Reliability as Argument

Jay Parkes, University of New Mexico

Reliability consists of both important social and scientific values and methods for evidencing those values, though in practice methods are often conflated with the values. With the two distinctly understood, a reliability argument can be made that articulates the particular reliability values most relevant to the particular measurement situation and then the most appropriate evidence and theory to support an argument for the presence of that value. The steps in making a reliability argument are explained and an extended example is given. The article is intended to provoke discussion, debate, and the development of additional reliability methodologies.

Keywords: classroom assessment, measurement theory, reliability

The reliability of scores has long been the sine qua non of sound educational and psychological measurement, the primary condition to be met in order for any assessment procedure to be deemed “sound measurement”. Though it is logical in many respects to address reliability concerns early in a measurement development process, some have questioned reliability’s preeminence among psychometric properties, calling reliability “privileged” (e.g., Moss, 1994, p. 6; Nichols & Smith, 1998, p. 25). The performance assessment movement of the 1980s and 1990s waned largely because large-scale performance assessment scores struggled to, but never did, achieve sufficient reliability (e.g., Linn, 1993, 1994; Mehrens, 1992). There were logistical, time, and resource constraints, too (Willson, 1991), but it is interesting to ponder whether those issues would have been addressable had performance assessment scores been more reliable. During that period, several different approaches to the poor reliability were argued and included everything from ways to raise the reliability estimates to arguments that reliability should be sacrificed to validity (Moss, 1994). The consistent response was: reliability is important and cannot be sacrificed (Brennan, 1998; Burger & Burger, 1994; Messick, 1994).

That response needs to be refined, however, by acknowledging that reliability consists of both important social and scientific values (Messick, 1994, 1995) and methods for evidencing those values. The problem that the performance assessment movement faced, that classroom assessment continues to highlight (e.g., Brookhart, 2003), and that is relevant to all of educational and psychological measurement in varying degrees, is a problem with reliability methodology, not one with the values of dependability, stability, accuracy, and consistency. Performance assessments and classroom assessments, in fact any assessment, must exhibit those important values (Burger & Burger, 1994), but the acceptable forms of evidence of those values, the reliability methodologies, could be broadened.

This article has three purposes:
1. to articulate some of the ways in which extant reliability methodologies bound demonstrations of reliability;
2. to propose reliability as argument in such a way as to expand the sources of evidence available for demonstrating the social and scientific value of reliability; and
3. to provoke discussion, debate, and, hopefully, development of new reliability methods and methodologies.

The first section of this article will articulate where the current boundaries are on reliability methodologies. The second section will propose that reliability, like validity, consist of an argument supported by evidence, which, while encompassing all current practice, would also permit the development and acceptance of other methodologies. The third section will demonstrate how this might work with a classroom example.

The Boundaries Set by Current Reliability Methodologies

Classical Test Theory, Generalizability Theory, and Item Response Theory are primarily statistical models, tools, which do certain jobs. The outcomes of the use of these tools—reliability coefficients, dependability coefficients, standard errors of measurement, information functions, agreement indices—serve as evidence of broader social and scientific values that are critically important in assessment. So a reliability coefficient is a piece of evidence that operationalizes the values of accuracy, dependability, stability, consistency, or precision. In practice and in rhetoric, however, the methodologies for evidencing reliability are often conflated with the social and scientific values of reliability. It is not uncommon to hear that the reliability coefficient is the reliability of a set of scores (or, even more anachronistically, the reliability of a test) (cf. Brennan, 1998; Thompson & Vacha-Haase, 2000). That is distinctly not the case. The difference can be highlighted in considering the question, “Can a one-item test produce a reliable score?” The conceptual answer is, “Sure! One question can produce a reliable score.” The existing methodological answer is, “Well, some kind of replication is needed before that reliability can be estimated.” In other words, at least two items would have to be used, or the same item asked on two occasions, or two raters must respond to the single item before an estimate of the reliability can be made (Brennan, 1998). So the reliability of a one-item test score and the method for estimating that reliability via a replication are distinctly different.
The conflation has some unintended consequences (Moss, 1994, 2004). As Schuwirth and van der Vleuten (2006) put it, “At the moment, we firmly believe that the educational community tends to see only the benefits of the theory, and that we tend to forget the sacrifices we need to make in this model” (p. 296). It can lead to a stunting of innovation in assessment (in large-scale as well as small-scale settings) because practitioners hold tightly to the methodologies when their intention really is to hold on to the values. Or as Nichols and Smith (1998) expressed it, using certain methods comes with “substantive expectations, or a theory, regarding cognition and instruction” (p. 25).

By conflating method with value, practitioners let the tools (the methodologies) define the jobs (sound measurement) and not the other way around (Mislevy, 1996). All of us would agree that, if we’ve got a job to do, say, to replace a switch plate on the wall, we examine the job first and then choose a tool. We look to see if the switch plate is being held in place with flathead screws or Phillips screws. Then we go to the toolbox and get the right screwdriver. None of us would ever (assuming we had both tools available to us) go to the toolbox first, grab the Phillips head screwdriver, walk to the switch plate, see flathead screws and say, “Well, those are lousy screws! I guess I can’t replace the switch plate after all!” And yet, in educational measurement, similar things have been done. If Classical Test Theory, Generalizability Theory, or Item Response Theory cannot be used, then the job cannot be done. The problem is then that a lot of important jobs do not get done (Moss, 2004): important educational measurement opportunities are abandoned as unsound. Or they proceed without any documentation of the reliability of the scores as though reliability simply did not matter in that instance.

Perhaps a concrete example taken from Parkes & Giron (2006) would help to illuminate these issues. Ms. Baca’s fifth grade class of 24 students at Justin Otter Elementary School is working on an integrated unit around the theme of plants. It is early spring, so the students go outside on the school grounds and in pairs select a living plant specimen. Over a period of 2 months, students observe the changes taking place with their specimens. They record the changes through a series of measurements and by photographing them with digital cameras.

Each week students photograph the plant parts and measure and record the length and width of the bud. They also learn how to input numbers into the computer to produce graphs that illustrate not only the size of the bud but the difference in size. They then graph the width and length of the bud.

Each week they analyze the plant changes, write a two-paragraph description of them and present in small groups of four the changes as well as what their predictions are for the plant the upcoming week. The team members in the group provide each other feedback on their descriptions. At the end of the 2 months, each student produces a written report about their observations, and the group does a presentation to the class about what has happened with their plant.

The grade for this project will consist of three parts: a grade for each student’s written report, a grade on the group presentation, and a grade for each individual on their collaborative skills.

- Written Report—This is scored by Ms. Baca using a rubric. The rubric is available to students from the beginning of the project. This represents 50% of the project grade.
- Group Presentation—This is scored by Ms. Baca and by classmates using a presentation rubric with comments. The rubric is available to students from the beginning of the project. A “trimmed average” (dropping the highest and lowest ratings) of these scores represents 30% of the project grade.
- Collaboration Skills—At least once each week, each student rates the other three team members on the Collaboration Rubric (see the appendix) and includes comments in reference to the small group work. At least twice during the unit, Ms. Baca also completes the rubric on each student. An average of all of these scores represents 20% of the project grade, but the earlier scores are weighted less than the later scores, and she drops the lowest set of scores.

How should Ms. Baca consider the reliability of the collaboration skills scores, for example? Ms. Baca most likely does not have training in even rudimentary psychometric methods for estimating reliability, and there’s good reason to suggest she should not be trained in them (Stiggins, 1991). Imagining, though, for just a moment that a g-study were conducted, it would have an occasion facet, a peer-rater nested within subject facet, a teacher-rater fully crossed with subject but nested within occasion facet, and a rubric dimension facet. This is possible, but complex, in generalizability theory. There will likely be missing data due to absences. It is also going to suffer from small sample sizes and hence large standard errors on variance components. There are additional, more specific, issues that will be addressed below, but in the end, the expectation would be a fairly inaccurate generalizability or dependability coefficient. Does that mean, however, that Ms. Baca’s evaluation of collaboration skills is, in fact, unreliable, and/or does it mean this is poor measurement and she should find another procedure? Under the currently available psychometric approaches to reliability, Ms. Baca’s assessment is in trouble. Similar complexities and problems would arise in a Classical Test Theory approach or an Item Response Theory approach to this measurement scenario.

**Boundaries of Current Methodologies**

So exactly how are current approaches to reliability evidence mismatched tools for this job? The assumptions and conditions on which they are built, their very strengths in many situations, cause them not to be appropriate for all situations (Moss, 1994, 2004).
One of the most powerful aspects of the classical test theory approach and its extension to generalizability theory is that it uses sampling theories. But that also comes with strong implications—and limitations—to the use of these theories. Thus the central tenet is that items, tasks, raters, occasions, etc., are each sampled from a universe of observations and that they are independent and interchangeable. If they are interchangeable, then they should only differ in their measurement of the true score by sampling error. But this almost never obtains in educational measurement, as human raters who have not undergone extensive rater training demonstrate. A favorite example is Diederich, French, and Carlton (1981), who had 500 essays read by 58 different judges and found that 94% of the essays received seven different scores on a 7-point scale. People are different and will see an essay in different ways. They are not, in fact, interchangeable. In many situations, like many large-scale writing assessment programs, making them more alike is very important, leading to extensive rater training and calibrating, etc. At a certain level, good writing is good writing (or good enough), the characteristics of which can be agreed upon by large numbers of people and which other humans can be trained to see and score. That has been the story of inter-rater reliability throughout its history. But consider Ms. Baca’s approach to assessing collaboration skills. There is a degree to which the sampling theory implications are appropriate. The class is learning what collaboration is; there is a rubric that helps to define it; there is a shared class definition of it and learning targets associated with it. Therefore, each student should be expected to be an interchangeable rater of another student’s collaboration skills and any variation among them is due to sampling error. And yet, two elements make “rater calibration” unrealistic if not antithetical in this scenario. First, these are learners of collaboration skills, so they are each going to be understanding collaboration differently as they continue to acquire the concept. Second, the vagaries of human interactions in groups will not be eradicated through calibration. Petty arguments, the fickleness of fifth grade friendships and loyalties, the change of moods on a daily basis, all mean that the rater by occasion interaction term will likely be large. Therefore it is unrealistic to expect rater calibration—here, the learning of collaboration skills—to overcome these effects totally. And some might argue that it is the perception of collaboration that is being assessed through peer ratings and thus those personality state and trait elements are part of the perception, that is construct-relevant. One student’s actions on a particular day may be perceived as helpful and collaborative by one peer but as bullying by another. Thus such variation may be true score variation (construct-relevant), not error variation. Or, put differently, this is not measurement error but different true scores (Nichols & Smith, 1998; Seddon, 1988). But current reliability methods have no way, really, of parsing the nuances inherent in that situation, of separating variation—perhaps within the same source—that is a reliability concern from that which is a validity concern. So does that make Ms. Baca’s assessment of collaboration skills unreliable, or does it mean there is no tool for capturing the reliability? Smith (2003) has wondered if internal consistency might actually be counterproductive in reaching the measurement goal. And as Wolf, Bixby, Glenn & Gardner have already written: “We will have to find a previously uncharted course between insisting on uniform judgments and mayhem” (1991, p. 63).

A second real restriction that is shared by Classical Test Theory, Generalizability Theory, and Item Response Theory is the assumption of unidimensionality. The implication of this assumption is that all constructs of interest are single continua or the orthogonal combination of single continua. For decades, that view was commensurate with the atomistic views of psychology (Mislevy, 1996; Shepard, 2000; Willson, 1991). As psychological constructs began to expand and to gain complexity from cognitive psychology to sociocognitive to sociocultural, unidimensionality matched less and less. Using the orthogonal combination approach helped ease the tension, but that is too clean a conceptualization of something like collaboration skills. Ms. Baca’s rubric has nine dimensions on it under three headings. But no one understands these nine dimensions on the rubric to be independent, interchangeable traits nor does anyone propose that collaboration decomposes to these nine traits orthogonally. And yet current reliability methodologies must assume so. In a generalizability study, for example, they must be viewed as either a random facet—interchangeably chosen from a larger universe of collaboration skills—or as a fixed facet—the only nine skills that could describe collaboration. As Wolf et al. suggest, there must be some middle way.

Ms. Baca is conducting good formative classroom assessment that, from an “assessment as learning” (Earl, 2003) standpoint has many commendable features. So should the tool (e.g. generalizability analysis) dictate the job (assessment of collaboration skills)? At the moment, what choice is there, other than to ignore reliability—value and evidence—entirely?

If the values of reliability were separated from the methodologies, however, a path forward takes shape: the values can be defended and upheld through the continued development of new methodologies.

Expanding Reliability Methodologies

In order to expand reliability methodologies one must move outside of the constraints of sampling theories and dimensionality assumptions. That necessitates recasting the idea of a replication, since replication is conceptually, not just computationally, inherent in the construct of reliability (Brennan, 2001). Words like dependability, consistency, and accuracy imply replication of performance. In traditional reliability methodology, however, replications tend to be defined structurally, that is, by counting the pieces of the assessment structure. Thus two multiple-choice items constitute a replication, as do two raters, two tasks, two occasions, two dimensions on a rubric, etc. Recently that began to change as more conceptual—structural replications were used, particularly in generalizability studies. Rellying on cognitive learning theories, items were categorized into the different kinds of thinking required to answer them, and those different categories were used as levels of a facet, for example (e.g., Nichols & Smith, 1998). Such a replication is structural because it is still item based, but the replication is defined through some conceptual tie among structures (i.e., items), not by virtue of the structures themselves. This approach in g-theory does help answer useful questions about sources of
variation and allows cognitive variables to be examined as a source of variation in scores. It also has simultaneously reliability and validity implications (Brennan & Johnson, 1995; Kane, 1982).

The problem remains that each rater in a collaborative group to some extent through her participation in the same group discussion has some common ground while at the same time has her own unique and specific perception of another student’s collaboration and see aspects that no other rater sees. And both aspects are construct-relevant. Such a situation requires a kind of “partitioning” that statistical models are not particularly good at because they rely on structural units.

Is it possible then to remove the structural aspect of replication entirely and speak of a conceptual replication? Not how many occasions were sampled throughout the 2-month period, but what commonalities and trends appear from those occasions? Can Ms. Baca and/or the student point to two examples within one group session and one example within a second session as most exemplary of the “listening to other teammates” and explain how his score for the project relies on those more than the other occasions?

Moving to conceptual replications, though, requires much greater care in the definition of a replication. It is no longer as straightforward as pointing to the eight group meetings during the project period as “replications”. This is where contextual factors and theoretical considerations become critical.

Ms. Baca is able to articulate contextual factors that do not make each group meeting an equal replication, and she can weight the occasions in her overall judgment in a nuanced way. She can weight later ratings more heavily than earlier ones. That makes each occasion unequal and non-interchangeable as structural replications.

And yet, collaboration skills are collaboration skills, and each group session does to some extent serve as a conceptual replication. A human judge can make the fine distinctions of “weighting” the different sessions—seeing which specific interactions within a group discussion serve as a replication and which do not—in making a judgment about the extent to which the student has mastered “making fair decisions”. Thus theory and contextual factors also help make sense of findings that are inconsistent across structural replications. Reliability is fundamentally about consistency or inconsistency of performance (Brennan, 2001), though Smith (2003) is thought-provoking here. But current statistical approaches only examine whether there is inconsistency, not why there is an inconsistency. Generalizability theory, for example, permits great articulation of sources of inconsistency (raters or dimensions or occasions and all their interactions, for example), but it does not offer an explanation for those inconsistencies. A generalizability analysis may reveal that task by occasion variation is the largest contributor to error variance, but it offers no explanation as to why that is. And that is where theory, “naturalistic analyses” (Mislevy, 1996), and context come into play. As Delandshire & Petrofsky (1994) express it: “Given the nature of interpretation, the variability in candidate performance, the variability in interpreting standards, the differences in judges’ positions in relation to the standards, however, it may be necessary to think of consistency in a way different from that traditionally conveyed in notions of reliability” (p. 16). Thus there is a role for theory and context as well as evidence providing the makings of a Reliability Argument.

**Making a Reliability Argument**

Validation is now considered a specialized form of scientific investigation wherein one relies on theory and evidence gathered with guidance from that theory to support an overall validity argument. The evidence does not, indeed cannot, exist on its own. Reliability should work the same way. There should be a reliability argument that, consistent with a theory base (Nichols & Smith, 1998), relies on evidence to make an argument about the consistency of judgments. No one would contend that, if it is impossible to calculate a correlation coefficient or a confirmatory factor analysis, there is not a valid inference from a test score. Theoretical arguments about the construct being assessed, qualitative studies of consequences, and passing rates, for example, are admitted in validation work. Why then, if it is impossible to calculate a reliability coefficient, is the information considered unreliable or that its reliability is “unknown”, which amounts to the same thing? If reliability, like validity before, were expanded beyond evidence to theories and arguments supported by evidence, a broader array of educational measurement and assessment situations could be demonstrated to be reliable enough to gain legitimacy, psychometrically sanctioned, widespread use.

Casting reliability in this fashion raises the issue of how reliability and validity are related and what is the interplay between them. More will be said at specific points as the idea of the reliability argument is developed, but there are some overarching issues that should be addressed now. There is no room here for the intricate issues about the relationship between the two as were offered in the long-running discussion following Moss’ (1994) paper. Nor is there room for a discussion of the philosophical and practical implications when the two “blur” (e.g., Brennan & Johnson, 1995, p. 12), as interesting as that discussion is. For the purposes of proposing reliability as argument, it should be sufficient to ask whether, though the implications that reliability has on the validity argument are well understood, the reverse has been as carefully considered. That is, how often are the implications of validity arguments and evidence for the reliability analysis considered? Generalizability theory does require this kind of thinking (e.g., Kane, 1999), but perhaps not explicitly. Shavelson, Baxter & Gao (1993), for example, discuss the person by task interaction term in a g-study in both the generalizability and validity sections of their paper. This connection has also arisen in an Item Response Theory application where Heller, Sheingold, and Myford (1998) demonstrated that raters’ use of extraneous criteria or failure to use relevant criteria in justifying their scores related to poor inter-rater reliability. Therefore, here at the outset, it is important to note that measurement decisions or arguments can have simultaneously validity and reliability implications. Inevitably, reliability and validity are going to become intermingled—not blurred because the implications for each remain distinct—as the concept of a reliability argument is advanced. But this should be seen as a great improvement. The flexibility provided in
a reliability argument that is not available in current methodologies should permit validity arguments and reliability arguments to complement each other, not limit each other, to complement and not contend. Much of what is proposed in this article requires that one start with the validity questions and work back to the reliability implications. These will be noted along the way.

So what exactly is a reliability argument? A reliability argument would have these six components:
1. A determination of the social and scientific values of dependability, consistency, accuracy, etc. most relevant to the scenario at hand.
2. Clear statements of the purpose and the context of the assessment.
3. The definition of a replication in the particular assessment.
4. A determination of the tolerance or level of reliability needed.
5. The evidence.
6. The judgment: Pulling it all together

**Which Values Apply?**

As was made clear at the outset, the social and scientific values are still important and requisite in assessment situations. The first step in making a reliability argument is to clarify which of those values is most important and appropriate in a given measurement situation so that in subsequent steps of the argument methodological decisions can be made to best evidence them. Take a venerable example, the value of stability in scores. Given the nature of psychological traits, for example, stability over time would be an important value whereas consistency among items at a single point in time, being inherent in both state and trait measurement, would be a little less central. In this example, the methods for evidencing these values are themselves venerable: a test-retest reliability coefficient (i.e., a coefficient of stability) for the former and Cronbach’s alpha (i.e., a coefficient of internal consistency) for the latter. There are some values under the umbrella of reliability, however, for which ready sources of evidence are not as available or accepted. For example, what of values like consensus and coherence (Bulterman-Bos, Terwel, Verloop, & Wardekker, 2002)?

Ms. Baca has her ratings as well as peer ratings of collaboration skills that might not initially agree. But what if Ms. Baca were to talk to the students about their ratings and/or have them talk to each other? They may come in to the conversation disagreeing about a student’s skills, but may reach some consensus after sharing their observations, their justifications for their scores, and their developing understanding of collaboration skills. The measurement community is used to this approach in the well-known iterative standard setting procedures and in rater calibration, for example. Consensus demonstrates a consistency of scores.

There is also the coherence of collaboration across a number of group discussions. The student herself might be asked to point out examples of her ability to listen to others. Or the other students could. Ms. Baca may also have used the collaboration rubric in previous years and thus would have a sense of its stability over different groups of students.

Thus the first step in making a reliability argument is to articulate which particular value is most appropriate in a situation. It is easy to think of methods at this stage (e.g., what is replicated?) rather than of values (e.g., Is consistency or stability more important?), but it is critical to stay focused on the value itself to avoid the confusion of method and values discussed earlier. The argument needs to progress from values to methods and not the other way around. Even in situations where current reliability methods are appropriate, this is necessary, though seldom done explicitly. Though internal consistency estimates are often employed for their relative ease, they demonstrate a different value of reliability than test-retest estimates do. In some measurement situations, test–retest estimation makes sense as an operationalization of the specific reliability value of stability, as was discussed, while in other situations, like measurement of learning growth, such a coefficient is antithetical. Learning, by definition, is unstable across occasions. Thus the careful thinking about the construct being assessed that is the hallmark of validity also has implications for the identification of the most germane reliability value and, subsequently, the choice of reliability analysis.

**Purpose and Context**

If tools will be chosen that fit the job at hand (Mislevy, 1996), then a clear statement of the “job” is necessary. In this phase of a reliability argument, the reasons the reliable information is needed and the situations in which the information will be both collected and used need to be articulated. The purpose of the assessment is central to the definition of validity in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) but also is a fundamental issue in reliability as well.

The purpose of the assessment dictates the degree of reliability required and the overall strength of the argument and evidence necessary. This aspect of reliability was brought to light during the performance assessment debates about exactly how important reliability is (e.g., Messick, 1994; Moss, 1992, 1994). In the peer assessment of collaboration skills, several choices Ms. Baca make would affect the answers here. She could have decided to make the ratings purely formative and not part of the project grade, or, alternatively, she could have made the ratings worth more than she did. It is currently unclear how much of a marking period grade this project represents. Each of these issues will have different requirements for the degree of requisite reliability, the strength of the argument, and the kinds of evidence presented.

Context is one of the key components in a reliability argument. As was presented earlier, traditional reliability methodologies have, at best, very crude ways of taking context into account, and hopefully the use of reliability arguments permits a more nuanced, supple, and pliant use of context in forming a judgment about how reliable a score is.

Context plays a large role in Ms. Baca’s evaluation of the reliability of the peer-ratings of collaboration skills. She will know what the interpersonal dynamics are among those students. She may know that one of them was having a “bad day” during the third observation. She may use the written justifications of ratings to provide additional context to those scores, and she could even ask the student herself about a particular set of scores, a common procedure in writing portfolios, and others like the AP art portfolio that Mislevy (1996) describes.

**Defining a Replication**

Replications also remain important, but now that they are conceptually rather than structurally defined, they are more
pliable and will fit into more situations. But this also requires that there be an explicit decision about what does constitute a replication.

Brennan (2001) is clear that even under current measurement theories researchers and measurement specialists have the obligation to define what constitutes a replication. With structural replications, however, that onus is easily handled. With conceptual replications, there needs to be more explication. When is this decision made? Who gets to decide? What contributes to the decision? These are also very much validity issues.

During the planning of assessment is the first opportunity to define what will constitute a replication. Ms. Baca made several decisions about the number of group discussions, the number of raters and ratings of collaboration skills, and the specific tasks assigned during those discussions. She determined ahead of time that she would ask the students for some justification.

There is a large role, too, for post hoc refining of replications. Ms. Baca can decide to ignore the ratings of each other from two peers who never got along. She can choose to emphasize one particular discussion when all students were truly focused.

Another consideration here is who gets to decide? Does Ms. Baca always? May the students contribute to the decision? Make the decision? Does strict consensus always rule? There are issues of power, authority, ownership, and learning involved here that should be thoughtfully navigated.

And now that it is clear that there is a decision, which has to be made sometime by someone, what should be the basis for that decision? Theory and context should be considered.

Theory is most important because it is the articulation of the construct to be measured and the guide, first, to what constitutes an instantiation, a sample, of the construct, thinking traditionally done as validity. So hopefully Ms. Baca has learned something about group dynamics and the parameters of true group tasks so that she can design tasks wherein true collaboration can be observed. And she needs this to be able to teach it to her students. Teachers may not always have a strong theoretical understanding of the content they are teaching or of their students, but there is evidence to suggest that they should (McMillan, 2003; Niemi, 1997).

Context is the local or immediate residence of the construct. What are the exact questions on the collaboration skills rubric, and what is the nature and quality of the interaction the rater has with the rated student? There is a Bayesian-like quality here, where the replication lies somewhere between the abstractness of the construct of collaboration and the single group discussion. Ms. Baca iterates between consideration of the construct and the group discussion itself to arrive at exactly how well it reflects collaboration.

When does this process end? When does Ms. Baca say “enough is enough”? When there is sufficient information (Smith, 2003) to reach the decision required. And again there is the issue of who gets to decide.

Required Tolerance

Not every educational decision is of equal importance, thus not every decision requires equal reliability. As Kane (1996) expressed it: “Ultimately, the magnitude of measurement error needs to be evaluated relative to the tolerance for errors in a particular context” (p. 356). In making a reliability argument, however, it is a very important consideration. How reliable is reliable enough for this decision?

Smith (2003) expresses this as sufficiency of information. When is there enough data to have reached a dependable, consistent conclusion? Has Ms. Baca seen enough group discussions herself and enough peer ratings?

This should be based first on the purpose of the assessment. In order to give specific feedback and guidance to the student on how to improve their collaboration skills Ms. Baca may need lots of information, detailed information. This may end up being a year-long assessment undertaking. If, however, she is more interested in the snapshot, summative purpose of assigning a report card grade, the reliability needs are quite different.

The Evidence

The last part of the argument is the evidence that supports the claim of reliable information. What data will one show that demonstrates reliability? Often this could be traditional reliability evidence like rater agreement indices, for example. But it might also include other forms. Mislevy (1996), albeit more broadly than reliability alone, proposes both statistical and naturalistic analyses. Ms. Baca explaining to a parent why she feels her assessment of the child’s collaboration skills is accurate might provide narrative evidence. That is, she might describe to the parent the four or five observations, formal and informal, which she made and describe the commonalities in collaboration skills she has observed.

All current reliability methodologies would be admissible provided they fit the situation. But any other information, which points to the consistency of the decision or scores, would also be admitted. Thus reliability studies become a specialized form of social science research just as validation has so that any social science research methodology could conceivably be employed as evidence of reliability.

Pulling It All Together

In a court trial, the attorneys spend time preparing their case: determining what evidence exists, what additional evidence to collect and to present, and threading it together with their argument about the guilt or innocence of the accused. Then, at the summation stage of the proceedings, they speak to the jury to pull it all together: what is their argument about what happened and how does the evidence support that argument? The same thing must occur in a reliability argument, though in much more mundane a fashion than Matlock or Perry Mason would make it happen. Ultimately, the person or entity responsible for the scores would make the argument. Ms. Baca will be able to explain to the student, administrators, or parents what her judgment about collaboration skills is, and what evidence supports that judgment.

Who is responsible for making this argument? The answer here gets sticky in some respects, but there is precedent in validity arguments as the Standards (AERA, APA & NCME, 1999) articulates the issue. There is a role for the developer of assessments and a role for users of assessment information. The developer has responsibility for aspects of the reliability argument relevant during the design phase and the initial intended uses. Anyone who moves beyond those initial intended uses then bears the responsibility of making a fresh argument.
for the new purpose and, if necessary, collecting additional evidence.

Ms. Baca’s Reliability Argument
Now that the components of a reliability argument have been articulated, perhaps an extended discussion of how Ms. Baca might make her argument for the overall project score might prove illustrative. Here is what the reliability argument might look like.

The Value Base
The total project score is a composite of three separate scores, and the value inherent in each is different. For the written report, the value is consistency across the different reports. Ms. Baca does not wish to be lenient for some and harsh for others or to permit any of the known rater effects to impact her scoring. With the group presentation scores, the value is consistency or stability across different raters. And for the collaboration scores, the value is consensus.

Purpose and Context
The purpose of the overall score is to contribute to a student’s grade primarily in science. Given the amount of time involved, it may well be a substantial portion of the grade.

Context has an impact on all three components of the assessment. For the written report, Ms. Baca is aware that context is working against her. She will use all of the advice she got in her classroom assessment class to prevent rater effects from having undue influence on students’ scores. She knows that seeing the students’ names, the order in which she reads the reports, and her own comfort with the rubric all may introduce error into the scoring process. For the group presentation rubrics, she knows that having students rate each other’s presentations is a good way for them to continue to internalize the criteria of a good presentation. She also knows that there will not be uniform understandings of the rubric among the students, that some students will be more attentive than others to the presentations, and that the typical friendships and animosities among classmates will all influence their scores, which is why she uses the trimmed mean approach. Finally, with the peer ratings of collaboration skills, she knows that the vagaries of daily life in the classroom mean that any particular student will be better on some days than on others and that students’ collaboration skills will improve as they do collaborate. That is why she is weighting later scores more than earlier ones. She also knows that interpersonal dimensions play a role. So she asks students to write a justification for their ratings on the back of each rubric. This confronts them with the criteria for good collaboration: holds them to a standard of giving evidence for their judgment; and provides her with additional context information when looking at the scores.

Replication
In the group project scores and in the collaboration scores the replication comes as raters. There is also a replication of occasion in the collaboration scores. Ms. Baca has built in the “trimming” to dull ratings that may be inaccurate due to some bias. With the written report, replications come as Ms. Baca is reading each report looking for instances of Procedures, for example. Though there is a section of the report labeled “procedures”, she is not viewing it as a single instance of procedure but rather is evaluating each statement within it. So the replication comes as she identifies instances within the procedures section from which she draws a conclusion regarding the quality of the procedures section.

Tolerance
Since this likely represents a large portion of a student’s grade, the overall score needs to be reliable. There is a need for this to be one of the more reliable judgments that Ms. Baca makes.

Evidence
Ms. Baca actually has quite a bit of evidence to present. She has all of the scores from the rubrics as well as the comments from them. She also has all of the procedures she has used—the trimmed means, the scoring procedures themselves, and the procedures for writing comments.

Judgment
In the end, Ms. Baca can look at the scores and determine that they are reliable. She could argue that she designed her procedures to enhance the accuracy and stability of the scores throughout. She would mention that she trained students in the use of the rubrics, that she required them to justify their scores, that she used a variety of techniques like having students put their names on the backs of their written reports to mask the students’ identities. Then she could move on to describe the way she trimmed means to prevent bias from influencing the average scores. And, in the end, she satisfies herself, her students, and anyone else who asks that she has reached a reliable score for their overall project.

Conclusion
Of course, teachers are not going to engage in all of these activities for every assessment. The purpose in proposing reliability arguments was not to make the lives of teachers even more complex than they are, nor was it to write a “how to” piece with practical advice. The goal in proposing reliability arguments here has been to broaden the conceptual underpinnings of reliability practice to make room for additional methods and methodologies to develop. There are large categories of good assessment being done that do not currently conform neatly to the strict assumptions of current reliability methodologies. That puts those assessors into one of two uncomfortable, if not untenable, positions. They must either make their square peg fit the psychometric round hole, or they must “ignore” reliability all together. Neither is satisfactory because there are always reliability concerns in measurement even when there are no methodologies to evidence them.

Conceptualizing reliability as argument permits all current approaches to exist and reminds practitioners that even with tried and true methodologies, they have an obligation to define what scientific value is most relevant to their measurement situation, what constitutes a replication therein, and what the purpose and appropriate tolerances are. But this conceptualizing also permits other methodologies that would address reliability in other kinds of assessment situations to be developed. Though classroom assessment has been the context of this present discussion, it is important to note that reliability arguments are actually necessary in any educational or psychological measurement or assessment situation. That similar discontent with existing
methodologies has been expressed in medical education (Schuwirth & van der Vleuten, 2006) demonstrates the universality of the issue. The boundaries of the current methodologies have been best viewed in the classroom context, however.

Finally, casting reliability as argument is an invitation for new methodologies to be developed and accepted. There is room immediately for debate and discussion about the need for reliability to be viewed as argument. Such a conceptualization has strengths and weaknesses that need exploration. One of the issues that requires additional explication is the point at which reliability and validity become hard to distinguish, or perhaps better stated, when something like “who gets to decide what constitutes a replication” or “theoretically what should constitute a replication” is simultaneously a reliability and a validity issue. There is also the need for additional examples, both practical and theoretical, of how reliability arguments would play out in everyday measurement scenarios.

Acknowledgments
The author would like to acknowledge the contributions of Tilia Giron of the University of New Mexico to the development of the Ms. Baca example as well as the reliance here on Parkes & Giron (2006) presented at the National Council on Measurement in Education. Further, the anonymous reviewers and editors Steve Ferrara and Susan D. Brookhart made very useful and influential comments.

References


Delandshire, G., & Petrosky, A. R. (1994). Capturing teachers’ knowledge: Performance assessment a) and post-structuralist epistemology, b) from a post-structuralist perspective, c) and post-structuralism, d) none of the above. Educational Researcher, 23(5), 11–18.


## Appendix: Collaboration Rubric*

<table>
<thead>
<tr>
<th></th>
<th>Beginning 1</th>
<th>Developing 2</th>
<th>Accomplished 3</th>
<th>Exemplary 4</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contribute</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research &amp; Gather</td>
<td>Does not collect any information that relates to the topic.</td>
<td>Collects very little information—some relates to the topic.</td>
<td>Collects some basic information—most relates to the topic.</td>
<td>Collects a great deal of information—all relates to the topic.</td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share</td>
<td>Does not relay any information to teammates.</td>
<td>Relays very little information—some relates to the topic.</td>
<td>Relays some basic information—most relates to the topic.</td>
<td>Relays a great deal of information—all relates to the topic.</td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Be Punctual</td>
<td>Does not hand in any assignments.</td>
<td>Hands in most assignments late.</td>
<td>Hands in most assignments on time.</td>
<td>Hands in all assignments on time.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Take Responsibility</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fulfill Team Role’s</td>
<td>Does not perform any duties of assigned team role.</td>
<td>Performs very little duties.</td>
<td>Performs nearly all duties.</td>
<td>Performs all duties of assigned team role.</td>
<td></td>
</tr>
<tr>
<td>Duties</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participate in</td>
<td>Does not speak during the science conference.</td>
<td>Either gives too little information or information which is irrelevant to the</td>
<td>Offers some information—most is relevant.</td>
<td>Offers a fair amount of important information—all is relevant.</td>
<td></td>
</tr>
<tr>
<td>Science Conference</td>
<td></td>
<td>topic.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share Equally</td>
<td>Always relies on others to do the work.</td>
<td>Rarely does the assigned work—often needs reminding.</td>
<td>Usually does the assigned work—rarely needs reminding.</td>
<td>Always does the assigned work without having to be reminded.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>**Value Others’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viewpoints</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listen to Other</td>
<td>Is always talking—never allows anyone else to speak.</td>
<td>Usually doing most of the talking—rarely allows others to speak.</td>
<td>Listens, but sometimes talks too much.</td>
<td>Listens and speaks a fair amount.</td>
<td></td>
</tr>
<tr>
<td>Teammates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teammates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Make Fair Decisions</td>
<td>Usually wants to have things their way.</td>
<td>Often sides with friends instead of considering all views.</td>
<td>Usually considers all views.</td>
<td>Always helps team to reach a fair decision.</td>
<td></td>
</tr>
</tbody>
</table>

*Adapted from a rubric retrieved on 3/6/06 from http://edweb.sdsu.edu/triton/tidepoolunit/Rubrics/collrubric.html.