Observer Calibration

A Tool for Maintaining Accurate and Reliable Classroom Observations

EDUCATIONAL TESTING SERVICE (ETS)

Jilliam Joe
Associate Research Scientist

Jaymie Kosa
Assessment Specialist

Jessica Tierney
Assessment Specialist

Cynthia Tocci
Executive Director, Research Division

About the Authors

Jilliam Joe

Jilliam Joe is an associate research scientist at ETS, where she conducts research to improve the quality of human scoring. Jilliam provided psychometric support for video-based classroom observation as part of the Measures of Effective Teaching (MET) project, a study funded by the Bill & Melinda Gates Foundation. She currently provides research direction and technical support for the Teachscape *Focus*™ Observation Training and Assessment (which is based on the Framework for Teaching Evaluation Instrument), a product developed jointly with Teachscape and Charlotte Danielson based on the expertise gained from the MET project. Jilliam has extensive experience in all phases of assessment development, research, and qualitative and statistical analysis, and has participated in several performance-based assessment projects. Her special interests are in rater cognition and the reliability of scoring for performance assessments. Jilliam received her Ph.D. in assessment and measurement from James Madison University.

Jaymie Kosa

Jaymie Kosa, a former middle school teacher, is an assessment specialist at ETS. She is currently responsible for developing performance assessments for initial teacher licensure and for Teachscape *Focus*. Previous experience includes serving as a master coder for the master-scored videos in Teachscape *Focus*; training raters for the Washington *ProTeach Portfolio* and the Kansas Teacher Leader Assessment; and serving as the content lead for the Protocol for Language Arts Teaching Observations (PLATO) instrument as part of the MET project. For 11 years, Jaymie taught language arts, reading/literature, media literacy and storytelling/performing arts. She also served on the National Board for Professional Teaching Standards and the National Commission for No Child Left Behind. In addition, she served on the Roosevelt Public School Board of Education for nine years. Jaymie holds an M.Ed. in language arts education from Rutgers University, and graduated Phi Beta Kappa from the University of Maryland, College Park, with a B.A in English language and literature.

Jessica Tierney

Jessica Tierney, a former elementary school teacher, is an assessment specialist at ETS. She is currently responsible for developing performance assessment for initial teacher licensure and for Teachscape *Focus*, which supports the Framework for Teaching Evaluation Instrument. Previous experience includes training master coders for Teachscape *Focus*; master scoring for the MET project; and assessing for *PRAXIS III*, an observation system for teacher licensure of first year teachers. Jessica taught for nine years at the elementary level, and served as adjunct professor at the University of Dayton. She holds an M.Ed. degree in curriculum and instruction from the University of Cincinnati, and a B.S. in elementary education from the University of Dayton.

Cindy Tocci

Cindy Tocci, a former middle school teacher, is an executive director in the research division at ETS and is responsible for facilitating research and development projects. Her current work in the area of teaching effectiveness includes Teachscape *Focus*. Previous experience includes work on measures of teacher effectiveness, classroom formative assessment projects and professional development to improve teacher practice. Her projects included scoring video of classrooms as part of the MET project, the Kansas Department of Education's evaluation frameworks for teachers, principals, and superintendents, the National Board for Professional Teaching Standards (NBPTS) assessments, and professional development programs for improving teacher practice called *Pathwise*®. Cindy received her Ph. D. in Education from Emory University.

Table of Contents

Executive Summary

After significant investment in initial observer training and certification, states and districts now need to consider how to support observers over time. Calibration is a process and a tool that helps observers maintain scoring accuracy and avoid rater "drift," provides teachers with assurances of accurate and fair assessments of their classroom practice, and allows states and districts to evaluate the overall quality of the observation process on an ongoing basis. The ideal calibration exercise provides observers with standardized and authentic observation experiences from which they can make fairly reliable inferences about the accuracy of their scoring. Online delivery of such exercises is an efficient way for states and districts to overcome logistical constraints associated with access to training and to manage data and track observer performance at individual and group levels.

Introduction

Swift and widespread changes are taking place in K–12 teaching evaluation in the United States. Many states and districts have invested significant resources into observer training and certification and are now considering how to support observers on an ongoing basis. The purpose of this paper is to present calibration as that next-step tool and process for maintaining high-quality observations after observers have completed training and certification.

In the context of performance-based assessment and scoring, to "calibrate" is to measure one's scoring performance against a standard for accuracy and, if necessary, take immediate and targeted action to align with that standard. Research has shown that the use of calibration as an intervention after training and certification significantly improves scoring accuracy and consistency (Cash, Hamre, Pianta, & Myers, 2012; Ricker-Pedley, 2011). In fact, frequent calibration (sometimes as frequent as every day) is a matter of course for large-scale constructed response or performance-based assessment scoring programs (Johnson, Penny, & Gordon, 2009), and for major large-scale admissions testing programs like the Graduate Record Examination, and for state-level teacher certification programs like Washington ProTeach.

The role of an observer calibration tool, therefore, is to provide activities that would help observers identify and correct errors in scoring, such as drift (defined as the gradual degradation of a rater's scoring accuracy over time). The process of calibrating to the scoring rubric reinforces behaviors that support accurate scoring, such as evaluating observed lessons against the rubric and training benchmarks rather than against one another.

Why Initial Training Is Not Enough

In areas of professional certification or licensure, certification satisfies the claim that an individual possesses the requisite knowledge, skills, and abilities to perform the task for which he or she has been trained. However, completion of training and certification is merely the first part of a practitioner's journey. It is not enough to support claims about the quality of the practitioner's skill once he or she is in the field. In many ways, this argument is true for observer training and certification. Initial training and certification are not enough to foster an observer's ability to score accurately and consistently over time while in the field. Periodic assessments and feedback opportunities are necessary. These frequent checkpoints ensure that observers are providing accurate and consistent scores to teachers so that observation data can be useful for improving teaching practice and for promoting student achievement.

This discussion about post-training calibration is important because even well-trained and practiced raters can exhibit undesirable scoring behaviors at times (Myford & Wolfe, 2003; Park & DeCarlo, 2011; Harik, Clauser, Grabovsky, Nugester, Swanson, & Nandakumar, 2009). They can become less able to consistently differentiate between the levels of performance defined by the rubric. Their scoring can also move away from the rubric standards (i.e., drift), gradually becoming more severe or lenient. Casabianca, Lockwood, and McCaffrey (2013) studied patterns of classroom observer scoring behavior over time using the Classroom Assessment Scoring System – Secondary, or CLASS-S (Pianta, Hamre, Haynes, Mintz, & LaParo, 2007). Classroom observation data were collected by 11 trained and certified observers for 458 teachers and 1,829 lessons over a period of two years. Observers calibrated with master-scored videos once every three weeks, and there was no minimum level of agreement required to continue participation in the study. Throughout the study, the level of feedback given to observers varied, and included: 1) No feedback; 2) feedback on agreement sent via email to observers; and 3) extensive review of discrepant scores with content experts. Differences in drift among the three types of feedback were not examined.

After controlling for naturally occurring changes in teaching quality over the two-year time period, the findings showed that variability in teachers' scores over time was largely due to changes in the severity of observers' scoring. For some domains of the CLASS-S instrument (e.g., Emotional Support), observers were lenient when they began scoring and became increasingly severe over time. For other domains (e.g., Classroom Observation), they became increasingly lenient over time. This findings provides further evidence that changes in scoring behavior can occur, even when observers are trained and certified. Casabianca et al. also found that observers who consistently disagreed with the master score by more than two points on the calibration exercises were also among those who consistently drifted.

Although the generalizability of these findings is limited by the study's sample size and other nuances, the results do reveal a few things about classroom observation that should be taken into consideration when designing an observation system. Primarily, if scoring quality is not monitored frequently through tools like calibration and, more importantly, if no intervention is provided to remediate or bar from practice observers who are consistently "off-scale," drift from the ideal scoring standard can have a progressively adverse impact on the quality of inferences made about classroom teaching practice based on observation data.

In most cases, such errors or effects on scoring have little to do with the observer's underlying ability to score—to learn and adopt a set of scoring guidelines and put aside personal preferences. A well-designed and technically sound certification test should uncover fundamental issues with the observer's understanding and use of the scoring rubric before they are even allowed to calibrate. However, scoring errors are mostly the result of factors that affect observers in unpredictable ways.

There are at least three ways in which drift can set in and impact scoring in practice. First, an observer's colleagues can play a role in influencing scoring drift. Take, for example, a group of observers who meet weekly to discuss their observation experiences. There may be one particularly opinionated and persuasive (as well as slightly inaccurate and biased) colleague who tends to influence the group's interpretation of the scoring rubric. As a result, individual observers who were certified as accurate have modified their internal performance standards over time to assuage their influential colleague, becoming more calibrated to the colleague's view of the rubric rather than with the rubric itself. It should be noted that group discussions are not inherently bad; at the same time, all conversations must be grounded in the rubric and anchored by the expert-scored examples of teaching practice presented during training.

Second, observers can inadvertently shift their expectations and their perceptions of the rubric based on the levels of practice represented in the group of teachers they observe throughout the year. If, for example, there are happen to be very few or no low-performing or high-performing teachers in an observer's school, an observer can drift from the standard of accuracy by forcing teaching practice to become more distributed than what it is in truth. Scoring is guided by standards of relative performance or personal preferences, rather than by the rubric. For example, with a rubric, such as the Framework for Teaching, that uses level performance descriptors like "unsatisfactory," "basic," "proficient," and "distinguished," "basic" performance can be perceived as "proficient," and "proficient" performance can become "distinguished."

Third, the daily stresses in a school day that produce fatigue or mental drain can influence an observer's ability to be vigilant about his or her biases and personal preferences, or the overall quality of their observations in general. With a weakened ability to attend to bias triggers, an observer's underlying biases can affect not only the interpretation of evidence, but also the kinds of evidence collected and, ultimately, the scores that are assigned.

While these examples may be unsettling to consider, they need to be acknowledged. If left unchecked, they undermine the validity of observation scores and feedback. Calibration is, arguably, one of the most effective tools for identifying inaccuracies in scoring and, in many cases, mitigating the effects of those inaccuracies on the quality of observation data.

Effective Observer Calibration

Before we discuss the characteristics of an ideal calibration exercise, it is important that we talk about using videos of classroom lessons as the foundation for calibration. At a minimum, calibration exercises use videos with true or master scores—those assigned by experts during a master-scoring process. The use of videos with master scores provides the means for a standardized calibration exercise for multiple observers. Observers have the opportunity to view and score a common set of lessons to assess where their scores are relative to the master score—the standard for accuracy. Then, depending on the stakes associated with calibration results and the type of feedback provided by the calibration tool (discussed in more detail later in the paper), they can discuss what they observed with their colleagues or an observation coach. The number of videos included in a calibration set ranges in practice from one 60-minute video of one teacher/one lesson to several 15-minute video clips of teaching by multiple teachers. It is advisable to use several (at least two) calibration videos for a reasonably reliable assessment of scoring accuracy. Developing calibration materials requires not only access to teachers who are willing to have their classes video recorded, but also master scorers who are able to dedicate several hours to scoring and, depending on the stakes for calibration, writing rationales for each video.

Master Scoring

As mentioned previously, true scores for each calibration video are obtained through a master-scoring process, which involves two highly trained experts who have demonstrated their knowledge of the observation instrument and proficiency in applying it. These experts independently review and score video-recorded lessons, noting and time-stamping essential evidence for each attribute of teaching practice contained in the scoring rubric, then assigning a score for each piece of evidence. Together, they draft a score rationale for each score that is assigned. Afterward, another pair of experts reviews the scores and draft rationales, and all four experts' scores are reconciled through discussion to reach a final score and final score rationales. The rationales provide a justification for each score and explain how the evidence from the video supports the score. Rationales can either be archived for auditing purposes or given as feedback to observers, if that is appropriate for the stakes of calibration.

Characteristics of an Ideal Calibration Exercise

An ideal calibration exercise provides observers with a standardized and authentic observation experience from which they can make reliable inferences about the accuracy of their scoring. Ideally, there would be more than one video to score. The exact number depends on the consequences assigned to how well an observer performs (consequences will be discussed later in this paper). Further, there would be enough feedback given to observers to guide their remediation efforts, if needed.

The calibration exercise is generally customized by grade band so that observers see calibration videos that reflect teaching in the grade levels they are assigned to observe. If, for example, an observer assesses teachers at the elementary level only, the video-recorded lessons should represent a mixture of elementary classrooms. For individuals who conduct observations across all grade levels, the calibration videos should represent classroom lessons at the elementary, middle school, and high school levels. Videos should represent a variety of content areas if the scoring rubric is not aligned with any particular academic content area. In addition, throughout the calibration exercises, observers are given opportunities to observe a range of teaching performance levels across the set of videos included in the exercise.

Calibration developers should ensure a wide level of diversity in the videos to prevent one ethnic or gender group from being consistently associated with a particular level of performance. For example, if only low-scoring male teachers are shown in the videos, developers may unintentionally create an expectation bias in observers that all males will be low scoring; similarly, if only high-scoring Hispanic teachers are shown in the videos, observers may think that all Hispanic teachers are high scoring.

Types of Calibration

Observer calibration can be used in a non-consequential or a consequential way. Non-consequential calibration—when there are no consequences for observers based on their performance in calibration exercises—is meant to help observers develop the habit of scoring accurately by providing as much feedback on their performance as possible. However, no expectations are set for performance that would prevent them from conducting live observations. Consequential calibration, on the other hand, holds observers accountable for scoring accurately by preventing those who have not demonstrated accuracy from conducting live observations based on established expectations or standards. The kinds of feedback given to observers in a consequential setting are limited. The consequences of calibration and level of performance feedback given to the observers are two important decisions that must be made when developing a calibration tool. The benefits and trade-offs are shown in Table 1 and discussed in more detail below.

TABLE 1. CALIBRATION APPROACHES

Calibration Approach	Feedback Given to Observer	Test Security	Consequences	External Intervention by District	Technical Standards (Reliability & Validity)	Development Resources Required
Non-consequential calibration	True scores, score rationales, and observer's agreement with true scores	True scores are exposed; videos cannot be reused	Observers permitted to continue observations when they deem themselves ready	Minimal	Minimal	High due to exposure of true scores
Consequential calibration	Pass or fail	True scores are secure; videos can be reused	Observers permitted to continue observations only when they meet the standard	Moderate	High	Moderate; overall demand for videos offset by reuse

NON-CONSEQUENTIAL CALIBRATION

Non-consequential calibration exercises are analogous to a series of formative assessments that precede a final examination (in this case, a recertification test). These exercises are designed to identify lapses in scoring accuracy and inform observers of the areas of performance that they most need to work on in order to score more accurately. Observers receive immediate feedback on the accuracy of their scores compared to the master scores, as well as, in the ideal case, score rationales for the master scores that offer a deeper understanding of the scoring rubric. Feedback to observers also includes specific and actionable steps that can be taken to improve or maintain scoring accuracy. These feedback messages are, in most cases, based on observers' level of scoring accuracy (agreement with the master score) and some general expectation for how accurately observers should score in practice. The feedback observers receive could link them to specific areas of training that correspond to the areas of weakness identified through the calibration exercise. It could also include guidance on how to examine scoring behavior for the influence of bias, or a recommendation to seek mentoring from a coach or another observer who has a firm understanding of the observation instrument.

There is a trade-off in providing master scores and score rationales in particular. Even though observers can benefit from specific information about their performance, once the master scores for the videos are revealed, the power of the exercise as a check on accuracy is significantly diminished. If the time between the initial and subsequent viewing of a video is short, it is likely that observers will recall their previous scores or impressions of the teacher in the video, and assign scores based on their memory of the videos rather than on an unbiased observation of the teaching in the video. This would not help observers address areas of scoring inaccuracy, and from a measurement perspective, it is unsound practice. Therefore, in a non-consequential setting, to ensure calibration experiences are not influenced by prior knowledge a fairly large pool of master-scored videos is needed to produce unique, but parallel, calibration exercises, and to ensure a reasonable amount of time elapses before a video is used again.

Also, when calibration is used in a non-consequential way and observers are permitted to continue conducting live classroom observations even if their calibration performance does not meet some pre-established standard for accuracy, the level of risk to both the observer and the teachers the observer is responsible for increases. A significant advantage, however, is information on an observer's scoring accuracy can be made available to district coordinators and others if desired, which sets the stage for observers to have constructive conversations with colleagues or supervisors about their areas of weakness and strength prior to continuing observations.

CONSEQUENTIAL CALIBRATION

The use of calibration as an accountability tool, as with consequential calibration, confirms observers' scoring accuracy before they are permitted to continue conducting live observations, thus decreasing the risk associated with scoring inaccuracy to both the observer and the teacher. Unlike non-consequential calibration, one potential disadvantage of consequential calibration is observers are not told the master scores and are only provided indicators of their scoring accuracy (e.g., "Pass/on target" or "Fail/off target"), which may not be enough information to guide a remediation effort. The advantage of consequential calibration is that since the master scores remain secure, the videos can be used again with the same observers. Videos should not be overused, however; over time, observers may talk to one another about the videos and discover a set of scores that would allow them to pass. There are other means of securing the master scores, such as administering multiple parallel forms of calibration in a proctored environment (for more in-depth discussion on test security, see McClellan, Atkinson, & Danielson, 2012).

Since, after failing a consequential calibration exercise observers are not permitted to continue with live observations, they should engage in some other activity to improve their performance before another attempt at calibration is made, and before they are permitted to continue live observations. Ideally, other opportunities or support mechanisms are provided to generate specific feedback on observers' scoring performance (e.g., paired observations or practicing scoring other lesson videos with colleagues). Calibration data can also be pooled across observers in the state or district, enabling the coordinator to determine if there are "hotspots" of scoring inaccuracy that need to be addressed at large, which could help in remediation efforts.

Consequential calibration requires more time to develop and administer in order to satisfy the standards for reliability and validity and support the decision to permit or bar an administrator from conducting live observations. To ensure observers' calibration scores are reliable (which is strongly recommended given the level of accountability in consequential calibration), observers should be required to enter multiple scores (either through several unique videos or multiple segments from one or two videos) for each component of the instrument. In determining whether to use the consequential calibration model, then, districts should consider the time investment and what supports exist to ensure observers have the best chance at success in light of the amount of investment.

Implementation of Calibration

Before implementing a calibration program, districts have decisions to make about not only the accountability level of calibration, but also how frequently to calibrate and whether to calibrate in person or online. These decisions must be balanced with resource and development constraints, data management, and other context-specific factors at the district or state level (e.g., regulations, etc.).

If calibration exercises will be developed by the district, the availability of local resources will determine how many calibration opportunities can be offered to observers and how those calibration results are to be used (i.e., for non-consequential or consequential purposes). A large pool of master-scored videos can yield enough samples to ensure reliability of calibration as an accountability tool as well as build several calibration exercises that allow observers to calibrate multiple times without seeing the same videos one after the other. The burden of capturing video may be reduced by building a library with neighboring districts—provided conditions of video recording, such as parental permissions, have been satisfied—or by taking advantage of publicly available video libraries.

It is hard to predict when scoring behaviors will fluctuate or lead to unreliable scores over the course of time. In large-scale, performance-based assessment, each day every rater is required to take a calibration assessment prior to scoring. Raters are issued a “Pass” or “Fail” result, and calibration is consequential—if after two attempts observers have not calibrated, they are not permitted to score for that day.

Daily calibration may not be a feasible practice for most districts since observations are not the only thing administrators and school leaders do in a day. It is recommended that, at a minimum, observers calibrate at least three times per school year. Calibration dates will depend on the district’s academic calendar, but ideally, calibration will occur after long breaks in between observations (e.g., summer and holiday breaks) and prior to conducting a large number of observations. The stakes associated with the observation (and more specifically for the teachers being observed) may also dictate how frequently an observer calibrates. For example, a district may decide to require additional calibrations or that observers practice scoring with master-scored videos prior to observing teachers who are in their probationary period, or those who are consistently low-performing teachers. That is, districts may implement additional requirements for observers when the stakes for a particular set of observations are higher. The time window between when observers have demonstrated scoring accuracy and when they conduct observations that will lead to high-stakes personnel decisions for teachers should be brief.

Besides frequency of calibration, the method in which calibration is delivered—in person or online—is another decision to be made. Conducting calibration in person with master scorers can strengthen observers’ sense of support and community, and provides a safe and professional setting to share with one another and address challenges. With a facilitator who focuses on the scoring rubric and benchmarks, rich conversations about teaching practice as defined by the rubric can take place, bringing to light common misconceptions or errors observers make in applying the rubric. To get the full benefit of the discussion as a learning opportunity, it is important that, prior to the group meeting, observers score the videos independently and without knowledge of the scores others have assigned.

The disadvantage to in-person calibration is that it requires a trained facilitator with deep knowledge of the observation protocol and rubric, along with master-scored videos. It can also constrain the frequency of calibrations since administrators may not be able to meet often due to the logistics of travel and scheduling. A variation on in-person calibration is phone- or web-based meetings where observers gather in a virtual environment to discuss their scores. In both cases, a trained facilitator must be present to moderate the conversation and ensure that it remains grounded in the observation instrument, the scores and rationales derived through master scoring, and best practices for accurate and reliable scoring.

Calibration can also be delivered in an online environment. It is far more efficient than in-person calibration, and it removes many of the logistical constraints associated with in-person calibration. For large state and district implementations, in particular, online calibration is a more efficient way to calibrate observers multiple times throughout the year. Another advantage of online calibration delivery is that it provides an efficient means for managing data and tracking observer performance. However, a considerable amount of design, engineering, and scoring expertise is needed to build an online calibration system. States or districts interested in delivering online calibration should explore existing software solutions for calibration.

Finally, data management must be a consideration. Depending on the frequency and type of calibration, it is possible for an individual observer to have a substantial amount of performance data. How to capture, store, and organize those data is an important consideration for analysis, reporting, and feedback to the observer, and also for timely intervention, if it is required. A well-built data management system will facilitate oversight of observer performance.

Conclusion

Calibration serves three important purposes. First, it is useful for detecting scoring errors that occur from day to day or over longer periods of time. Second, it is a means by which observers can receive ongoing feedback about their scoring performance. Third, the process of calibrating inspires confidence in the observation system. For teachers, observer calibration helps to ensure accurate and fair assessments of their classroom teaching practice—assessments that will inform their overall evaluation and what professional development is necessary. For observers, the calibration process helps to maintain scoring accuracy and develop self-confidence in their observation skills. It can ensure that post-observation conferences with teachers are fruitful and are grounded in a reliable assessment of that teacher's classroom practice. Finally, for coordinators at the state and district levels, calibration is a means to evaluate the overall quality of the observation process on an ongoing basis, and it supports the validity of personnel decisions that are made based on information gathered from observations.

The purpose of this paper is to convey the idea that initial training and certification are necessary, but they are not adequate for providing ongoing assurance that the classroom observations are accurate and reliable. Given that the consequences of observation scores are far more litigious than in times past, particularly as the stakes of teacher evaluation increase, frequent checking of observers' maintenance of their scoring skills is not an unreasonable proposition. Calibration, when used as an intervention, realigns observers to a common set of standards so that scores remain meaningful and valid for the intended purposes.

References

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2013). Rater drift in classroom observation scores. *Manuscript submitted for publication.*

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*(3), 529-542.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43-58.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: The Guilford Press.

McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluator training and certification: Lessons learned from the Measures of Effective Teaching project*. San Francisco, CA: Teachscape.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Park, Y. S. (2011). *Reliability estimates for Danielson Framework for Teacher component scores*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Park, Y. S., & DeCarlo, L. T. (2011). *Effects of classification accuracy under rater drift via latent class signal detection theory* and item response theory. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ricker-Pedley, K. L. (2011). *An examination of the link between rater calibration performance and subsequent scoring accuracy in Graduate Record Examinations (GRE) Writing* (ETS RR-11-03). Princeton, NJ: Educational Testing Service.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

877.98.TEACH
info@teachscape.com
71 Stevenson St., 20th Floor, San Francisco, CA 94105
www.teachscape.com