University of Florida College of Medicine
Program Evaluation Committee
February 26, 2021

Attendees

Dr. John Aris, Dr. Eric Black, Dr. Miguel Chuquilin, Dr. Lou Ann Cooper, Dr. Frank Genuardi, Dr. Grant Harrell, Dr. Heather Harrell, Dr. Matthew Ryan, Oliver Shore, Dr. Ashley Wright

Recorded by: Wendi Miller, Course Manager

Lack of Reliability in Course Exams in the Basic Sciences

Dr. Cooper and Dr. Harrell and other committee members have been discussing the lack of reliability in course exams in the basic sciences. There are notable exceptions such as Dr. Peter Sayeski. The first two exams of Foundations are fine. Dr. Cooper suspects the last two exams are troubling because of the small number of items. She encourages Dr. Aris to bump up the number of items.

Looking at best practices and assessment, the program evaluation committee needs to provide a snapshot of what the college is doing now and how we want to improve.

Dr. Cooper said that there have been problems with Gator Evals in the MS1 courses. Dr. Cooper wanted to see if the they were meeting our needs in the basic sciences. Dr. Cooper shows some examples of evaluations from the 2019-2020 and 2010-2021 academic years.

She shared her screen showing the Genetics and Health evaluations. She is less troubled by the instructor evaluation questions. She thinks this isn't the best way to evaluate the course director (Gator Evals) who doesn't do much teaching. It is fine for those course directors who do a lot of teaching in the course it is fine. She said that the questions don't fit the model. The Gator Evals staff said that we could add questions.

Looking at the example, Dr. Cooper said that the course went up from 4.3 overall to 4.5 overall. She thinks this is a pattern that we will see throughout last year to this year. She doesn't think students know how to evaluate courses in a strictly online curriculum. She thinks it bumps up evaluations. She thinks it might be legitimate based on how well the course directors handles the Zoom meetings and the questions the students have based on the Zoom.

Dr. Cooper shows the Foundations of Medicine evaluations. She said that with respect to course improvement and evaluation of the course, the comments are always more insightful and more helpful for improving the course than the numbers. But when comparing the longitudinally within a course and compare between courses, the numbers can show a pattern.

Dr. Cooper said that there are four questions don't target what is done in medical education with respect to the courses. The Provost Office has said repeatedly that we don't know how to write questions. Dr. Harrell asked if we could change the questions or not use them. Dr. Cooper said that we can add more questions.

Dr. Cooper said that she is fine with others thinking the four questions capture what we're looking for in course improvement. Dr. Cooper shared the standardized questions that were used for pre-clinical courses. There is a lot of data showing the performance of the courses over time within a course and between courses. The questions were selected based on what is seen in the accreditation standards at the time. The standards were focuses on learning objectives and learning activities and methods of evaluation. The standards also require mid-term course formative feedback which his difficult in the basic sciences.

The Gator Evals overall evaluation shows a sum of all the items over all the questions and the responses. That is not the same thing as a gestalt about how the course went. She thinks that Gator Evals were designed to elicit positive feedback

rather than corrective feedback and ways to improve. Dr. Black asked if Tim Brophy or Jon Jordi have examples about questions that other programs may be using.

Gator Evals is showing a higher for the faculty evaluations. Several faculty are over 4.5. Our faculty scores tend to be higher than their department averages, higher than their college averages, and higher than the university averages. This shows we have very good people in place to direct these courses.

Dr. Cooper asked if the committee would like to add additional questions that flesh out things that we think are more important for improving our courses. Dr. Genuardi said that the open ended questions are what really drive intervention much more so than the numeric scores.

Dr. Wright said that she utilizes the narrative feedback much more. She is concerned about feedback fatigue, would adding questions to the evaluation make it lengthier and provide enough value on adding that effort. Dr. Cooper agrees that this is true for faculty questions because you need to have a comparison to what everybody else in your department is scoring. She notes that she uses the numerical data but the comments are selected to include or emphasize. But it is harder to plan for improvement based on many pages of comments.

Dr. Black asked in the chat if we could add a qualitative question about organization.

Dr. Harrell said if we added "please rate the course overall" that's not really adding that question because the current way we're doing it the evaluation not asking that that it's just a generated number. Dr. Harrell said that is the most useful number of their gestalt of the course. Students would only answer that one question. She likes the old questions got at the organization because that can pick up specific problem which is a course director specific issue. She like the learning objective questions are useful, particularly for the phase one courses. She is in favor of adding the first three questions from the previous phase 1 course evaluations.

In Gator Evals students have to answer every question.

Dr. Cooper said that the Gator Evals question prompts aren't as thorough as we'd like. Our prompts asked to students to give specific comments on the strengths of the course, the weaknesses of the course, and then give suggestions for improvement. Gator Evals asks students what is something the student has learned from the course that they will take forward into their career.

Dr. Wright said that she got very positive feedback from the last two years, less formative and actionable feedback because of the way the prompts were written. She also always liked question 9: Overall, the instructors were sensitive to individual student differences such as gender, race, religion, sexual orientation, socioeconomic factors, ethnic origin, and students with disabilities. She says that sometimes something would come up she would handle it and see it in comments but then if the students overall still felt that the instructors met the criteria that was something helpful to look at.

Dr. Cooper said that question 9 was often paired with "so and so was didn't respect my time" or "so and so called on me in class and embarrassed me". Dr. Wright said that this could be helpful sometimes as faculty might have done something unintentionally and that they were unaware of how offensive their actions are.

The Gator Evals staff would like all the answers to be the same – strongly agree and strongly disagree instead of excellent and poor. Dr. Harrell said that with the accreditation looming, we will have an opportunity to pushback at Gator Evals to say that their evaluation is not getting the data we need for our accreditation.

Dr. Cooper said looking at question 2, please rate the effectiveness of the course director on the organization and management of the course, that it might be okay for the course director with significant teaching role to use their faculty core evaluation and Gator Eval as their evaluation. Students will usually comment on the faculty as the course director. Dr. Harrell said that this is problematic for promotion and tenure because one role may bring a course director down and make the comparative data not look right. She would like to find ways to prompt students to evaluate course

directors on their instruction ability. They will be able to evaluate the course director on the course overall in a separate option.

Dr. Cooper asked the committee where they would like to put the course director evaluation, whether it needs to go in the course evaluation or as a separate evaluation. She said in the past there was a form that just asked for the course director evaluation that gave space for comments. This could be something that we could continue to do. She worries that if the course director evaluation is added to the course evaluation in Gator Evals it might be hard to find the course director's evaluation when they need it. Dr. Harrell said that if the course director evaluation is separate it gives comparative data against other course directors as opposed to against teaching faculty.

Dr. Black asked if the course director evaluation could be set up in a Qualtrics survey attached to Canvas, that Dr. Cooper could directly inject Qualtrics into Canvas bypassing the faculty's ability to access the data but creating the opportunity for review.

Dr. Harrell said that she likes questions 1, 2, and 3 but that 4, 5, and 6 is more about what the curriculum committee should be assessing not what individual students are assessing as she thinks it's hard for them to give formative feedback. Formative feedback doesn't fit well with phase one courses. She said that the specific strengths and weaknesses questions and suggestions for improvement in the course director's evaluation.

Dr. Cooper said that she would try to get this going but that it would not start until August.

To decrease the burden on our students for evaluations, we give the core faculty, the people who do the bulk of the teaching, the core faculty, the regular Gator Evals questions. For the faculty that give one or two lectures, we evaluate lecture by lecture, assigning a third of the students to do each lecture.

Dr. Cooper showed a slide on Genetics and Health instructor evaluation in New Innovations. The first column Dr. Cooper highlighted is the lecture evaluation for the people who give one or two lectures. The people who get the core evaluations are in the second column. She said that we have been asked in the past if the students' evaluation is based on synchronous lectures or asynchronous lectures. This evaluation is short, four questions that gets at the things that a lecturer should do. The lecturers don't have much to say about the course content or organization, so we should not ask the students about that person's ability to organize and structure content. The questions are:

- Overall is an effective lecturer, facilitator, or discussion group leader
- Organized content logically
- Presented useful information at the appropriate level
- Encouraged critical thinking/active participation.

There are open-ended questions:

- Specific strengths of this lecture/event
- Specific weaknesses of this lecture/event
- Suggestions for improvement

We have received good responses from students. We received 39 responses for this particular evaluation. If we send out the evaluation to 45 students we still get a good response. Statistically, we can get a stable mean at 40. The mean at 40 is not going to change much from the mean at 120.

Having this kind of summary is a quick way to look at which of the faculty the students felt did a good job and which may need to look at more closely for ways in which they can improve.

Some faculty give a single lecture in multiple courses, but the evaluations aren't going into a place where the faculty can pull them for promotion and tenure unless he or she knows where the evaluations are at. Dr. Aris said that the importance of the lecture evaluation for the chairman is less important than the value of the data for accreditation. Dr. Cooper said that the faculty evaluations aren't used for accreditation. Dr. Aris said that the only time faculty evaluation

data is really significant is if there is a red flag, if the numbers are really low, but it's his opinion that the basic science chairs don't pay close attention to the numerical data.

Dr. Cooper said that we've tried to solve the problem by creating new FFE courses in Canvas that allow faculty to go in and look at their individual faculty's evaluations, and allows the individual faculty to look at their evaluations. Dr. Harrell asked if there was a way to test it and report back on how it works. Dr. Cooper asked Dr. Aris if any of his lecturers have used this service.

Dr. Aris said that when his faculty contact him for evaluations for promotion and tenure. He sends the evaluations out to the faculty automatically. He said some respond to say thank you and others don't respond at all. He said that the ones that respond say that the narrative comments are the most relevant, the scores are less relevant.

Reliability of Exams

Dr. Harrell asked if the meeting could jump to reliability of exams as she needs to leave early. Dr. Cooper said yes. She explains that the Cutter Richardson 20 and Coefficient Alpha are the same thing for dichotomous data and for data that scored right/wrong. In estimates internal consistency of test items from a single administration so there's no test/retest. Reliability refers to reproducibility of scores.

Low values are usually due to an excess of very easy or very hard items, this is one of our issues with our exams in the basic sciences. We have an inordinate number of items that are 90% or above in terms of difficulty. Are the items too easy or are the students learning the information?

Poorly written test items don't discriminate and you can give low reliability estimate based on violating the condition that the items test a unified body of content.

We are trying to spread out the distribution so that we get a variability and performance to be able to say a students Is ranked 20 verses a students who ranked 75.

Test length: more items the less reliable the test is. The more similar in abilities our students are, the less reliable your test is likely to be. It is recommended that we have a coefficient of .70 and above for locally constructed classroom tests. We approach that in some courses and some exams and then in other courses we are not doing so well.

Dr. Cooper is concerned more for courses that have one exam and the grade for that course is based on that one exam with a low reliability with that test.

Genetics has a 70-item exam in 2016 and 2017. The reliability coefficient was much higher than it has been the last couple of years. We are seeing that across the years that more students seems to be failing. We may have high means we still have a good percentage of students failing. Dr. Cooper said that exam one in Foundations is not the same as exam four. Students have gotten used to how the information is being delivered, how to study that information by the time exam four rolls around. It is not unusual to have more failures in the first two exams than the last two. She congratulates Dr. Aris for his test writing abilities.

Exam one has a an average of 62.8 items between 2016 and 2020 with an average of 8.4 students failing those exams. Exam two has an average of 56.8 items between 2016 and 2020 with an average of 13.2 students failing those exams. Dr. Cooper said she was not going to quibble too much with a reliability coefficient of .60 given that we have such a homogeneous population and a small standard deviation. She noted that in 2017 24 students failed exam two.

Dr. Cooper asked Dr. Aris if the content areas in the units have been relatively consistent in both the amount of time spent on them and subject matter covered. Dr. Aris said she was correct.

Dr. Cooper said that the two questions per hour of content may be wrong. Dr. Aris said that units 1 and 2 are nominally four weeks each but that the content is delivered in three weeks, and units 3 and 4 content is delivered in one week or maybe six days of instruction. Dr. Cooper said to improve reliability, Dr. Aris could bump up the questions per hour of

content to three questions. She said that by the time the students get to exams 3 and 4, fewer students are having trouble.

Looking at Research and Discovery, Dr. Cooper said is another course that is problematic because it only has one exam. The reliability is in the .4 to .5 range, she worries that we're making decision on people passing and failing based on a test that is not very reliable.

Dr. Cooper said that setting the threshold for passing at 75 is capturing people who are two standard deviations below the mean. Looking at the standard deviation in the mean, almost in every case, you're right at 75 when you take two standard deviations below the mean. The distribution scores are skewed towards the higher scores and that there's a long tail of the distribution. Dr. Cooper said that the course director for R&D could add more questions or different kinds of questions, but that this would be something discussed in the assessment committee.

Dr. Cooper's next slide is for Fundamentals of Microbiology and Immunology. Dr. Cooper said that when looking at a course with a course director who has been in that role for a long time, they do a good job with assessments. She uses Dr. Gulig's course as an example as it shows a reliability of at least a .6.

Dr. Harrell said that there is not a culture of blue printing an exam and deciding the key content. She said that if you were like Dr. Gulig who give a lot of the content, you're likely doing exam blue printed, you have a faculty member who is experienced writing the questions. She says that Genetics has a lot of faculty giving content which brings in a lot of different factors and a lot of people who don't have experience writing questions. They are operating on the assumption that all the questions are equally weighted, which may not be true. She thinks this factors into some of the differences we see on courses that have fewer lecturers and who are experienced writers.

Dr. Black asked how many items have been recycled by year. She doesn't but asked Dr. Aris if he could comment considering how consistent the number of items are for each year. Dr. Aris said that he recycles about 90% of the items. He asks the faculty to review the questions and to either revise or rewrite them. Dr. Black asked if Dr. Aris knew how well that was replicated. Dr. Aris said he doesn't know but he knows that Dr. Gulig has a bank of questions. Dr. Cooper said that he used to write a new exam every exam.

Dr. Wright said that in ICM it is very similar that about 90 – 95% remain. She too sends out the questions to the faculty. They might get new questions if there is new faculty.

Dr. Aris said that there is a lot of background work being done on exam security. Dr. Black said that exam security is always a concern but he doesn't see if worked out well. Dr. Chuquillin said that the mean doesn't change. Dr. Black said that we would see something indicative of that, students who have been performing poorly suddenly spike upwards. Dr. Black suggested that we track the items that are recycled and which don't, are we paying attention to the quality of the items.

Dr. Cooper said if you are sending a bad item to the faculty and they say go ahead and use it, do you then use it? Dr. Aris said that if an item has a difficulty index and a point by serial that aren't good, then he discusses with the faculty to revise the item. He then keeps the revised item.

Dr. Cooper suggests combining exams 3 and 4 to create a longer test. Dr. Aris said they either did that originally or it was discussed but the items goes up pretty high. She said if you had 80 items he'd have to worry about whether the reliability was going to be affected by having more than one body of content. Dr. Aris liked the idea of increasing the items by 10 or 15 and will work on it for the next academic year.

Summary:

- The committee supports adding questions 1, 2, 3, and 9:
    - Please rate the course overall
    - Please rate the effectiveness of the course director on the organization and management of the course.
    - The learning objectives of the course were clear

    - Overall, the instructors were sensitive to individual student differences such as gender, race, religion, sexual orientation, socioeconomic factors, ethnic origin, and students with disabilities.
- Changes to the evaluations won't happen until August