Module 6: Consensus-Based Assessment

by Dr. Tim Brophy

Hello, my name is Tim Brophy, Director of Institutional Assessment here at the University of Florida. Welcome to Module 6 of our Passport to Great Teaching series. This module is on consensus based assessment approaches. There are three primary goals in this module. First, we're going to take some time to explore the emerging conception of individualized assessment. Then I'm going to describe consensual assessment and consensus moderation, and then explain and operationalize consensus moderation as an assessment process that we can use in the work we do every day here at the University.

So there is an assessment quandary that's posed by creative and open ended responses. Let's explore that for a little bit. We know that data collected as evidence of learning is really delimited by the assessment type. You collect certain amounts of data depending, or certain types of data depending on the kind of assessment that you're conducting. For example, if you are giving an examination, you're going to collect scores. If you're assessing something using a rubric or preset criteria with certain levels of achievement, you're going to collect that information.

Now a consensus based methods present a non-delimited approach to assessment. So we're going to discuss that in this module. Here's the quandary. We know that we really cannot standardize the assessment of creative or open ended responses. We can expect certain parameters to be explored in the response, but we can't really standardize it, because that would go-- with that would fly in the face of why we actually ask our students to create something new or write something in an open ended way.

We also know there is no such thing as a standardized student. We know that everyday when our students walk in our rooms. But there's also no standardized response to an open ended or creative assessment task. So even our best attempts, though, at developing standardized rubrics, or scoring approaches, really are somewhat limited in their scope, especially in this context, and their transferability to practice. And some still actually have raised questions of validity.

So if variability, we continue to systematically ignore our variability in these kinds of open ended and, creative responses individuals then become kind of synonymous with statistical averages, which doesn't really make sense in these kinds of assessments. So faculty and researchers then lose their ability to account for this, the very processes that they're seeking to help describe and explain the phenomena that they're looking for. In this case, it's something new, something open ended, a creative response or creative prompt of some kind.

So let's think for just a minute about the science of the individual that's put forth by Rose and other scholars. This is an approach to understanding and analyzing human behavior

that's really based on the precept that individuals behave, learn, and develop in distinctive ways. And they show patterns of variability that are not really captured by these statistical models that we have. You can't find a mean, or a median, or mode, or a reliability coefficient that corresponds with standard statistical models when we're looking at these kinds of measures. Especially when we're considering individualized types of responses.

These authors also ask us to consider that human beings are really dynamic systems. Wouldn't that assumes that behavior's actively organized and context dependent? That is, we behave in certain ways depending on the context that we're in, therefore, variability as a response, in the responses that we give us humans to various open ended and creative tasks is going to be a natural outcome of the work. And that's kind of what we expect.

Also, these authors argue that learning is really not a linear progression, and that you go through in a universal type of sequence, where the starting and end point are predetermined because of the individuality of each person who is going through this process of learning and developing. They're going to go through it at different rates and different paces, and produce different levels of work at different stages of their development. And that we understand intuitively. So what happens is that creative and open ended kinds of tasks that we want to give our students, and their responses, kind of fall into this category of really not being statistically analyzable.

So what do we do? Well, we need to reconsider the approach to assessment for these types of responses, based on what we understand about human variability in response and human variability in all our behaviors. The standard methodology where we aggregate our data, then analyze it using some kind of commonly used statistical methodologies, instead we probably want to reconsider and analyze the information first, and then see how it actually can be categorized and aggregated in some way that makes meaning, and is meaningful for us as professors in these kinds of situations.

So why would we then use consensus based assessments? I've just given you some of the conceptions that underlie this idea. Well, we know that most of the assessments that we design are measured by some kind of predetermined set of criteria, or by us determining or developing a set of questions that are worth a number of points. The number of questions that students get correct adds up to a number of points and total score. That's applied to some kind of scale.

So we know that this is how we determine achievement. But not everything that we do as professors in our work falls into that kind of a category. Because there are some assessments for which the establishment of preset criteria or scoring procedures really is counterproductive. Because what happens is by setting these criteria in advance, we actually constrain the response the students give us to fit those criteria. So by conforming, then they therefore conflict with our original intention to get something creative and new.

So the kinds of assessments we're talking about are those that really examine individual distinctiveness or creativity, and therefore, result in an expected amount of response

variability. So these kinds of assessments include things like interpretations of either musical performances, or interpretations of dramatic readings or poetry. Or also creative writing, such as writing new stories, novels, poems, et cetera. Or musical performances, or musical works such as compositions, improvisations, choreography, paintings, sculpture, ceramics, any other fine art. As well as the development of new theories or logical arguments and so forth, whatever requires the student to put together their knowledge ensemble in a new way and create something new for us to read or to experience.

So in these cases, setting criteria in advance for these types of assessments could lead to response conformity, and that conflicts with the intended purpose of the creative or open ended assessment. We want to capture the individuality of the respondent, but we can't do that if we've preset the criteria for everybody to conform to. So there is the conflict.

But we want to talk now about consensual assessment. Because this is a process by which we can measure creativity in a different way. Now Teresa Amabile in 1996, first put forth this idea. So it just rests really on the belief that the validity evidence that professors can obtain from the information they collect from students in these types of assessments is really strongest when the experts-- us as professors, or those who we call in to do the assessment if they come in from the outside-- will use their subjective judgments. So this gets a little bit interesting, though, because raters are going to use a set of predetermined criteria or dimensions or expectations that the response is supposed to contain, but determine the levels of achievement based on a scale they themselves develop. And they're based on their subjective interpretation of the work that they're reviewing.

Now the judges in these situations measure these creative products of interest or these open ended responses in complete isolation. They don't collaborate with one another. And it is the responsibility of the person who has designed the assessment, who collects the information, to triangulate all of that information in a way that makes sense, and also captures the variability of response that the judges may have. And then make a subjective holistic judgment based on that information.

Now in this case, this type of approach, consensual assessment, is indeed critical. I mean, excuse me, interrater reliability is critical. And it's been shown to be fairly strong, acceptable in some recent studies, but it's not been shown to be acceptable in other studies. So the jury is still out on the application of those kinds of statistical methodologies to determine the reliability in this particular approach.

So now I'm going to talk to you a little bit about consensus moderation. This is another way to rethink our approach to these types of measures. When we're measuring creative learning or open ended or creative works in some way, so we can actually accommodate individual variability. And in a way, this is an extension of and slight variation of consensual assessment that we just discussed. So let's review first what we mean by-- why we set criteria for assessment to begin with.

We know that we can do this in one of two ways, and we talked about this in a previous module in this series. So you want to go back and review that if you would like. But there are two ways we do this. Analytically, we can do this in a rubric that lists criteria, describes levels of achievement fully, so that we can place students in levels of a set of achievement based on our subjective judgment based on those criteria.

Then there's the holistic rubric where we specify the levels of achievement, but we include the description of the achievement. I should say the description of the achievement includes multiple criteria. So those are the two ways, analytically and holistically, we can develop rubrics. So why do we do this?

Well, there are some reasons why we do it. First of all, we believe that students have a right to know how the quality of their work is going to be judged. So that's a perfectly reasonable reason to develop these preset criteria. Also, students' responses to the same tasks should be assessed according to the same criteria. And we believe that, and that's a fairness issue. Also, the criteria provide guidance to the students so they can actually build their response to conform to the criteria, which is good in some situations.

And then fixed criteria can add to the objectivity of the responses. So it does reduce or eliminate subjectivity to a degree. And then the set of criteria do provide, though, a convenient and economical way to provide feedback to our students. So by saying, you didn't meet this criterion. Or you could have done this to meet this criterion better, then we're able to provide some feedback and kind of gives us something to hang on to when we're doing that. So these are all reasons why we do this, and they're all good reasons.

However, we need to reconsider the use of preset criteria in certain instances. And when we set the criteria in advance, what happens? Well, first of all, while they look like a good idea at the outset, they become really quite difficult to separate. Because often the criteria and creative and open ended responses, students can demonstrate them simultaneously. Or one can be demonstrated more than another. So it becomes very interesting how the individual students' responses can match the criteria that we've set.

Also, a lot of our actual feedback to students can fall outside the established criteria. One of the things that is most interesting is when we sit down with a rubric and we have a piece of work that a student has written or created that we're using this rubric based assessment tool to evaluate and assess. We often find that there are elements in the work that didn't quite make it into the criterion list that we have in our rubric. So our responses often fall outside of those criteria. So the criteria we were specified in advance really aren't helping in all instances, and some of those that I've just described.

So let's define this idea of consensus moderation. Well, we know the consensus is just simply reaching a general or common understanding. And then moderation by definition is the lessening of extremes. Now there are ways that moderation can be done. There are several ways. One is you can average different readings or coded judgments, however, they are coded. Perhaps in a rubric, you give a number of a level, has a number associated

with it, 4, 3, 2 or 1. You can average those. We could remove the most discordant judgments, simply strike them, and averaging the remainder.

Then we could also accept the middle judgment, whatever that may be. We kind of say, OK, we have four different sets of judgments here from four different reviewers. And we'll take the middle one or the median. And then the other way, the fourth way that we're going to double in into more fully here, is actually discussing the responses amongst the reviewers until a consensus is reached.

So consensus moderation is the result of successful consensus seeking which makes-- which reduces discord, I should say. And by that means, then it moderates the final result. So when you have a group of people who get together to discuss the student work of interest, they're going to discuss their opinions of it, their assessments of it. And they're going to moderate and come to consensus on the overall judgment.

So let's talk about consensus moderation as an assessment process that we can apply. Well first of all, there have to be multiple experts. Those are us as the faculty, or those we bring in to do the judgments for us when we're reviewing an artifact or a work. And then we have to be open to the qualities in the work that are observed. So there is no attempt to steer the student toward any particular qualities that we're seeking. We want the student to create this and to develop this as an individual expression.

And then we do as assessors make a holistic judgment about the level of proficiency or competence in this process. But we do it as a group in a consensual way. Continuing on, the reasons for our judgments follow the judgment. This invokes the criteria that we've established. So we give our reasons. So we make a judgment of our-- make a judgment on our decision of whether or not the work meets our expectations or what we were looking for.

Then we give our reasons. And those reasons will invoke the criteria that we have applied. The criteria are salient to the judgment because the idea of this is that the individuals who are assessing the work and moderating with their colleagues have an idea of the limitless number of criteria that could be applied to the work. But the work itself brings forth the criteria that are relevant for judgment.

Now let's talk just a little bit about justification. It's not the same as the rationale that the judges can apply. Because justification will set out the grounds or the reasons for the judgment, whereas the rationale is a defense of an assessor's judgment. So if somebody challenges my judgment or my justification for giving a particular decision on the assessment of a particular creative or open ended work, then they challenge that, I can give my rationale for that. So I simply am describing and defending my judgment in that way.

So students, though, need to be inducted into this process. They're not familiar with it. As a matter of fact, they're so used to certain established assessment models in our courses

and our daily work that we have to help them learn to monitor and control the quality of their own performances and productions while their productions are in progress. So this model, it fits very well in the arts, in design, in theory building, this is what these content areas, and professors in these content areas, do naturally.

So let's think just a moment about reliability and validity again, which we covered in an earlier module. Now we know that the consensus moderation process, as well as consensual assessment, can provide strong validity evidence, and that there is discussion amongst the assessors to arrive at consensus. So the validity evidence comes forth in that discussion. So the interrater reliability is also high when consensus is reached. So we can all agree that the reliability is there.

So let's think for a minute about consensus based approaches based on the discussion I've just shared with you. So think about this. Where in your assessment of student work and in your course, or in your program, is consensual assessment or consensus moderation the best approach? And then, how might you triangulate your other assessment data from exams or quizzes, et cetera, with the results of a consensus based assessment? So I'm going to leave you with those ideas to think about and show you the references that I referred to in this particular presentation. Thank you.