion. Failure to help

students succeed in such assessments may impede their career progression or promotion (e.g., resulting in discontinuing their employment).

Students as Assessment Agencies

Formative assessment is not what only teachers do. Students engage in formative assessment as they study or complete classroom or assessment activities. Students are considered independent leaders or agencies of their learning. Through self-regulation and self-assessment, for example, students invest their time, available resources, and effort to learn a given lesson and content before, during, and after the class. They do not always wait for their teachers to instruct them on what to do (see Phakiti, 2018). Students are known to set personal and classroom learning goals, plan their actions to achieve them, monitor and evaluate their learning progress, achievement, and difficulties, and predict the likelihood of their learning success. Examples of student assessment include self-assessment and self-testing. Students can self-test by completing classroom activities independently before receiving answers or teacher feedback. They may independently complete practice tests outside the classroom before checking the answer keys. They may evaluate their work using the assessment criteria before submitting it to their teachers. They may even test their memory of what they have studied before the formal examination (e.g., memorising and recalling information, such as vocabulary and grammar rules).

Students can collaborate with their classmates to engage in assessment activities by forming study groups and peer-assessing one another. Peer assessment occurs when students act as critical peer assessors based on teachers' specified assessment criteria.

The Five Golden Rules for Classroom Assessment

The following are the five golden rules for classroom assessment.

Golden Rule #1: Choose the Suitable Types of Language Assessment Tasks

Brown (2012a) points out that using the wrong types of assessment is irresponsible from a professional perspective and can lead to inappropriate and harmful decision-making regarding students' performance or achievement. Simply put, assessment tasks should be related to the target learning outcomes, abilities, or skills. They should be appropriately designed to assess the intended language learning outcomes (Brown, 2005; Brown & Abeywickrama, 2019; Brown & Trace, 2017). For example, content difficulty and task complexity should be at the level students have been learning in the classroom. Students should be familiar with assessment tasks (e.g., tasks are similar to those carried out during classroom learning). It is also essential to align formats of assessment delivery with classroom activities. If students have been learning to complete a task in pairs, assessments related to such a task should also be carried out in pairs. Suppose students do not use computers as part of their classroom activities. In that case, asking them to take a computer-based assessment may not be appropriate since students may not have essential computer skills to demonstrate what they have learnt from the classroom. Some students may be better at using computers than others, giving them an advantage to do better in the assessment tasks than other students.

Brown (2012a) suggests that teachers can avoid choosing the wrong types of assessments by understanding the differences between (1) standardised language proficiency assessment

and classroom language assessment and (2) types of task responses often used in standardised assessment and those in the classroom activities (e.g., selected, constructed and personal responses; see also Brown & Trace, 2017).

According to Brown (2005), standardised language proficiency tests are used to measure students' general language ability that is not necessarily connected to any particular classroom syllabi, learning outcomes, or prior learning experiences. Topics and situations are more general than those in classroom assessment. Language proficiency tests aim to place students on a proficiency scale (e.g., beginner, intermediate, or advanced) relative to other test takers. Classroom summative assessment seeks to determine how much and how well students have achieved the learning outcomes in a particular course (e.g., as reflected by percentages of performance and awarded grades). In such assessments, teachers aim to compare students' performance with the learning outcomes rather than comparing students with their peers.

Second, standardised and classroom assessments can have similar task responses. It is, therefore, essential to understand that different types of task responses are appropriate to other kinds of language knowledge or skills being assessed. For example, selected-response tasks such as multiple-choice, true-false, and matching techniques help determine the amount of language knowledge (vocabulary, grammar, and punctuation) and comprehension in receptive skills such as reading and listening. Constructed-response tasks such as short-answer and written or spoken responses are suitable for assessing productive language skills in which students can apply and use the knowledge or skills they have. Constructed-response, syllabus-based tasks, such as portfolios, project-based tasks, and assigned group work, may assess a mix of language skills (known as integrated language tasks). In classroom assessment, integrated language tasks are more complex since students can personalise their own choices of texts or resources to produce their language responses. Students do not complete them under a testing condition. In standardised assessment (e.g., TOEFL), an integrated task may ask students to write an essay in response to spoken and written texts, combining information from them with their personal viewpoints. Although both classroom personal-response tasks and integrated test tasks require integrated language skills (e.g., discourse synthesis), personal-response tasks are unsuitable for administering under standardised testing conditions (e.g., due to pedagogical reasons for promoting autonomy in learning and feedback on drafts and progress).

A key implication for recognising the types of responses is the need to use the correct task responses for the right skills of interest. For example, using a multiple-choice task for assessing speaking would not be suitable because it does not match the speaking processes. Similarly, using a grammar test to evaluate students' reading comprehension would not be appropriate because reading requires more than grammatical knowledge. Therefore, classroom assessment should reflect the links between what to assess, assessment types, delivery formats (e.g., paper- or computer-based), and classroom learning outcomes and activities (Brown, 2012a).

Golden Rule #2: Recognise the Two Types of Assessment Interpretations

Score or performance interpretations refer to inferring the meaning of test scores or assessment performance beyond raw test scores or assessment to a more abstract level, such as proficiency or achievement levels (see Bachman & Damböck, 2017; Bachman & Palmer, 2010). Assessment interpretations are, therefore, the outcome of such score or performance interpretation leading to decision-making about students' learning achievement. In language

assessment, *norm-referenced* and *criterion-referenced interpretations* have been considered and adopted in educational assessment (Brown, 2005, 2014a, 2014b). Teachers need to understand the fundamental concepts and distinctions between these two types of interpretations because they play a crucial role in the decision-making of student learning.

The term 'norm' is derived from the concept of the *normal distribution* in statistics that has a bell-curved shape when all scores are displayed together. The middle of the distribution represents the average scores of all students. In the normal distribution, many students will be spread and placed around the average score, with few at the left and right end of the distribution. Some will perform very poorly (e.g., low proficiency level) and exceptionally well (e.g., high proficiency level), while many perform at an average level. Based on this principle, the norm-referenced interpretation sees students' performance relative to others. That is, students are ranked in order. A norm-referenced interpretation is helpful when there is a need to select students or applicants, especially when a few places or positions are available (e.g., those in the top 5% are selected).

The term 'criterion' is derived from *standards* that determine students' performance levels. In a real-life context, we can think of a situation in which people take a driving license test, which measures driving knowledge and skills, such as speed management, road positioning, decision-making, responding to hazards, and vehicle control. If driving test candidates competently demonstrate these skills, they should be able to pass the driving test, thereby obtaining a driving license. The decision is made on a pass-fail basis. The driving test official is not interested in relating different candidates' performances to one another before deciding whether they should be granted a driving license. In this example, the standards for judging whether people should pass the driving test are used as the criteria for decision-making. Hence, this kind of interpretation is known as criterion-referenced.

In classroom language teaching, standards or criteria are commonly related to the learning outcomes and specific skills included in a given language syllabus or those taught across different lessons. For example, if a criterion in reading comprehension skills is "Students can correctly identify the main topic of a written text" and if students can demonstrate that they can do so correctly at a required level (e.g., at least 70% correct), they have met this criterion or standard.

Brown (2005) further articulates that a norm-referenced interpretation leads to *relative* decision-making (i.e., "a student's performance is ordered and compared to those of other students in percentile terms" p. 3). In contrast, a criterion-referenced interpretation results in *absolute* decision-making (i.e., "a student's performance is compared only to the amount, or percentage, of material learned" in the given classroom, p. 3). The notion of *percentile* is related to the place of students relative to all students. That is, a given percentile of a student indicates the proportion of students who score above or below this student. For example, students with a *percentile* score of 80 perform better than 80 out of 100 but worse than 20 out of 100. By contrast, the notion of percentage is related to the degree to which a given student has fulfilled the criteria or standards (i.e., the percentage of what they have achieved in the classroom). If the syllabus specifies that students should do at least 80% of the overall assessment tasks, those who have achieved this percentage should pass the course. Teachers do not need to compare students' performance to one another to pass or fail them. All students who have achieved at least 80% should pass the course.

Brown (2005) has stressed that a criterion-referenced interpretation is more suitable for classroom assessment of learning which mainly measures "well-defined and fairly specific instructional objectives" (p. 2). A norm-referenced interpretation, by contrast, is suitable for standardised language proficiency tests because it "measures global language abilities" (p. 2). Brown (2005) profoundly stresses that "in order to test appropriately, administrators and teachers must be very clear about their purpose for making a given decision and then match the correct type of test to that purpose" (p. 7).

Golden Rule #3: Consider and Apply the Basic Assumptions of Classical Test Theory for Analysis of the Assessment of Learning

According to Brown (2012e, 2014a, 2014b, 2022), classical test theory is a measurement theory developed in the twentieth century (hence classical) that has a set of assumptions about test scores or observed performance and factors that influence them (namely error). Brown (2022) points out that while it has been overshadowed by other measurement theories (e.g., criterion-referenced testing, generalizability theory, and item response theory), classical test theory is not a thing of the past, as often believed by some people. He reassures that it is "alive and well" (p. 450) and is "the dominant psychometric theory actually applied to the problems of language testing in real-world language teaching situations" (p. 449).

According to Brown (2005), classical test theory can define students' observed scores or performance from a given classroom assessment task (100%) as composed of two types of scores. The first is *a true score* that reflects their actual or target ability described in the learning outcomes gathered by the test or assessment task. The second is an error score resulting from interferences of factors unrelated to the intended interest ability. Classical test theory specifies various sources of error in measurement (Brown, 2005, 2022). Some of these are systematic, pre-identifiable, and controllable when known as they are part of the testing or assessment program, and others are purely random or unpredictable. Sources of score or performance errors are, for example, the testing or assessment environments (e.g., noise and room temperature that interfere with students' test-taking processes), administration procedures (e.g., instructions and instruments), test and assessment types and methods (e.g., types of responses, techniques or tasks, quality of questions or tasks, time allowance, and test security), scoring methods (e.g., scoring or rating procedures, quality of the rating scales, rater bias), and students' factors (e.g., health and motivation), to name but a few.

Classical test theory assumes that students' scores or performance from any given test or assessment task are likely to contain errors that affect interpretation and decision-making accuracy or trustworthiness. Such errors need to be reduced as much as possible. Based on classical test theory, if score variation is largely based on the true score, little error will affect the score or performance interpretation. This situation helps teachers make sound decisions on students' learning or achievement levels. However, if score variation comprises a large proportion of error, score or performance interpretation will be significantly affected. This situation does not help teachers make sound and fair decisions for their students.

Therefore, several measures and guidelines have been developed to help eliminate or minimise errors from testing and assessment (see Brown, 2005). For example, various sources of errors presented above are considered and reduced before the assessment administration. The five golden rules presented in this article can help minimise the influences of errors

interfering with students' test scores or assessment performance. Other analytic measures include several types of reliability analysis (e.g., Spearman-Brown prophecy formula and K-R20 and K-R21 formulas). In classical test theory, *reliability* can be understood as the measure of score or scoring consistency. It is related to the concept of precision. If a test score is precisely consistent with the target learning outcome or ability, a given test or assessment is considered reliable since errors are marginal. Different types of test responses (e.g., selected versus constructed response tasks) require different types of reliability analysis and item and faculty analysis. For example, in selected-response questions, reliability analysis is performed on students' responses, whereas in constructed-response tasks such as an essay task that require subjective judgments, reliability analysis is performed on raters' scores (e.g., intra-rater and inter-rater reliability; see Brown, 2014a). In selected-response tasks, *item facility and discrimination analysis* can identify the extent to which a given question is easy or difficult and whether it effectively discriminates students with different abilities (e.g., students who have mastered the learning outcome can answer it correctly, whereas those who have not mastered it should not be able to answer it correctly).

Classroom assessment of language learning can apply the core principles of classical test theory to reduce errors on tests or assessment tasks. However, since teachers are likely to use a criterion-referenced interpretation of their students' scores or performance when making decisions about their achievement levels, it is also useful to understand the concept of decision *dependability*, analogous to reliability in classical test theory. Dependability is developed separately from classical test theory (part of the *generalisability theory*). Classroom assessment is associated with the consistency of teachers' decisions in passing or failing students based on students' performance, for example, through a cut point (see Brown, 2014b, Fulcher, 2010). Various dependability analyses have been developed to ensure decision dependability, for example, threshold loss agreement approaches, squared error loss agreement approaches, and kappa estimates.

Golden Rule #4: Develop Well-defined Assessment Criteria and Rubrics for Judging Students' Performance

Assessment is an integral part of syllabus design or curriculum development. Assessment for and of learning ensures that students learn what they are supposed to know and are learning successfully. Given this, well-defined assessment criteria are needed for teachers to use and students to guide their language learning in the classroom. Classroom assessment of learning is likely to include a combination of various language tests or tasks, so multiple sets of specific criteria need to be defined and concretely developed.

Rubrics can be used in both norm- and criterion-referenced interpretations for judging students' test scores or assessment performance (Brown, 2012b). Brown (2012c) published a comprehensive edited volume that promotes using and developing rubrics for various classroom language assessment contexts. Brown (2012d, p. 1) defines a rubric as:

> a grid set up in one of the two ways: (a) with scores along one axis of the grid and language behavior descriptors inside the grid for what each score means in terms of language performance … or (b) with language categories along one axis and scores along

the other axis and language behaviors descriptors inside the grid for what each score within each category means in terms of language performance.

Brown (2012d) points out that based on the features of such rubrics, teachers can use a rubric for scoring students' language abilities or skills and for providing feedback on their progress in learning them. Brown (2012b) provides a comprehensive assessment rubric guideline. He formulates three questions for guiding teachers when constructing an assessment rubric. We answer these questions rather briefly below.

What is the Difference Between Holistic and Analytic Rubrics?

Based on Brown (2012b), a *holistic rubric* uses a single general scale (e.g., 1 to 5; see the (a) definition of a rubric above) to create a global score for students' performance, whereas an *analytic rubric* gives separate scores for different aspects of students' performance (e.g., organisation, logical development of ideas, grammar, punctuation, spelling, and mechanics; see the (a) definition of a rubric above). Although holistic and analytic rubrics can be used in a complementary way in classroom assessment (e.g., some simple tasks can be assessed using a holistic rubric while some more complex tasks can be assessed using an analytic rubric), teachers need to decide which they will use in advance since it is not practical to use both types for one assessment task. It is important to know the advantages and disadvantages of these rubrics. Brown (2012b) discusses some advantages and disadvantages of both rubrics. For example, a holistic rubric is easier and quicker to use than an analytic rubric, so it is often used in standardised language proficiency tests due to the need to produce test scores timely. An analytic rubric is more time-consuming than a holistic rubric, but it is useful for feedback provision to students, so it is suitable for classroom assessment. Decisions on the type of rubrics should, therefore, be made early.

What are the Steps in the Rubric Development Process?

Brown (2012b) also presents six major stages needed for developing an effective rubric:

1. *Planning* includes defining an assessment purpose or goal, considering classroom materials and activities, brainstorming ideas with colleagues (if available), deciding which type of rubrics is suitable, choosing aspects of language or criteria for judging performance levels (e.g., task fulfillment, development of ideas, organisation, language use, and mechanics), and deciding the range of scale (e.g., 1 to 5, which can describe students' language responses or levels).
2. *Designing the rubric* is related to creating rubrics based on the planning stage. This includes creating grids that place scores on the vertical axis and assessment aspects on the horizontal axis. Associated descriptive information (also referred to as a descriptor) describes a progressive criterion level placed in each score grid. For example, descriptors for an essay 'organisation' criterion are: (1) poorly organised, (2) not very well-organised, (3) somewhat organised, (4) fairly well-organised, and (5) well-organised.
3. *Planning the assessment procedures and using the rubric* is related to a broad assessment development (e.g., choosing a task or prompt to elicit performance, writing clear instructions about what students are supposed to do, arranging a date, time, and place for students to complete the assessment, assessment administration, and training teachers or

raters to use the rubrics). The focus of training is to ensure that raters can consistently and similarly differentiate students' responses based on descriptors in the rubric. Referring to McNamara (1996), Brown (2012b) noted that raters can differ in their scores given their individual differences. Rater training should not be about forcing raters to be unnaturally homogeneous in their ratings. Very harsh, lenient or unreliable raters may be realised during training, prompting an action to resolve (e.g., by showing ratings by an experienced or expert rater and reasons for a score given).

4. *Evaluating the reliability/fairness of the rubrics*: Reliability is consistency in scoring. *Intrarater reliability* is a rater's consistency in assigning a score to students based on different descriptors of a given criterion in the rubric (e.g., organisation). Intrarater reliability can be described as internal consistency of a given rater (e.g., the same score is assigned a second time). *Interrater reliability* refers to the agreement between two raters in providing a score to the same student based on a given criterion. Statistical calculations such as agreement rates and Cohen's kappa can be calculated for investigating interrater reliability.
5. *Evaluating the quality of the rubric* includes examining the level of appropriateness of the rubric based on experience of rating students' responses. The purpose of rubric evaluation is to judge if it is suitable for the assessment purpose and usable for a given situation. Brown (2012b, p. 27) provided a list of guided questions for evaluating a rubric, for example: (a) is this rubric appropriate for the assessment purpose?; (b) is it too specific or too general?; (c) are all descriptors clearly stated; and (d) are there missing descriptors that illustrate students' performance. In terms of rubric usability, questions to be asked include: (a) is it user-friendly (e.g., in terms of layouts and spaces for comments); and (b) is the wording of the descriptors appropriate for the target students (e.g., according to ages, educational levels, and cultural background).
6. *Planning feedback and revising for periodically useful ratings* include explaining the meaning of the scores students receive. It is useful to show students the scores they receive in relation to the rubric used. An analytic rubric is useful for feedback on specific areas (e.g., content, cohesion, and organisation) because students will know in which areas they are strong or weak. Feedback may be provided as an overall written or oral impression of students' responses. The last stage in this cycle is revisiting the rubric to check for any updates.

What Resources Can Help in Creating Rubrics?

While Brown (2012b) is hesitant to include internet websites for recommending resources for creating rubrics as they are prone to disappearing over the long term, the following websites with updated URLs remain active at the time of this article writing.

- Teachnology: https://www.teach-nology.com/web_tools/rubrics/
- Rubistar: http://rubistar.4teachers.org/index.php

Brown (2012b) strongly encourages teachers to use rubrics in their teaching and assessment and remarks that:

> Rubrics are an essential tool for all language teachers in this age of communicative and task-based teaching and assessment – a tool that allows us to efficiently communicate

> with our students what we are looking for in the productive language abilities of speaking and writing and then effectively assess those abilities when the time comes for giving students feedback, for grading, for placement into new courses, and so forth (p. 7).

Golden Rule #5: Use Assessment of Language Learning as a Source of Feedback to Improve Student Learning and Existing and Future Assessment Tasks

Our fifth golden rule is about using feedback based on assessment tasks. Students should learn something from participating in language assessment activities (Turner & Purpura, 2017). Brown (2013) points out that what distinguishes assessment from classroom activities is that it furnishes purposeful feedback that can help shed light on the effectiveness of language teaching and learning. We endorse this viewpoint and intend to promote the use of feedback for improving both student learning and assessment use.

Although students receive feedback on their learning in the classroom (e.g., assessment for learning), feedback from assessment of learning is invaluable for students. According to Brown and Trace (2017), students can be provided feedback in the forms of numerical scores, scales, or letter grades, which provide comprehensible numerical information about the level of their performance success (e.g., high scores mean that they have correctly and appropriately addressed the given test or assessment tasks, whereas low scores indicate that there are problems with their performance that should be improved). Therefore, accurate scores are critical to language assessment because they allow students to be realistic about their current knowledge, ability, or skill level. Providing an average score to students is also useful as it will enable them to understand their performance relative to other students (e.g., Brown, 2009). Numerical feedback can help students reinforce their learning and develop a step to advance their language skills. Numerical feedback can appeal to students when scores can be mapped out over time, giving them a reference to their ongoing and longitudinal sense of their learning achievement (Brown & Trace, 2017).

Nonetheless, numerical feedback alone cannot help students improve their learning or performance. Without teachers' assistance to make it meaningful to students, numerical feedback does not go much further. In our experiences as language teachers and educators, we have learnt that in many classroom situations, after students are informed about their test scores or grades, they rarely receive further constructive feedback that supports their language learning. Brown and Trace (2017) also pointed out the need to provide summative feedback for students. Students should be able to make sense of test scores or performance at least, so giving them just their scores is not sufficient. Therefore, qualitative feedback such as oral or more detailed written comments (e.g., reasons for their scores) is essential as it informs students about their strengths and weaknesses and socially connects them to teachers (Brown, 2013; Brown & Trace, 2017).

Different types of feedback can be considered after classroom assessment of learning. For example, *corrective feedback* provides information about the nature of accuracy, fluency, and appropriacy of language performance. That is, students can gain specific information about how to improve their language use, for instance. *Explanatory feedback* utilises some positive feedback on students' performance, but highlights what they fall short in regarding their performance. *Metacognitive feedback* provides information about their cognitive processing ability and its suitability for performing a given assessment task or how it inhibits task

completion. For example, students may learn that they need to improve their essay outline before completing an essay task. They may realise that their way of writing is prone to plagiarism issues. Brown and Trace (2017) postulate that "combined with feedback after the assessment, learners can then evaluate how well their particular learning strategies worked in preparing them for the test, which in turn can help them determine whether or not to continue using those strategies" (p. 499).

In addition to using assessment of learning as a feedback source for students, teachers can use it to improve their teaching approaches and assessment methods (e.g., Brown, 2013; Brown & Trace, 2017). An obvious benefit of classroom assessment of learning is information about students' scores or performance that teachers can use to confirm their teaching practice (e.g., when their students perform well) as well as inform decisions about what and how to improve their teaching and assessment approaches further (e.g., refining the current language syllabus, classroom activities, and current assessment practices). Golden rule # 3, for example, is particularly beneficial for teachers to realise what may contribute to errors in their assessment. Reliability and dependability analysis can prompt teachers to revisit earlier test or assessment development stages by correcting or revising previously developed assessments and creating new ones.

We agree with Brown and Trace (2017, p. 502) that "sound classroom assessment is much more than just scores and grades, but rather is a tool that teachers can use to support and promote their students' learning" and "teachers should seriously consider ways to use their assessment items and tasks to affect their students' future performances and displays of language ability".

Conclusion

Assessment is necessary for good teaching and learning. We have learnt from JD Brown that it is also about transforming and improving students' lives. Classroom language assessment is not just about designing, creating, and using testing and assessment tools for a given intended purpose. At the heart of classroom language assessment (whether formative or summative assessment) should be its role and influence in promoting equality and social justice for students. Assessment is at the centre stage of the language classroom. It requires teachers' and students' involvement in its design and practice. We would like to conclude our article with this poem by one of our Language Testing and Assessment students --Cassandra Polden, who permitted us to publish her poem that captures critical elements of the five golden rules.

How on earth do we hope to know
One's capabilities and proficiency from the get-go
Informal, formal; the assessor may choose
What may provide the most holistic views
Awareness of options aligned with purpose
Formative, summative; designed to inform and motivate, not hurt us
Teacher's judgment employed with best intentions
If necessary can provide suitable targeted interventions.

ORCID

https://orcid.org/0000-0002-7929-8924

https://orcid.org/0009-0003-6763-602X

Acknowledgements

We would like to thank the Editors for their feedback on the earlier drafts of this article. It helped us clarify our content substantively. Thanks also go to our Language Testing and Assessment (EDPJ5026) students at The University of Sydney for their active participation that led us to conceptualise ideas about the five golden rules that are linked to JD Brown's contributions.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

Bachman, L. F., & Damböck, B. (2017). *Language assessment for classroom teachers*. Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Black, P., & Wiliam, D. (2018) Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practice* (3rd ed.). McGraw-Hill.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. McGraw-Hill.

Brown, J. D. (2009). Using a spreadsheet program to record, organize, analyze, and understand your classroom assessments. In C. Coombe, P. Davidson, & D. Lloyd (Eds.), *The fundamentals of language assessment: A practical guide for teachers* (2nd ed.) (pp. 59–70). TESOL Arabia.

Brown, J. D. (2012a). Choosing the right type of assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *Cambridge guide to second language assessment* (pp. 133–9). Cambridge University Press.

Brown, J. D. (2012b). Developing rubrics for language assessment. In J. D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages (pp. 13–31).* National Foreign Language Resource Center.

Brown, J. D. (Ed.) (2012c). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. National Foreign Language Resource Center.

Brown, J. D. (2012d). Introduction to rubric-based assessment. In J. D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages (pp. 1-9)*. National Foreign Language Resource Center.

Brown, J. D. (2012e). What teachers need to know about test analysis. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *Cambridge guide to second language assessment* (pp. 105–112). Cambridge University Press.

Brown, J. D. (Ed.) (2013). *New ways of classroom assessment, revised*. Teachers of English to Speakers of Other Languages.

Brown, J. D. (2014a). Classical test reliability. In A. J. Kunnan (Ed), *Companion to Language Assessment*, John Wiley & Sons. https://doi-org.ezproxy.library.sydney.edu.au/10.1002/9781118411360.wbcla054

Brown, J. D. (2014b). Score dependability and decision consistency. In A. J. Kunnan (Ed), *Companion to Language Assessment*. John Wiley & Sons. https://doi-org.ezproxy.library.sydney.edu.au/10.1002/9781118411360.wbcla085

Brown, J. D. (2022). Classical test theory. In G. Fulcher & L. Harding (Eds.), *Routledge handbook of language testing* (2nd ed., pp. 447-461). Routledge.

Brown, J. D., & Trace, J. (2017). Fifteen ways to improve classroom assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning Volume III* (pp. 490-505). Routledge.

Fulcher, G. (2010). *Practical language testing*. Routledge.

Green, A. (2020). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Assessment Quarterly*, 4(1), 6–36. http://doi.org/10.1080/15434300701348318.

McNamara, T. (1996). *Measuring second language performance*. Longman.

Phakiti, A. (2018). Assessing higher-order thinking skills in language learning. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching*. Wiley.

Phakiti, A., & Leung, C. (forthcoming, 2024). *Assessment for language teaching*. Cambridge University Press.

Popham, W. J. (2017). *Classroom assessment: What teachers need to know* (8th ed.). Pearson.

Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N.H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 257–71). Springer.

Turner, C. E. (2012). Classroom assessment. In G. Fulcher & F. Davidson, *Routledge handbook of language testing* (pp. 65–78). Routledge.

Turner, C. E., & Purpura, J. E. (2017). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255-273). De Gruyter Mouton.