Spring 5-20-2016

Quantifying the Qualitative: Increasing the Reliability of Subjective Language Assessments

Jeremy W. Armes
laughingman8@gmail.com

Follow this and additional works at: https://repository.usfca.edu/capstone

Part of the Curriculum and Instruction Commons, Elementary Education and Teaching Commons, International and Comparative Education Commons, and the Other Teacher Education and Professional Development Commons

Recommended Citation
https://repository.usfca.edu/capstone/338

This Project/Capstone is brought to you for free and open access by the Theses, Dissertations, Capstones and Projects at USF Scholarship: a digital repository @ Gleeson Library | Geschke Center. It has been accepted for inclusion in Master's Projects and Capstones by an authorized administrator of USF Scholarship: a digital repository @ Gleeson Library | Geschke Center. For more information, please contact repository@usfca.edu.
University of San Francisco

Quantifying the Qualitative:
Increasing the Reliability of Subjective Language Assessments

A Field Project Proposal Presented to
The Faculty of the School of Education
International and Multicultural Education Department

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts in Teaching English as a Second Language

By
Jeremy W. Armes
May 2016
Quantifying the Qualitative:
Increasing the Reliability of Subjective Language Assessments

In Partial Fulfillment of the Requirements for the Degree

MASTER OF ARTS

in

TEACHING ENGLISH AS A SECOND LANGUAGE

by
Jeremy W. Armes
May 2016

UNIVERSITY OF SAN FRANCISCO

Under the guidance and approval of the committee, and approval by all the members, this field project has been accepted in partial fulfillment of the requirements for the degree.

Approved:

_____________________________  ______________________
Instructor/Chairperson              Date
TABLE OF CONTENTS

Acknowledgements..........................................................................................................iv

Abstract .............................................................................................................................v

Chapter I – Introduction......................................................................................................1

   Statement of the Problem..................................................................................................2
   Purpose of the Project .........................................................................................................3
   Theoretical Framework ......................................................................................................4
   Significance of the Project ..................................................................................................5
   Limitations of the Project ..................................................................................................6
   Definition of Terms ...........................................................................................................7

Chapter II – Review of the Literature................................................................................9

   Introduction.......................................................................................................................10
   Rating Procedures .............................................................................................................11
   Rater Behaviors ................................................................................................................14
   Task Factors ......................................................................................................................16
   Student Factors ................................................................................................................17
   Conclusion .........................................................................................................................19

Chapter III – The Project and Its Development ...............................................................22

   Brief Description of the Project .......................................................................................23
   Development of the Project ...............................................................................................24

Chapter IV – Conclusions and Recommendations .......................................................26

   Conclusions .......................................................................................................................27
   Recommendations ............................................................................................................28

References ...........................................................................................................................29

Appendixes .........................................................................................................................31

   The Project ......................................................................................................................1
   Introduction .......................................................................................................................3
   Things to look for ..............................................................................................................5
   Module I ............................................................................................................................6
   Module II ..........................................................................................................................17
   Module III .........................................................................................................................24
Module I slideshow .................................................................31
Module II slideshow ..............................................................43
Module III slideshow .............................................................50
ACKNOWLEDGEMENTS

This project would not have been possible without the direction, assistance, and encouragement of Dr. Sedique Popal. He has been a profound influence on my academic and professional career. His content knowledge, teaching techniques, and sense of humor always inspired me to do more than I thought I was capable of. I will forever be indebted to him, and hope one day to have the same kind of impact on my students.
ABSTRACT
This project explores how to increase the reliability of subjective language assessments. What are the major factors that affect reliability, and what can educators do to maximize it? Research in the field suggests four key factors that have the greatest effect on reliability. The first factor is the procedures used to administer and score the assessment. The more standardized the process is, the higher the reliability will be. The second factor is rater experience, both in the profession and with the assessment in question. Experience is positively correlated with reliability. Novice teachers who get training and feedback in a scoring rubric can become almost as reliable as veteran educators, so training cycles should be a component of any assessment program. The third factor that affects reliability is the type of task the students are asked to do. Tasks that integrate more than one language skill are more reliable than those that use only one skill. The fourth factor is student experience with the testing format. More experience is positively correlated with better achievement, so educators must make sure their students are familiar with the tasks they are asked to do.
Chapter 1

Introduction

Statement of the Problem

Purpose of the Project

Theoretical Framework

Significance of the Project

Limitations

Definition of Terms
CHAPTER 1
INTRODUCTION

Statement of the Problem

Subjective assessments can not be quantified in the same way as objective assessments. Scores on subjective assessments can vary dramatically depending on the rater’s experience, training, and bias (Attali, 2015). Novice and experienced teachers are often unaware of their own biases, making them difficult to correct (Schaefer, 2008). This means that scores on subjective assessments is not always reliable, and that any teaching decisions made from them may be suspect. Institutions and educators need to find ways to ensure the reliability of their subjective assessments.

There are ways to increase the reliability of scoring on subjective assessments. Training in the rubric, and experience grading a test, increase reliability significantly (Kim, H-J 2015). Cycles of training, scoring, and feedback can make novice raters almost as reliable as veteran teachers, and can make veteran teachers even more reliable (Barkaoui, 2010).

Raters are not the only source of score unreliability. The type and number of tasks students are asked to do can have a measurable effect on performance (Gebril, 2009). The types of rating scales can affect scoring (Barkaoui, 2010). Procedures for test administration affect the outcome of tests (Cassady, 2005). How a test is used may conflict with how it was originally intended, leading to negative washback for teachers and students (Luxia, 2005). Educators make instructional decisions based on the results of their assessments, so it is vital to ensure that the results are accurate and useful.

There is a need for a comprehensive, easy to understand guide to making subjective assessments as reliable and valid as possible.
Purpose of the Project

This guide will help teachers and administrators to understand the strengths and limitations of their tools, and help them to make changes in administration and scoring to make them even more valid and reliable. The end result of this project will be a handbook for teachers and administrators to use in quantifying their qualitative assessments. It will consist of four parts. The first part will be a survey of current research in increasing assessment reliability. The purpose of this section is to give teachers and administrators the background knowledge they will need to make informed decisions about their training in and administration of assessments. In addition, it will provide a common vocabulary for educators to use when discussing this topic. It will serve as an introduction for new teachers, and a review for more experienced ones. It will be broken into two parts, discussing issues with the reliability of tests themselves, and the educators who rate them.

The second section addresses how to increase inter-rater reliability through training and testing experience. Raters who go through regular cycles of feedback and training become much more reliable, whether they are novices or veteran educators.

The third and final section will discuss ways to modify assessment tasks and procedures to increase their validity. This will focus attention on reducing potential sources of bias, as well as student anxiety. How to change independent tasks into integrative ones will take up a large portion of this section, since integrated tasks are a more reliable indicator of language proficiency.

The primary audience for this project will be teachers and administrators in ESL programs at all levels. While the author’s experience is in primary education, he hopes to create a
product that will be useful at all levels of education. He also hopes that it can be useful to institutions that are creating their own assessments.

Theoretical Framework

Communicative Language Teaching (CLT) is the primary theoretical framework behind this project. According to CLT the primary purpose of language teaching and learning is to be able to communicate competently and effectively in the target language. This approach was developed in the 1970s and 1980s by Dell Hymes in response to concerns that students could do very well on tests and in the classroom, but were unable to communicate in authentic situations outside of the classroom. Researchers came to understand that linguistic competence requires more than simply mastering linguistic structures. CLT emphasizes language functions over language forms (Larsen-Freeman, pp. 122-31).

There are a few things to keep in mind when using CLT. Materials should be authentic. That is, they should reflect how language is used in real-world situations. Newspapers, broadcasts, role-plays, and language games are all examples of authentic materials that are used in CLT methods. CLT is most often used to justify teaching techniques, and rarely for assessments. The author feels that the principles of CLT form a strong foundation for assessment creation, as well. If authentic communication is the goal of language teaching and learning, then assessments should include opportunities to demonstrate that(Larsen-Freeman, p. 133).

Howard Gardner’s theory of multiple intelligences is also useful to consider when designing and interpreting assessment data. His theory states that people have different learning styles, with preferences towards certain kinds of input. A visual learner would get more information from a written text or a graphic than from listening to a lecture. Someone with a preference for auditory information would learn more from listening than reading. We must
remember this when designing assessment tasks, and use multiple measures to make our decisions. A written test may not give us useful information about a student’s proficiency, if they have a strong preference for verbal information. We must always look to assess our students in the ways that they learn best, so that we can get reliable information from them.

Significance of the Project

Some skills must be assessed qualitatively. Writing, speaking, and listening tests rely on subjective judgements to determine performance. Students’ performance on these tests determines what classes they are sorted into, what educational interventions they receive, and whether or not they meet proficiency standards. These are not abstract concerns for these students, but the difference between success and failure in school or work. Teachers need to make sure that their subjective judgements are reliable and fair.

In an era where students’ performance on assessment is being used to evaluate the quality of a teacher, it is vital that the assessments reflect the actual knowledge of the student. Unreliable scores make valid comparisons between teachers and teaching methods nearly impossible.

If teachers who work with EL students had a resource available that could help them reduce or eliminate the unreliability of the assessments they are using, their students could benefit dramatically. Teachers would be able to accurately assess their students abilities, and plan targeted interventions. In addition, educators would be able to use this tool to push for necessary changes to their testing regime. They would have the knowledge and vocabulary to effectively describe the problems with a given assessment, as well as recommend needed changes.

Limitations
This project is not without its limitations. The parts dealing with testing circumstances and administration will be useful for teachers who have to give any kind of assessments. Standardizing how tests are given helps make scores from different testing sessions comparable and generalizable. However, because this project concerns itself primarily with subjective assessments, it will not be as useful for objective tests. A large part of the project is concerned with increasing the reliability of raters by reducing their subjective biases. There are no subjective factors to scoring a multiple choice test or true/false tests, so this project is of little to no use for teachers that use those kinds of tests exclusively.

Furthermore, the recommendations in the project may not be able to be implemented by institutions. Educators may not have a choice in what kinds of assessment tasks they give, or how they give them. There is very little assistance this project will be able to give to people in those situations.

A final limitation to the project is that there are no “one-size-fits-all” solutions to the problems of subjective assessments. Every institution has its own unique needs. Every teacher has their own strengths, weaknesses, and biases. The very nature of a subjective assessment is that it cannot be quantified. Readers must view the author’s recommendations as guidelines for improving their practice, and not as a set of iron-clad rules. They will need to adapt what is relevant for their situation, and disregard what is not.

Definition of Terms

Achievement test: An evaluation of whether learning objectives have been met. This is generally given at the end of a unit or course (Brown, p. 346).
<table>
<thead>
<tr>
<th><strong>Analytic Scale:</strong></th>
<th>A rating scale that gives individual scores to different aspects of a student’s performance (Barkaoui, 2010, p. 55).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Assessment:</strong></td>
<td>The process of deterring a student’s competence by using multiple measures. This is an ongoing process (Brown, p. 346).</td>
</tr>
<tr>
<td><strong>Authenticity:</strong></td>
<td>The correlation between test items and language tasks being tested. Assessments must be authentic to be useful to teacher and student (Brown, p. 346).</td>
</tr>
<tr>
<td><strong>Competence:</strong></td>
<td>Underlying ability to use a language or language skill, often contrasted with performance (Brown, p. 347).</td>
</tr>
<tr>
<td><strong>Diagnostic test:</strong></td>
<td>A test designed to evaluate a specific language skill or aspect (Brown, p. 348).</td>
</tr>
<tr>
<td><strong>Formative:</strong></td>
<td>An assessment that evaluates students’ skills in order to help them grow. These are usually given at the beginning of a course or unit. (Brown, p. 348).</td>
</tr>
<tr>
<td><strong>Holistic scale:</strong></td>
<td>A rating scale that takes all aspects of student performance into account to give a single score. (Barkaoui, 2010 p. 55)</td>
</tr>
<tr>
<td><strong>Performance:</strong></td>
<td>Actual use of language skills in context, usually contrasted with competence (Brown, p. 351).</td>
</tr>
<tr>
<td><strong>Placement Test:</strong></td>
<td>A test designed to sort students into leveled language programs (Brown, p. 351).</td>
</tr>
<tr>
<td><strong>Practicality:</strong></td>
<td>A measure of how easy a test is to administer. It includes cost, administration, ease of use, and scoring procedures. A test’s usefulness is strongly linked to its practicality (Bachman, 2005 p. 25).</td>
</tr>
<tr>
<td><strong>Proficiency Test:</strong></td>
<td>A test designed to measure a student’s overall language ability outside of the demands of a particular curriculum, course, or language skill.</td>
</tr>
</tbody>
</table>
**Reliability:** A measure of how consistent test results are, and how accurately they measure student performance (Rezaei, 2010 p 19).

**Rubric:** A guide to rating subjective assessment that consists of descriptive performance targets for each band (Rezaei, 2010 p 19).

**Summative:** An assessment that evaluates student’s growth over time, usually given at the end of a course (Brown, p 353).

**Test:** A subset of assessment, testing a specific skill at a discrete point in time.

**Validity:** A measure of how well the test measures what it is supposed to measure, and how useful it is for the purpose of assessment. There are many kinds of validity, including **construct** does it have a model that explains how to interpret results?), **content** (does the test assess what the students have been learning?), **face** (does the test-taker view the test as fair and useful?), and **predictive** (how accurate the test is in predicting future performance) (Brown, p. 353).

**Washback:** A measure of how effective the test’s feedback is to students and assessors. Washback helps teachers and learners evaluate their teaching and learning (Luxia, 2005 p. 143).
Chapter II
Review of Literature

Overview

Rating Procedures

Rater Behaviors

Task Factors

Student Factors

Conclusion

Bibliography
CHAPTER II
REVIEW OF THE LITERATURE

Introduction

The usefulness of an assessment increases as its reliability, construct validity, and practicality do. Arguments for test use rely on the explicit links between score interpretations and teaching decisions made based on them. Therefore, teachers and administrators must ensure that the assessments they use, and the procedures they use to administer them, are as relevant to the course content as possible. A test may be used for more than one teaching decision if it aligns well with course objectives. However, a highly practical and valid test may not be reliable if it is used for a purpose other than what it was designed for (Bachman, 2005).

The interpretation of test results affects classroom instruction as well as teacher and program evaluation. The purpose of an assessment must be clear to educators and students alike. Misuse of an assessment leads to negative washback that affects students and programs in unintended and often negative ways. For example, Qi Luxia found that teachers and administrators in China were using a summative evaluation, the National Matriculation English Test (NMET), as an admission test. The original intent of the NMET was to reform Chinese secondary ELT to align more closely with communicative language teaching. However, the selection function for the test undermined its intended role as an agent of educational change. Teachers viewed it as a grammar test rather than an integrated skills test, and taught largely grammar and test preparation, rather than communication. The selection function was so powerful that teachers and schools were evaluated based on student performance on the NMET, which gave an even stronger incentive to teach to the test rather than change teaching approaches. This eliminated and positive washback from the exam, and is a cautionary tale for
educators. The washback from an assessment should affect teaching and learning in positive ways. Teachers should be able to use the results of an assessment to plan instructional interventions. Students should be able to use the results to identify their areas of strength and weakness. Washback can be beneficial to teachers and students, as long as the purpose of the test aligns with the nature of the course. If there are conflicts between them, the usefulness of the test is compromised. Marking schemes must reflect the function of the test as well as the content of the course. There must be clear communication between test designers, test administrators, and students in order to facilitate this alignment (Luxia, 2005).

In order to maximize the usefulness of an assessment, educators must understand what variables affect reliability, validity, and practicality. The variables can be broadly categorized into four areas. First, how rating procedures affect the scoring of the task. Second, how rater experience, training, and biases can affect the scoring of assessments (Kim, H-J, 2015; Attali, 2015). Third, how the type of task affects student performance. In general, integrated tasks are more reliable and valid than independent tasks (Lee, H-K 2007, Lee, H-W, 2006). Fourth, how examinees’ background experience and language proficiency affect the reliability of scores. In general, language proficiency and background knowledge are correlated with better performance (Lee, H-K, 2007). Understanding these factors, and learning to adjust them in response to feedback, will allow educators to make the best possible decisions for their students.

Rating Procedures

The usefulness of a test is strongly linked to how practical it is to administer (Bachman, 2005, p 25). Tests that are practical in administration and scoring are more likely to be used as intended, giving teachers and students useful feedback. Two major factors affect test practicality:
the number of raters for a given task, and the procedures used to score the test (Brown, H.D. 2010).

Adding additional raters to a task increases the time and cost to administer it, which negatively affects its practicality. Research tends to agree that adding raters increases reliability significantly, even if the raters are novices or not well-trained in the rating scale (Barkaoui, K. 2010). Most studies find a significant increase in reliability when the number of raters on a given task goes from 1 to 2. While there can be additional benefits from adding a third and fourth rater, the marginal improvement is small. There is little to no benefit from increasing the number of raters past four (Gebril, A. 2009, pp. 509-511). These findings suggest that two raters for a task is optimal for practicality considerations, with up to four if the institution can afford it.

Performance bands and rubric scales can be useful in diagnosing and monitoring student progress. They allow for comparisons to be made across students and classrooms. There is some debate over how effective they are at providing reliable and consistent results, but most studies agree that the more experienced raters are with the test and the scoring guide, the more consistent and reliable they are at scoring the test. Yigal Attali found that novice and experienced raters of a standardized writing assessment did not differ much in the severity or consistency of their ratings if they were given similar training and feedback. The novice raters improved so much during the training process that the statistical differences between the experienced raters and them was negligible (Attali, Y. 2015). Kim found similar results for a speaking assessment. Novices (having no experience in ESL instruction or assessment), developing raters (having two to three years of experience), and experienced raters (defined as having five or more years of experience in ESL instruction) went through three rounds of scoring, training, and feedback. Each session was separated by a month. After the third session, the novice raters were nearly as consistent and
objective as the experienced raters were in the first session (Kim, H-J. 2015). These findings suggest that training in scoring and immediate feedback are as valuable as years of teaching and grading experience. They also imply that significant gains can be made with a few sessions over a short time, regardless of the teaching experience or educational background of the raters.

The scale used to grade tests can have an effect on the practicality of the scoring process. There are two categories of scales used in scoring, analytic and holistic. Holistic scales consider all parts of task performance to arrive at a single score for a student. Analytic scales rate each element of a task separately, assigning sub-scores to individual traits which are often summed to produce something resembling a holistic score (Barkaoui, K. 2010). Holistic scales are more practical to use, generally requiring less training and expense to use reliably. Analytic scales are less practical to use, requiring more time and expense (Knoch, U. 2011). Educators must balance these practicality considerations with how the scale used affects the rating process.

When novice and experienced raters use holistic scales, they tend to give more weight to rhetoric, language use, and ideas rather than mechanical criteria. Raters using analytic scales will tend to rate all criteria with similar weight, giving a more balanced and reliable assessment of performance (Barkaoui, 2010). These biases can be corrected with regular training and feedback (Kim, H-J, 2015). Educators must keep these findings in mind when deciding what kind of scoring scale to implement, and how to train their raters in the scale.

When deciding what kind of scale to use, it is also important to consider how many performance bands or proficiency levels should be included. If too few are used, the rater may not be able to draw meaningful distinctions between categories. If there are too many levels, the differences between them may be too insignificant for real educational use. Raters can generally distinguish between seven points, plus or minus two (Knoch, 2011). A scale of six points is often
used, and provides enough differentiation to make meaningful teaching decisions (Schafer, 2008). They should be developed to provide information that is relevant to the test and how the performance should be interpreted. Qualifiers such as “consistently” and “often” should be replaced with more empirical distinctions when possible. The more detailed and specific a rubric is, the more reliable it becomes (Knoch, 2011). Raters must be trained in the use of performance bands for maximum effect (Papageorgiou, et al 2015).

Rater Behaviors

In addition to practicality considerations, educators must be aware of the biases that raters themselves bring to the scoring process. Edward Schafer defines bias as the tendency of a rater to rate higher or lower than indicated by a student’s performance. Novice and experienced raters alike have biases towards tests, tasks, student groups, and rating scales (Schafer, E. 2008). Raters can also be biased in their style of rating. Ute Knoch found that raters tend to fall into one of four rating styles. They will assign more weight to either: 1. mechanical errors, 2. topic presentation, 3. personal reaction, or 4. scoring criteria (Knoch, U. 2011). Raters are often unaware of their own biases, which makes bias identification and correction an important part of the evaluation of assessment and rater alike. Anti-bias training can reduce bias and increase intra-rater reliability, though it is unlikely to eliminate bias completely (Schafer, E. 2008).

There are a number of studies that suggest that training in scoring a specific assessment has a stronger effect on rater reliability than teaching experience and educational background. These studies were discussed in the previous section. Training and feedback can correct the biases of novice and experienced raters alike (Barkaoui, 2010; Attali, 2016). There are other variables to consider when discussing how raters affect the interpretation of assessments.
Hyun-Jung Kim summarizes Tom Lumley’s six variables of raters as follows: teaching experience; rater’s experience with the test; training in the rating of the test; experience in rating other tests; educational background of the rater; and internal consistency (Kim, H-J. 2015 p. 241). These variables have been studied extensively in an effort to determine how raters influence the assessment process. In general, teaching and grading experience and educational achievement are positively correlated to reliability and consistency (Kim, H-J 2015). The differences in reliability between novice and experienced raters tend to narrow significantly with training in the rating scale and experience grading a specific test. Training may not eliminate score variability entirely, but is a significant factor in increasing rater reliability (Schafer 2008). Administrators should ensure that regular rubric training and scoring feedback are part of their assessment program, no matter how experienced their educators.

Raters’ experience affects how reliably they use the two kinds of rubrics. Novice raters tend to have more reliable scores when they use an analytic rubric. It forces attention on all areas of a rating scale and gives trainers a useful guide to feedback sessions. Novices do not always use the full range of scoring criteria as much as experienced raters, although this can be corrected with training (Barkaoui, 2010). Experienced raters use holistic scales more reliably and consistently, utilizing their background knowledge and greater experience to arrive at reliable results. Holistic scales feel more authentic to raters, which may affect their attitudes towards the assessment and the rating process (Knoch, 2011). Administrators should consider their raters’ teaching and rating experience when interpreting the scores they give on assessments.

Task Factors

Objective tasks are more practical to administer and easier to score than subjective tasks. Multiple choice tests can be scored by computers, which can run sophisticated analyses of the
results. In H. Douglas Brown’s book, *Language Assessment: Principles and Classroom Practice*, he discusses the strengths of objective tests. These tests have high face validity. Students and teachers consider them reliable as long as test items correlate well with the purpose of the test. Standardized tests go through significant validity and reliability testing. As long as the testing procedures are consistent, the results of such tests can be considered accurate and useful for making educational decisions. Educators that want to design their own objective assessment tasks must carefully design them for maximum validity. Regularly analyzing item facility, item difficulty, and distractors will ensure that a test gives relevant and timely information (Brown, 2010).

Not every skill can be measured using an objective test. Listening, speaking, and writing must be measured subjectively, using rubrics or other scoring guides (Brown, 2010). The nature of a subjective task can have a significant effect on its reliability as an assessment tool. The first kind of task is *independent*, which measures only one language skill (reading, writing, speaking, or listening). Performance on an independent task is strongly correlated with a students’ content knowledge and language proficiency, and not necessarily underlying competence. This can prevent scores from being generalizable across different student populations (Lee, Y-W. 2006).

*Integrated* tasks use some combination of reading/listening with writing/speaking to assess a student. For example, a student may have to listen to or read a brief lesson on a subject, and then write or speak about it. A student’s background knowledge is less of a factor in this kind of task, because the reading or listening selection should contain the information necessary to complete the speaking or writing prompt. This more accurately reflects real-world language use situations, and is a more authentic assessment (Lee, H-K. 2007, Lee, Y-W. 2006). Tasks that
integrate more than one skill are more reliable indicators of true competence. Educators should use integrated tasks whenever possible.

Student Factors

Students bring different levels of background knowledge and language proficiency to assessment tasks. Educators must understand how these factors affect performance on assessments. This topic is complicated and the effects of individual factors are difficult to isolate and interpret. Current research supports two broad generalizations about student background and performance.

First, target language proficiency (as measured by tests such as TOEFL) is usually a better predictor of performance on reading, writing, and speaking tests than background knowledge in a given content area (Lee & Anderson, 2007). This makes intuitive sense, since a student who cannot understand or express themselves in the target language will not be able to demonstrate their knowledge. This effect is stronger in undergraduate students than in graduate students, implying that educational level also affects performance on reading and writing tasks. (Lee & Anderson, 2007).

Second, student performance will vary depending on whether a task is passage-dependent (answers to test items are contained within a reading or listening passage) or passage-independent (answers to test items require independently obtained knowledge). Passage-dependent, or integrated, test items, are usually more reliable at assessing language competence than independent items. The reasons for this were discussed earlier in the task factors section (Lee & Anderson, 2007).

There are certain factors that affect student performance that are independent of language proficiency or background knowledge. Students’ experience with a test format can affect their
performance. Unfamiliar formats or unclear directions can negatively affect scores. Educators should ensure that directions on tasks are clear and unambiguous (Brown, 2010). Students should also be given the opportunity for practice in the format of a test before they have to take it for a grade. Students who take practice tests and get feedback on their performance can increase their performance on summative evaluations up to 12% (Cassady and Gridley, 2005). The implications for educational practice cannot be overlooked. Educators need to give students practice with test formats, and feedback on their performance so that students are aware of their areas of strength and weakness.

Cassaday and Gridley looked at how the format of a test can affect student anxiety about it. They found that student test anxiety has measurable effects on student performance. (Cassady & Gridley, 2005). Students who self-report high levels of test anxiety tend to do less well on tests than students who do not. This is not because of any actual difference in language competence or background knowledge, but the real effect of the affective filter on learning performance. Ongoing formative assessment with timely feedback reduces anxiety by giving students experience with test format as well as positive testing experiences. There is evidence that allowing students to decide when and how to take a test can reduce their anxiety towards it and give a more reliable measure of competence. Cassidy and Gridley also found that students who were allowed to take a test online, in a given time frame, reported lower levels of anxiety than those who had to take the test in a classroom at a proscribed time. Educators should do what they can to reduce student anxiety, so that they can get a true measure of their performance. (Cassady & Gridley, 2005). One criticism I have of this study is that it involved only undergraduate students in an educational psychology course. We may find that students in other educational settings or who have are less proficient with computers would report different levels
of anxiety around online tests. I would recommend that educators keep their student populations in mind when deciding what kind of testing format offers the least amount of anxiety.

Conclusion

There are a number of conclusions to draw from the literature. The first point is that assessments should be closely aligned with their stated purpose. If it is not, then any data collected from it is not valid. Test designers should make the purpose of an assessment unambiguous, and institutions should ensure that the assessments are being used as intended. The potential negative washback from misuse of assessment data has profound educational consequences. Students may not receive instruction that meets their needs, and teachers and institutions may be unfairly evaluated. Educators and administrators have an important role in guarding against these consequences.

A second point is the importance of rater training in increasing the validity and reliability of assessments. Rater teaching and grading experience is correlated with reliability, but raters of all experience levels benefit from training in specific rubrics and tests. Educators should have ongoing training and time to discuss how they will interpret rubric scoring criteria in order to increase inter-rater reliability. Anti-bias training is important for raters of all experience levels, since people are not always aware of the biases they bring to the task.

Of course, raters are not the only factor in the reliability and validity of test results. The type of assessment task has a measurable effect on student performance. Success at independent tasks, such as answering a writing or speaking prompt, depends largely on student language proficiency and background knowledge. This can mask a student’s true competence. Independent tasks generally do not reflect real-world communication situations and should be avoided whenever possible. Tasks that integrate two or more skills give a more reliable picture of a
student’s abilities. For example, it is better to give an essay prompt after a reading or listening passage. This gives context for the communication and helps correct the bias towards students with existing content knowledge. Educators should also be careful not to draw strong conclusions from just one assessment. Multiple measures of performance should be used to draw a more complete picture of student competence.

The final point to consider is the student being tested. Students do not come to assessments as blank slates. They may have significant anxiety around testing that negatively affects their performance. There are some ways to mitigate this anxiety. Anxiety is lower when assessment tasks are familiar, so teachers should give students practice tests that mimic the assessment. In addition, they should give timely feedback to students so that they are aware of their strengths and challenges. There is also evidence to suggest that giving students a testing window, rather than giving it at a specific time and place, can reduce anxiety and give a more accurate view of student abilities.
Chapter III

The Project and its Development

Description of the Project

Development of the Project
CHAPTER III
THE PROJECT AND ITS DEVELOPMENT

Description of the Project

The project is a handbook for education professionals who want to increase the reliability, validity, and usefulness of their subjective assessments. There are three modules, each of which focuses on a different aspect of the assessment process. While the modules are designed to be used independently of each other, the first module contains necessary background information and vocabulary and should not be skipped.

Module I will cover the key concepts and specific vocabulary of assessment. It is designed as an introduction for new teachers, and a refresher course for more experienced ones. It summarizes the most recent research into the subject as of this writing. This module will give educators a common vocabulary to discuss their assessment needs. The second and third modules are designed to be used independently of each other. Module II covers how to increase inter-rater reliability through cycles of scoring, discussion, and training. It will include sections on using collaborative processes to determine grading norms, identifying rater biases and correcting for them, and how to analyze and adjust rubrics and performance bands to make scoring more accurately reflect student competence. Module III looks at ways to change assessment tasks and procedures to limit two factors that affect assessment reliability and validity: bias and student anxiety. Special attention is paid to developing integrated tasks, using alternative assessment procedures, and measuring competence using multiple measures.

Users will only need to use the sections that are relevant to their needs. Modules are presented in two formats: a detailed training document, and a slideshow summarizing each
module. Users can choose which format best meets their training needs, but they will be much more effective if they are used together.

Development of the Project

The inspiration for this project came from the author’s many years of experience in administering and scoring standardized assessments in the California primary school system. He noticed that teachers would often score the same assessment in wildly different ways. There were many reasons for the score variances. Teachers might give more weight to different categories, or disregard some categories as irrelevant to the assessment. They would suggest that assessment performance did not reflect a student’s true competence, and adjust the score to reflect that. These differences in scoring interpretations made it difficult to make educational decisions based on the assessment data, or to make comparisons between different approaches to teaching the material. The author’s school district did not address this issue in any meaningful way, so he decided to

To develop the project, the author read peer-reviewed articles on how to increase the validity and reliability of assessments. He found three themes that researchers seemed to agree on. First, that increasing rater reliability through training and experience was key to increasing scoring validity. He was surprised to find that effective training over a short period of time could make a novice rater nearly as reliable as someone with years of experience, and that even veteran raters needed training to overcome their biases. The second major finding was that the nature of the assessment task itself could have a significant effect on examinee performance. The third finding was that the procedures used to administer a test could have a significant impact on test performance.
The nature of educational institutions is that they will rarely have the same needs. This is why the author decided to split the training document into three different sections, each addressing one of the themes addressed above. Institutions would be able to use whatever sections they required. The modules are designed so that they will be useful to teachers and administrators of all experience levels. For new teachers, or individuals who are going to be training their staff, there are text sections that give detailed information about the subjects in question. While the author encourages everyone to familiarize themselves with the contents of those sections, he understands how valuable teacher time is. That is why each text section is accompanied by helpful summaries of the most important points in the text. Veteran educators may want to review these summaries to ensure that their professional knowledge is up to date. In addition, the modules were split into two different versions, a text document and a slideshow presentation. This maximizes its utility as a training document, containing everything a presenter would need to teach these findings, and requiring little advance preparation. The author hopes that this project will lead to better educational outcomes for students.
Chapter IV

Conclusions and Recommendations

Conclusion

Recommendations
CHAPTER IV
CONCLUSIONS AND RECOMMENDATIONS

Conclusions

Teachers and institutions must do everything possible to ensure that their assessments are reliable. If teachers base their interventions on unreliable data, then students will not get the best possible educational outcomes. If administrators base their evaluations of teachers on unreliable test data, then teachers are not going to be fairly evaluated. If prospective students use unreliable assessment data to judge a program, then the program may not be fairly evaluated. All of the stakeholders in the assessment process have a strong interest in getting the most reliable data that they can.

This factors that affect reliability must be considered in total. Teacher experience in the profession and with the test has a measurable impact on reliability. Training in a rubric has a similar effect. The type of task students are asked to complete can subtly bias the results. Administration and scoring procedures must be standardized in order to get the highest inter-rater reliability. Finally, bias must be identified and corrected for to get the best results.

It took the author a significant amount of time to research these topics and develop his recommendations. Most educators do not have similar amounts of time to devote to this, so they need something that can summarize the research in an easy-to-understand, comprehensive format.

This project meets that need. Each of its modules is comprehensive and readable. It summarizes the most recent research on the topic, and its recommendations are supported not only by the research but also by the author’s decade of experience in primary education. It is designed for ease of use and should be relevant to educators at all levels.
The author would like to stress that the process of making assessments reliable is a process that may not have an endpoint. Educators should be prepared to do this work for an extended period of time. An assessment may never become perfectly reliable, but that does not mean we shouldn’t try to improve it as much as possible.

Recommendations

It is highly recommended that institutions use all three modules when designing their own program. The modules have been designed to be integrated together into a comprehensive training program. The author understands that this may not always be possible due to the many demands on educators’ time. If that is the case, institutions should have a clear idea of their needs and pick the modules that meet those needs. Educators should also be prepared to revisit these modules on a regular basis throughout the teaching year. This work is most effective when done on an ongoing basis, with regular times set aside for training and feedback. Teachers need time to reflect on their practice and collaborate with their colleagues in order to improve in their teaching, especially when it comes to standardizing assessment procedures. Do not neglect this very important part of the assessment cycle.

This project could be improved by using authentic materials as sample problems. The author hopes that anyone who uses this project will use their own assessment materials to illustrate the teaching points.
REFERENCES


Appendixes

Field Project

Module I

Module II

Module III

Module I Slideshow

Module II Slideshow

Module III Slideshow
Field Project

Module I: Overview

Module II: Rater Training

Module III: Task Adjustment
Quantifying the Qualitative:

A handbook for educators interested in increasing the reliability of their subjective assessments
Introduction

Thank you for choosing to objectify your subjective language assessments! The author hopes that this document will help you to reduce or eliminate the sources of unreliability in your assessment tasks and procedures. While the author has written this document primarily for teachers of second language acquisition, the research and recommendations are valid for most if not all kinds of subjective assessment. We must use assessment data to drive our instruction, but if our assessments do not give us accurate information about our students, we will not be able to give them what they need. The goal of any assessment program is to ensure that you can draw the most accurate conclusions from your assessment data as you can, in order to make the best instructional decisions for your students. After you are finished with this document, you will have all the information you need to transform your assessments into the most reliable versions they can be.

This document has been designed to make it easy for your institution to adapt it to your unique needs. Each module can be used independently of the others, so that you can use the ones that are most relevant to your needs. Each module is presented both as a detailed text document for in-depth study and a slideshow for summary and review.

Module I will get you up-to-date on the most recent research in assessment reliability. It contains the vocabulary and background knowledge that you will need to make informed decisions about the task at hand. It is recommended that you do not skip this module unless you and your staff are already well-versed in this information.

Module II contains information about how to increase inter- and intra-rater reliability. It explains the various factors affecting reliability, and how to control for them. In addition, it
discusses how to implement a continuous program of training and evaluation for your staff. This module draws heavily from the author’s experience with professional learning communities (PLCs), and he will explain how this model of collaboration is uniquely suited to this task.

Module III deals with how to change assessment tasks and procedures for greatest reliability. It will explain why integrated tasks are preferable to independent ones. It will also address how to change administrative procedures to minimize their effect on the results of the test. While it is the shortest of the three modules, that does not reflect its importance. Test tasks and administrative procedures can be a significant source of unreliability, and it does not take a tremendous amount of work to adjust them.

At the end of each module, there is a brief quiz about its major teaching points. An answer key is provided at the end of the document. The author encourages the use of these quizzes to check your understanding of the subject, or as discussion questions for your training sessions.

At the end of the document, you will find an annotated bibliography for use in further professional development.
Some Things to Look For

This training module has been designed so that important information in it is easy to identify. At various points in the text, there are text boxes that contain important information for novices and veterans alike. The type of information you are looking for will be in a uniquely formatted text box like the examples below:

Learning Goals
I. Will be in boxes that look like this

Vocabulary will be in text boxes that look like this

Summaries of the previous paragraphs will be in boxes that look like this
Module I: Introduction and Overview

Goals for this module are:
I. To understand the concept of reliability in assessment
II. Describe sources of unreliability in tests and raters
III. Discuss how to increase test and rater reliability

Whether you are an experienced teacher with years of teaching experience, or a novice just about to teach your first class, you are going to use subjective assessments regularly. You need to know what makes them useful, and how to maximize their utility. In order to do that, you must be familiar with some terminology used in assessments. After completing this module, you will understand the concepts of test and rater reliability. You will be able to discuss what factors affect reliability, and how to adjust for them. You will also understand what potential sources of bias there are in tests and raters, and have some idea of how to correct for them.

Raters refers to anyone who scores a test task.
Tasks refer to anything asked of a student during a test

What does the author mean when he talks about reliability? In testing, reliability is a measure of how consistent scores are across time and tasks, and how accurately they reflect student competence. High reliability allows us to make confident judgements about student performance and make informed teaching decisions. Unreliable data is not useful to educators or students. Therefore, we must maximize the reliability of the test we use in order to make the best
decisions for our students. In the next section, we will look at the factors affecting the reliability of tests.

**Reliability** means **consistency** and **accuracy** in test scores. **Test** refers to any assessment of a skill at a specific point in time.

**Factors affecting test reliability**

There are several factors that affect the reliability of a test. In this section, we will look at the sources of unreliability in a test and its administration. **Test length, item difficulty, student population, test circumstances, and test administration** will be each be discussed in detail. The discussion will identify research-based ways to make your testing as reliable as is possible.
Our first item for discussion is **test length**. All other factors being equal, longer tests are more reliable than shorter ones. The more items there are on a particular test, the less likely it is that one item will affect the overall score. Students have multiple opportunities to demonstrate their competence in a given skill or content area, giving teachers a more complete view of their abilities.

What is an ideal test length for purposes of reliability? Research tends to support including at least four tasks for each skill being tested. There is evidence that increasing the number of tasks past four has diminishing returns on reliability, and can negatively affect practicality. Do not base your opinion of a student’s abilities on just one task! Make sure to give your students many opportunities to demonstrate their knowledge.

**Item difficulty** is important to consider in any test. Test items that are too easy or too difficult do not give educators much useful information to base their teaching decisions on. Item difficulty is more relevant to objective tasks, such as multiple choice questions. It is not completely irrelevant to subjective testing, however. If students consistently underperform or over perform on a particular test item, it may need to be revised. The difficulty of a reading or
listening passage may need to be adjusted, or the comprehension questions changed to require higher-level thinking.

What does it mean for an item to be too difficult or too easy? When it comes to multiple choice questions, if fewer than 15% of students get a question correct, it’s too difficult. If more than 85% get it correct it’s too easy. The difficulty “sweet spot” is between 30% and 70% correct responses. With subjective assessments, the lines are somewhat less clear. Institutions will have to determine what difficulty benchmarks are appropriate for their assessments. This discussion should be ongoing, and relevant to the needs of the institution and the purpose of the test.

For example, if the purpose of a test is to place students into programs or to determine language proficiency, it may be appropriate to have a wide range of question difficulty. You need to be able to sort your students into appropriate categories, and so it is appropriate to have questions of very high or very low difficulty as long as those questions serve that purpose. Teachers should review their data regularly to make sure that this is the case, and adjust or remove any questions that are no longer useful.

The population being tested also affects test results. Homogenous student groups will give more consistent testing results than heterogenous ones. This factor is of particular importance in most language classrooms, since students tend to come from a wide variety of
language and cultural backgrounds. In these classes, teachers must be closely monitor the results of their assessments to help identify hidden biases in the test or the students.

When considering your student population, be aware of and sensitive to their cultural and linguistic backgrounds. They may have difficult answering questions on taboo subjects in their society, or on topics that are politically difficult in their native lands. You will not get reliable, accurate data on these questions, because your students will be too anxious or offended to answer them. Choose test questions carefully and always be respectful of your students.

A fourth factor is testing circumstances. If you want reliable, consistent results on assessments, then you need to standardize the assessment conditions as much as possible. Student performance can be affected by any number of factors. Testing areas should be clean,
well-lit, and comfortable. Distractions should be minimized as much as possible. You want your students’ focus to be on the assessment, not on their surroundings.

Teachers also need to be careful to administer a test consistently every time they give it. They should use the same language when giving directions. It may even be helpful to use a script until you are comfortable with your consistency. Using the same language each time makes it unlikely that a teacher will confuse their students or give more information than they should have for the task.

Teachers should also be consistent in how they give their students access to the testing materials. Do they pass out the test booklets and pencils first, or give directions first? Are students allowed to get materials on their own, or is there a proctor who delivers them? What kind of assistance are you allowed to give? Is the test closed, or are notes and other study materials allowed? In order for an institution to ensure the reliability of their testing results, its
educators must collaborate to find the answers to these questions. Even small changes in the administration of a test can lead to large score variances.

Making raters more reliable

The reliability of a subjective assessment’s scores depends largely on the reliability of the rater who scored it. Increasing rater reliability should be a primary goal of any educational institution. There is a large body of research into what affects rater reliability, and the following section will detail some of the most significant findings of that research.

The first variable we will discuss is the number of raters on a given task. In general, the more teachers who are scoring a given task, the more reliable the final score will be. Discussing the reasons behind score variance, and deciding what constitutes proficiency at each benchmark level, lead to better understanding of the rubric and increased accuracy of scores. There are significant gains in reliability when adding just one additional rater to a task. While there are some reliability gains from using three or four raters per task, there are considerable diminishing returns after the second. For the purposes of practicality, two raters per work sample should be the standard. It would be ideal for an experienced rater to be paired with a less experienced one, but circumstances do not always allow for that. Institutions need to put in place procedures that allow for teachers to collaborate on their scoring.
One of the biggest factors affecting rater reliability is **teaching experience**. Raters with several years of experience are generally more reliable scorers than novice raters. This shouldn’t come as a surprise, since scoring subjective assessments is a skill like any other. Novice teachers benefit from collaboration with more experienced teachers, who can help them identify their biases and blind spots. However, this does not mean that veteran teachers are always perfect! They can have biases and blind spots, too. This is why training matters.

**Training in a rubric** is strongly correlated with reliable, accurate scoring. This is true no matter how experienced or inexperienced the rater is. Enough training and feedback can reduce if not eliminate the gap in reliability caused by experience. Research suggests that three cycles of scoring, feedback, and training over three to six months closes the gap significantly. Regular trainings should be scheduled for as long as an assessment is in use. Training cycles will be discussed in more detail in the module II.

For the best combination of **reliability** and **practicality**, use two raters for each work sample.
Rater bias is another roadblock to reliable scoring. In this context, I am not referring to bias in the form of prejudice against groups of people based on external characteristics. Instead, I am referring to bias for or against parts of the grading process. Teachers are generally biased towards one of four grading styles. They tend to give the most scoring weight to either: 1. errors in production; 2. the content of the message; 3. their personal reaction to the content; or 4. the specific grading criteria. I will discuss each of these biases in more detail in the following section.
Raters that have a bias towards production errors may think they are being objective, but their scores can be as unreliable as someone with another rating style. They may give a high score to a paper that is technically perfect but does not answer the prompt completely (or at all!), or give a low score to a student who gave a detailed answer that is filled with mechanical errors.

Raters with a bias towards the content of a message have the opposite problem. Their emphasis on the message can obscure serious issues with production. The score they give to a student may not reflect their true competence, and can mask that student’s true educational needs.

In both of the above cases, the score is not a reliable indicator of the student’s competence and any educational inferences drawn from it could be wrong. Raters need to understand their own biases before they can work on being more reliable assessors. Module II will address training in identifying and overcoming biases.

Raters can be biased towards:
- The mechanics (production errors, intelligibility)
- The content (the message intended by the student)
- The scoring guide

Unrecognized bias leads to unreliable scoring!
Quiz for Module I:

1. What is reliability in assessments?
2. Name at least three things that can affect a test’s reliability.
3. Name at least three factors that can affect a rater’s reliability.
4. At a minimum, how many task should a test have to maximize its reliability?
5. Describe the external factors that can affect the results of a test.
6. Why are multiple raters for each test task more reliable than just one?
7. What are some ways in which a rater can be biased?
8. How can different forms of bias affect the reliability of a rater’s scores?
The reliability and usefulness of a score on a subjective assessment is positively correlated with two things: the experience of the rater, and the amount of training that they have received in the rubric for that assessment. Novice teachers tend to show the greatest improvement from training, but even veteran educators benefit from collaborative feedback on their scoring. For these reasons, you should ensure that your educators have regular training and feedback on the assessments that they administer.

In this module, we will cover the kind of training cycle you should put into place for your assessments. By the end of the module, you should be able to describe why educators of all experience levels need regular training. You will know how to conduct a thorough needs assessment of your educators, and will learn some procedures that can be put in place to make this kind of cycle self-sustaining.

Just as needs assessment is a central part of any teacher’s classroom practice, so it must be the central part of any training program. The first question we must ask ourselves is this: what are the training needs of the teachers I will be working with? The answer to that question will determine how you proceed. In order to find the answer to that question, you’ll have to ask quite
a few more. In the next section we will discuss the importance of knowing exactly how much experience your teachers have.

You need to consider what kinds of experience that your educators have, not only in the profession but also with this particular assessment. Their familiarity with your institutions grading norms is another factor you need to consider. You training needs will be influenced by the population you are training, so you need to know as much as possible about their experience in all of these areas.

The more experience a teacher has in the profession, the more reliable they are in rating subjective assessments. Veteran teachers, generally defined in the literature as having five or more years of experience in the field, will have had many opportunities to administer, score, and
receive feedback on the assessments they use. They may have taken leadership roles in the development of tests, and in the training of new teachers. Their input will be valuable in the discussions that will follow.

The more experience a rater has in administering and scoring a test, the more reliable and consistent they will be. If your teacher have never administered this particular assessment before, then they will need detailed instruction in its administration. Teachers who have a lot of experience with the specific assessment might need a brief refresher on the process, just to make sure that they are following it faithfully. If your teachers are a mixture of new and experienced, you will need a collaborative discussion on how you will standardize how the test is given. Everyone involved in the testing must know the answers to the following questions, so that we can reduce or eliminate testing conditions as a variable that affects student performance:

I. How much time is given for each test task?
II. How are the directions for each task given?
III. How are testing materials distributed?
IV. What kind of assistance is the teacher allowed to give?

The answers to these questions are so important, it’s probably a good idea to devote at least one training session to them. You’ll also want to revisit them throughout the school year, and revise them based on the feedback from your test administrators.

The next part of your needs assessment deals with understanding the scoring rubric and other grading criteria. Subjective assessments are sometimes difficult to score reliably because teachers misunderstand or misinterpret the scoring guide. This is compounded by the fact that
rubrics can be written ambiguous or imprecise ways. Additionally, teachers may not agree on what work needs to look like to meet each performance standard. This affects inter-rater reliability, making it more difficult to judge the effectiveness of different educational approaches. For these reasons, you should make rubric training a significant part of this process.

The less experience an educator has in the profession, the less likely they are to use the different parts of a rubric appropriately. They may not consistently use all parts of a rating scale. For example, they may never use the highest and lowest parts of the scale, even if the scoring criteria would make those scores appropriate. They may not be very consistent in how they award their scores. They will need detailed instruction in how to interpret the rubric, and should be monitored closely to make sure they are scoring their assessments accurately. They may need more frequent feedback that veteran teachers, so be prepared to schedule more frequent sessions with them.

The first training session should be devoted to analyzing the language used to describe each performance level of the rubric. You’ll need to answer the following questions about the rubric in order to determine what, if anything, needs to change:

I. Is it easy to determine why each level is described as it is?
A. Look at the differences in language used from level to level. Is the language vague, and open to excessive interpretation? An abundance of qualifiers such as “rarely,” “often,” or “consistently” can make reliable scoring difficult.

B. Look for a natural progression of performance expectations from level to level. If the jumps in performance expectations seem inconsistent or confusing, the rubric may need to be adjusted.

II. Does the progression from lowest to highest proficiency on the rubric align with the classroom learning goals?

A. A rubric is useless to making educational decisions if it is not relevant to what you are teaching. If the answer to this question is “no,” then you will need to seriously evaluate either the assessment or the course objectives.

III. Is the rubric detailed and specific enough for our educational needs?

A. Holistic rubrics give one score for the assessment that is supposed to represent overall proficiency. There is no breakdown of the score into individual language components, such as grammatical proficiency or vocabulary usage. This makes them useful for summative evaluations, but less useful for ongoing formative assessment.

B. Analytic rubrics give multiple scores for an assessment based on the language skills that are being tested. They tend to give more diagnostic information that holistic rubrics, making them useful for identifying the specific skills that students are excelling at, or need extra support in. Teachers tend to be more reliable when using analytic rubrics.

The answers to these questions are best arrived in a collaborative discussion with the teachers. They must have input into how scoring is interpreted. They have valuable insight to
give about how rubrics relate to course content. They know what content is being delivered to
their students, and how they are teaching it. They know how to align the scoring criteria on a
rubric with the course content and their own expectations. If they see a mismatch between the
course and the rubric, they will lose confidence in the assessment as a tool for guiding
instruction. They will resent having to administer it, and may passively or actively sabotage the
process. This can be avoided by giving them a voice in how the assessment is scored.

No institution has an unlimited amount of time for training and feedback. You must know
how much you can afford to devote to this, and over what time frame. It may be difficult or
impossible to address all of your staff’s needs, so identify the most urgent ones and plan to teach
those first. Keep in mind that you will need multiple sessions over time to effectively train your
teachers. Research suggests a minimum of three sessions for each training topic, with feedback
sessions after each training. The exact number of sessions, and the time between them, depends
on your needs.

How will the training be delivered? Will you be working with individuals, small groups,
or an entire staff? Will you have staff meeting time reserved for the training, or will you have to
schedule it during the work day? How will you ensure regular re-training sessions for the staff?
Remember, this is a process that is never finished. Ongoing professional development is crucial
to effective teaching.
Quiz for Module II:

1. What is a needs assessment, and why is it important to this process?
2. What three areas do you need to know your teachers’ levels of experience?
3. Why is it important to know your teachers’ experience in these areas?
4. What specific item might novice teachers need more training in?
5. What factors in the administration of a test should you attempt to standardize?
6. What questions should you ask when determining the usefulness of a scoring rubric?
When we are assessing our students, we are asking the question: what language skills have they mastered, and what do they still need to be taught? Therefore, we must make sure that the tasks we give them will give us reliable answers to that question. What difference does the type of task make to assessment reliability? Quite a bit, as it turns out. The way a task is written can bias it towards certain groups of students. If a task is biased, it means that student performance does not accurately reflect their competence. This makes any educational decision that is based on that performance unreliable. You must do everything possible to eliminate potential sources of bias, so that you can trust the results of your assessments. After this module, you will be able to explain why integrated tasks are more reliable assessments than independent ones. You will understand how to adjust an independent task to make it more integrated.

*Independent tasks* are very common in subjective tests. These are stand-alone questions that are not linked to a text or listening passage. They can be short answer questions, essay prompts, or anything in between. What makes them *independent* is that a student must answer them with the knowledge and information they already have. This might seem like a good thing,
at first glance. Don’t we want to know what our students can do independently? Isn’t that the point of an assessment? What, exactly, is wrong with this kind of task?

Independent tasks are not linked to a reading or listening passage. They require the student to answer using only the knowledge they already possess.

From a communicative language teaching perspective, the problem with an independent task is that it is not an authentic example of communication. It tells us nothing about a student’s ability to use language in a real-world context. Their performance on such a task has no predictive value, and thus is not useful for educational purposes. In addition, the task is heavily biased towards students who have significant prior knowledge of the subject matter. This may not be relevant if the task is a summative evaluation of a course. In that case, you want your students to demonstrate their knowledge and an independent task will be appropriate. It is definitely relevant if the task is for placement or proficiency evaluation! In that case you want to make sure your evaluation is as objective as possible. For that, you need an integrated task.

Integrated tasks combine two different language skills for a more comprehensive evaluation. Reading a selection and answering content questions is one example of an integrated task. Listening to a passage, then speaking about it with an evaluator is another example. These tasks are more reliable indicators of a student’s abilities because they more accurately reflect how language is used in real-world situations. They also correct the bias towards prior knowledge that independent tasks often have. Ideally, the content knowledge a student needs to answer the questions is contained within the listening or reading passage. This makes the test a
more accurate indicator of linguistic competence than background knowledge, and makes the results more accurately reflect student abilities.

**Integrated tasks are questions that are linked to a listening or reading passage. They integrate multiple skills.**

You should have a strong preference for including integrated tasks in your assessments. This is not always the easiest adjustment to make. Finding a leveled passage that is appropriate for your students and has the relevant information may be difficult. You may need to write it yourself, or adapt something from a published work. If you cannot do this, then you need to make sure that your independent questions are as unbiased as possible. You will need to identify the potential biases, and look at how you might adapt the question to reduce or eliminate that bias. Here are a few questions you should ask yourself when determining if a question is biased in some way:

1. Is specific content knowledge required to answer this question? If so, do your students have the appropriate knowledge?

2. Is the question biased towards specific cultural knowledge? For example, does it require familiarity with cultural norms and practices? How many of your students possess this knowledge?

3. Does the question contain language that is biased towards any particular ethnic or religious group? Is there any kind of implicit or explicit stereotyping?

Identifying and correcting bias is often a difficult process because we are mostly unaware of what our own biases are. You may find it useful to do further research on this topic. For a
good overview of the significant issues, go to edglossary.org/test-bias. For more detailed information along with recommendations for reducing and eliminating all kinds of biases in testing, look at the Pearson handbook *Assessment bias: how to banish it, 2nd edition*, by W. James Popham. It is available online and free of charge at ati.pearson.com/downloads/chapters/Popham_Bias_BK04.pdf.

Quiz for Module III

1. What is an *independent task*?
2. What is an *integrated task*?
3. Why are integrated tasks more reliable than independent ones?
4. What can you do to make independent tasks integrated?
5. What are some questions you should ask to determine the *biases* of test questions?
Answer Keys for Module Quizzes
Answer Key for Module I Quiz:

1. Reliability means consistency and accuracy in test scores.
2. There are many factors affecting test reliability, including test length, item difficulty, student population, testing circumstances, and administrative procedures.
3. Rater reliability is affected by experience in the profession, experience with the specific test, and biases.
4. Tests should have four tasks to maximize their reliability.
5. Some external factors affecting reliability include temperature, noise, light level, and time of day the test is given.
6. Using multiple raters is more reliable because collaborative discussion leads to more consistent results. Raters can also help each other recognize and correct for their biases.
7. Raters can be biased towards specific kinds of test tasks or grading styles. For example, they might give more weight to mechanics rather than content.
8. Rater bias affects the reliability of scores because they may not reflect the true competence of a student.
Answer key for Module II Quiz

1. A needs assessment is a process to determine the distance between “what is” and “what should be.” Any effective educational intervention requires an accurate needs assessment.

2. You need to know a teacher’s experience in the profession, with administering the test in question, and at the institution itself.

3. You need to know your teachers’ experience in each of these areas because there are very specific training needs for each one.

4. Novice teachers need extra training in how to interpret the scoring criteria on a rubric.

5. You need to standardize how much time you give for each task, how you give directions for each task, how you distribute materials, and how you give assistance.

6. Are the scoring criteria easy to understand? Do proficiency levels align with course objectives? Is the type of rubric appropriate for the task and the institution?
Answer Key for Module III Quiz:

1. An independent task is a stand-alone question, without any reading, listening, or speaking passage to give it context.

2. An integrated task assesses at least two language skills: reading, writing, speaking, and listening.

3. Integrated tasks are more reliable assessments because they are more authentic examples of communication. Context, in the form of a reading or listening passage, also helps eliminate potential bias.

4. Integrate independent tasks by pairing them with a complementary skill:
   Reading and writing tasks should be paired with listening or speaking tasks.

5. What kind of content knowledge is required to answer the questions? What
Biases must be identified and corrected for in order for a test to be useful.
• Detecting biases is often difficult.
• We are usually unaware of our own biases.
• Our goal should be to identify and eliminate any potential biases