Improving Assessment Practices

David Eubanks

FURMAN UNIVERSITY
david.eubanks@furman.edu

These are annotated slides from the 2019 SACSCOC Summer Institute, “Improving Assessment Practices” 7/23/2019.

Revision history:
7/26/2019 First complete draft
7/29/2019 Draft for distribution after feedback

This slide deck is edited down from the presentation version and some new material and commentary has been added. It was produced quickly and should be seen as an informal discussion. The original slides had some content about practical tips for getting assessment done, which is somewhat tangential to the main topic. If you want that material, email me.

Acknowledgments

I have SACSCOC President Belle Wheelan to thank for the invitation to do this session, stemming from an email I sent about the learning assessment standards. This should not be interpreted as an endorsement of the contents of this presentation, but it does demonstrate that our Association welcomes intellectual debate on important issues.
Summary

We must reform assessment as it is practiced for compliance reports. The current culture of practice, including expectations, language, and peer review standards for assessment reports, delivers too little in return for the effort. Rules about what must and must not be done stifle innovation and reward conformity. While we churn away at report-writing, critical questions about student success go unanswered. While student learning outcomes (SLOs) are a wonderful tool for language-based analysis of a curriculum, data work with SLOs is problematic:

- measuring learning is difficult, but those difficulties are generally ignored in favor of completing reports,
- compounded by the need for many reports (i.e. a large number of poor research projects instead of a few good ones), and
- the fact that most programs cannot generate the prescribed types of data in enough volume to properly measure averages or proportions, which is
- exacerbated by unwritten rules that restrict data types to a few types, but yet
- excludes faculty from “just knowing” about their students and programs, requiring decisions to be made from the poor data.

It may be that the current system is the best we could have done under the circumstances, because of external forces. Undoubtedly, we should thank leaders of higher education—including the regional accreditors—for saving us from No College Student Left Behind or some
worse fate. But now we should pause to assess where we are, where we want to be, and how to get there. The current language from the Department of Education about peer review of student learning calls for change. Leaders across the spectrum of outcomes assessment (NILOA, AALHE, and AAC&U) are calling for dialog to fix a broken system. It’s time for assessment to be assessed, starting with the goals we hope to achieve.

It is also time for assessment to become a mature discipline. The burden of being a de facto professional organization for assessment practice has admirably been carried by our Association for years, but assessment is no longer an infant—it’s a twenty-something living in the basement. We can already see tension between assessment’s professional organizations and the standards used for compliance. Papers and public talks from the ranks of those organizations are calling for less bureaucratization and more value to students.

The tension works the other way too: the rules of assessment reporting conflict with what I see as a philosophical bedrock Principle: let the institution make its case in the context of its mission. Rules for report-writing set up so many barriers to effective empirical work that there is no case-making, only box-checking. This is out of step with our ideals, burdens institutions with an assessment “tax”, and—most critically at this moment—has an enormous opportunity cost: we have significant problems to solve that require data science to solve them, we need a free exchange of ideas. We need innovation.

---

The plenary sessions at the Summer Institute touched on themes that are inspirational, achievable, and desirable: goals like an integrated system for ensuring student success, for treating students and their aspirations and fates individually, and paying attention to students who are most vulnerable.

Achieving these ambitions means enlarging our concept of outcomes from learning to success. If a student doesn’t have enough to eat, the solution isn’t about pedagogy. In compliance terms, the chasm between 8.1 and 8.2 needs to be bridged, and institutions need the ability to do that in ways that suit their missions.

We must address the value for cost of education, which entails a more integrated and comprehensive view of the experiences and effects of earning a degree than can be accomplished with a typical SLO methods. Our goal should be an integrated system of understanding individual student success. We can encourage progress toward that goal by realigning the expectations for 8.2—by rethinking what that standard means and how it is reviewed.

---

This slide deck and commentary contradicts advice you heard in other sessions at the Summer Institute. It probably also contradicts what you read in books on assessment, and what the peer reviewers will tell you is the way things are done. My intention is that the arguments below be taken as an academic debate intended to reveal the usefulness (or lack) of those practices. If assessment cannot engage in critical self-analysis it leaves us with nothing but obedience.

---
I have spent my professional life in the South and am a SACSCOC patriot. I would love to see our Association lead the way to solve some of higher education’s thorniest problems. The real payoff, of course, is improving the lots of our institutions and their students.
After casting a wide net, Trudy Banta and Charlie Blaich did not find many good examples of demonstrated improvements from assessment methods. This is a higher bar than reporting requires, since they searched for demonstrated value from the process, not just an action intended to improve the situation. Still, this finding from two established figures in the higher education assessment world should be surprising. Institutions had been producing reports for decades by this point.

Another signal of a problem is that faith in higher education seems to keep declining, and I see no evidence that our assessment practices even have a response, let alone an effective one. The tools in use are outmatched by the needs we have—we need better methods.

Looking for Outcomes (2011)

We scoured current literature, consulted experienced colleagues, and reviewed our own experiences, but we could identify only a handful of examples of the use of assessment findings in stimulating improvements.

My article in *Intersection*, which had a circulation of only about 200 assessment staffers, was intended to prompt an internal conversation about our methods. But it quickly sparked a series of public exchanges that didn't seem to resolve anything. That isn't surprising, since assessment can mean different things (as in my triangle diagram on the first slide), and it's easy to talk past one another in public discourse. What is surprising is that there was no sign of internal debate, just a weary circling of wagons.

I have not seen any real response to the challenge that most “measurement” of SLOs is of poor quality. The replies have dodged the question or made excuses for why we don’t have to care about statistics.
This candid assessment of results “minor tweaks” appeared in Linda Suskie’s blog, which she posted to the ASSESS email list. It describes what I’ve seen many times in assessment reporting.

The “tweak the syllabus” approach is a common way to demonstrate an action for improvement. It’s a natural consequence of poor data and oversimplified causal models, e.g. find the lowest number in the assessment data and add some nominal fix. Phil Hill, who is an education technology consultant, calls this kind of thing the “more cow-bell” approach, a term I borrowed for the talk. But since that is an obscure reference, I’ll call it the “top-up method” in these notes: if the level of something is too low, we add more of it until we have enough. This is a common sense approach, meaning that it assumes there is a simple and direct method of doing the top-up, like putting gas in the car. Much of the time, however, student success has multiple causes and variations among individuals, which makes the top-up approach too simple to be useful.

Here’s an example of top-up thinking: students arrive at the university with a skill deficiency in math. So we make them take a remedial class in math. Now they are assumed to know the math.

That’s too simple a model. It ignores the reasons why students don’t know math, how they learn it, and what they’ll need it for.
The three main assessment organizations leveled public criticism of assessment practices in Stitt-Bergh, Wehlburg, Rhodes, & Jankowski (2019) and called for a dialog. They don’t diagnose the problem in any detail, but they noted the need to make assessment more useful to students, particularly those who may be underserved by higher education.
Does current assessment practice help institutions prove they are doing a good job?

“The answers were pretty uniformly no, despite all the activity colleges have engaged in during the last decade.”

"There are good reasons why faculty hate it. It's real and it's earned," Jankowski [NILOA] said.


At the WASC meeting, sessions intentionally invited criticism of assessment practices, and the NILOA director went further than the joint article just cited, specifying the bureaucracy of the assessment process as a problem.
In 2019 the Department of Education released a new Handbook for accreditors seems to call for less expense and difficulty in peer review system for SLOs. Quoted above is similar language from the proposed new rules. It takes aim at the peer review process, including other language challenging the need for consulting and software and “complicated rubrics.” I don’t know what the overall effect of the new rules would be—I only looked at the SLO-related pieces.

Summary
Assessment faces criticism from all quarters. It’s clearly time to review what we are doing.
This diagram may be useful in separating conversations in assessment that are about different things. Most offices have some kind of research project going on, maybe to improve retention, which relies on large data samples that are carefully analyzed. That way of knowing the world is what I’m calling empirical research, and is characterized by an underlying skepticism about the data and its meaning.

If empirical research is the “left brain” of assessment, the top right box is the “right brain.” It represents the complex decision-making that drives changes in academic programs. We hire experts in disciplines, and they understand their curricula, courses, and students better than anyone else. Assessment work can facilitate useful conversations that might not happen otherwise, asking questions like “what do we want our graduates to be able to do?” and “does the curriculum support that?” More focused conversations, often with the teaching/learning staff, may look for improvements at the course or assignment level. External review of programs also falls into this box, where colleagues from another institution review all the important aspects of a program to reach conclusions. The evidence might include examples of student work chosen to show the range of student performance.

The empirical and evaluative tool kits are very different. The former is analytical in the sense of breaking big problems into smaller pieces. However, this requires formalization and reduction that are not as helpful for evaluative decision-making. I think there is an unfortunate bias against subjective expert decision-making, and this can be compounded by a distrust of faculty that has emerged in the assessment community. I attribute that to the tension over report-writing. Some within assessment circles openly disparage the idea that faculty could intuit student learning, implying that a formal assessment is the only valid epistemology. That is dangerous, damaging, and demonstrably wrong. I think we would be better off acknowledging their expertise and trying to maximize the use of it, rather than over-emphasizing measurement, especially when measurement has little chance of success.
In contrast to empiricism, report-writing is designed to crank out required elements for compliance. Whereas empirical research adopts a skeptical attitude toward measurement (the data has to convince after interrogation), report-writing assumes that measuring learning is easy if you follow a few simple rules. You can’t accomplish the report-writing function without making that assumption due to the number of projects and small data sets.

Exacerbating the problem is the requirement of using measurement techniques instead of evaluation methods, even when the latter would be more effective. For example, a small program with about one graduate per year can still be required to submit what looks like a statistical analysis based on N=1, or imperil the institution’s 8.2 submission. A simple review of the curriculum by faculty (ideally including an external one) would be more beneficial.
I would love to have a heart-rate monitor to see what the average increase is when this slide comes up in a room full of assessment people. This is the specification for our public job review. If your compliance certification has no recommendations here, it’s a huge relief. If you flunk, it can be very stressful, even crushing. If the top two triangles on my diagram are like left-brain/right-brain, the report-writing function is the amygdala.

The important elements for this discussion are the three components of a program report.

**Outcome Statements**

Outcome statements can be very broad (e.g. critical thinking) or as specific as a single test item on a narrow topic. There’s a hidden ideological filter here. I doubt that you can get away with declaring that your learning outcome is “students should master the course material in program courses,” even though this is perfectly reasonable. The ideological issue is that outcomes statements have been elevated to a rhetorical status that far exceeds their usefulness. People in assessment have debates about the right verbs to use, for example (recently it was about whether “demonstrate” was appropriate, on the ASSESS email list).

The most debilitating aspect of the focus on learning outcomes is that it excludes other kinds of student success, like grades, course completion, retention, graduation, success after college, ability to repay loans, thriving in life (e.g. as addressed by Gallup’s index), affinity for the institution, and so on. A successful program that increases class attendance in a gateway course, which lowers the DWF rate, would be considered a success by any reasonable standard, but probably not for SLO reporting.

Outcomes statements are better suited as a good tool for working with faculty in the evaluation box (the right brain of the assessment triangle), but this function doesn’t require precise language. Fussing about the language of outcomes gives assessment the veneer of an academic discipline, but it almost always falls apart at the empirical stage.
My recommendation is to let programs define student outcomes as broadly as they want to, to include student success measures. This would allow integration of university goals and more effective means of accomplishing them. Let the relative value of different kinds of outcomes be determined by objective worth in the market of ideas. We need that flexibility to deal with the significant issues faced by higher education.

Assessments
The “extent to which” is almost always interpreted as measurement, and assessment culture is filled with references to measurement of student learning. Therefore, I am only going to focus here on quantitative approaches to assessment. In practice, these data are usually collected by rubric-scoring student work or by testing students. A single item on a single test might be used as an indicator of a student’s generalizable ability in some area, a procedural convenience that cannot be justified empirically--such a small amount of information is undoubtedly influenced by all sorts of factors that we are not interested in measuring.

As Trudy Banta points out, the researchers at the College Board spend huge amounts of money on the SAT instrument, which has only modest reliability (the College Board publishes validity studies, which you can find online). How much reliability could we expect from a single item on a single test? Or a rubric rating of a single paper.

Measurement is hard. Generating reports for compliance would be impossible if the recommended standards from statistics and measurement theory were applied, so instead of that we have a set of rules for what can and cannot be done. These rules have little to do with getting the right answer to questions, but make it easy to rate compliance with a checkbox approach.

Trudy Banta called for a scholarship of assessment at the first conference I went to in 2001, and while there are journals on assessment that include some nice research articles, actual report-writing practice—driven by the amygdala—is usually like a folk understanding of the world, looking for water by waving a forked stick.

Seeking Improvement
The new 8.2 is an improvement over the old CS 3.3.1.1 in that it clarifies that reports don’t have to demonstrate an improvement with data; they just have to show that actions that could reasonably be assumed to lead to improvement have begun. Language here is important. A tip from an experienced IE person I know is that you should never write “we are planning to...” in the report, but rather “we have decided that...” The distinction of future tense (bad) and past tense (good) makes the difference. I think this is probably accurate, but it’s unfortunate that reviews can turn on such trivia.

Two aspects of the “seeking improvement” requirement are problematic. One is that the decisions about what to improve must be linked back to the data from the assessments. I’ve already described in general terms why we should be very skeptical of the usefulness of that data, and this requirement effectively closes off the evaluative power of the program faculty. They are in effect not allowed to know something (in this context) unless it is verified by the data. Good faculty make improvements to their courses and programs all the time, but these are often not useful for reporting because there’s no epistemological cover in the form of assessment data. It may not exist, may not be good enough to be useful, or may contradict what the faculty think is the case. Normally this last case would be considered a challenge to the
validity of the data, but because of a bias against faculty judgment, it works the other way around: the assumption of validity stems from using an approved process, not critical inquiry. For example, if faculty unanimously believe that the academic program would be improved by infusing more theory throughout the curriculum, this has no official usefulness unless such is “blessed” by formal support by numerical measures. This inverts the actual usefulness of the data, from playing a support role to being the sole way of knowing allowable. If you try to make improvements without the nominal warrant of an assessment measure that points in the same direction, you run the risk of a citation for the lack.

The second problem is that no research project—even when well-designed—is guaranteed to tell us something interesting. The data resulting from the assessment process may be ambiguous, untrustworthy, or tell us that everything is fine. There are no exceptions for the data-driven actions for improvement. In practice, you can probably get away with 25% or so of the programs failing to produce such actions (depending on the review committee), but the requirement is that all programs meet this standard.

The unfortunate consequence of this combination of requirements is that the best strategy for compliance is to generate numbers by whatever means are allowable, choose the smallest one and “improve” that outcome with one of Linda Suskie’s “minor tweaks.” For example, if you use a rubric with five dimensions, one of them will be lower than the others, and you can use the top-up method, e.g. “since the rubric sub-score on Style was lower than we would like, we will add a new lecture and assignment devoted to that in the new syllabus.” I’ve never seen anyone say what is being given up when the new material is added.

Undoubtedly, many reports unintentionally use this strategy. As we will see, small sample sizes, natural variation in outcome levels, and measurement error will lead to outcomes measures that have a lot of statistical uncertainty. The result is a pedagogical random walk.
In addition to the Association’s voted-in standards that appear in 8.2, there are unwritten rules that must be complied with if you want to get your reports approved. This is unusual in that standards generally speak for themselves, and institutions have the opportunity to make the case however they like. Of course, institutions have enough rope to hang themselves, as the saying goes, and should pay attention to traditions and norms. But hard rules are generally either in the standards themselves or in a formal policy, e.g. with academic qualifications.

Grades
There is a blanket proscription against using course grades as measures of learning for program assessment. A common argument is that grades conflate specific learning goals with each other and with non-learning factors like conscientiousness in getting work in on time. On the first point, it is quite reasonable that a program would want to have a general outcome like “Students will demonstrate competency related to coursework,” and assess that with grades. This has no conflict with a specific outcome, since the outcome is whatever is learned in the course.

I have never seen that approach attempted, probably because the prohibition against grades is well known. Another argument against using course grades as assessment data is that the outcome is not specific enough to use the assessment measure to understand how to make improvements. In fact, outcomes are usually specific enough so that the top-up method of improvement will work: if critical thinking is low, top it up with more critical thinking exercises.

If our goal is increasing student success, the ban on grades is a significant barrier, and it has undoubtedly already costs us a great deal. The history of assessment would have been very different if at the beginning we had resolved to use grades as the primary outcome measure, and then spend our time trying to align grading practices with desired course outcomes. Instead, we have created a new parallel grading scheme via assessment that is supposed to be superior but in most cases cannot live up to that promise.
The ban on grades breaks the rule of data science that you don’t throw out data until you’re certain it’s not useful. The ban restricts many routes of inquiry that institutions could use to “make their case,” and does so outside the language of the standard.

Rather than making arguments about the relative merits of grades versus assessments, I’ll give a real counterexample to the primary claim that grades can’t tell us about learning. That’s a little later.

Self-Reports
We are not allowed to use student self-reports of learning, e.g. a survey item like “rate how much math you learned in this course” as a primary measure of learning.

The prohibition against all forms of self-reported data dismisses the value of subjective self-reported information, despite the fact that many academic fields that depend on this type of data for research, e.g. personality theory. Most of the ways in which we navigate the work depend on individual subjective assessments. The proscription against such data may be another symptom of the zeal for standardize-testing methods at the expense of more useful ways of knowing.

These graphs were not on a slide during the presentation, because I didn’t have enough time for this much detail. They compare freshman survey self-concept ratings in mathematics ability (1=low, 5=high) to their grade averages in certain math courses, disaggregated by gender (I eliminated any data points with N < 30). Notice that the lines go up in all cases—certainly an indication that students have some predictive power over their future ability to succeed in a mathematics course. The simplest explanation is that they can assess their abilities with some
accuracy. Also note that the Female lines are always higher than the Male lines. This suggests that on women underestimate their abilities relative to men, by about half a grade point on average. This is troubling, because it probably means that young women are closing off possible majors and career paths because of illusory barriers (“I’m not good enough to…”). In the big picture of student success, this is concerning and opportunity to make improvements (as Harvey Mudd College did with Computer Science). Such improvements are invisible to the usual SLO-centered activity of standard 8.2a.

The example clearly shows that on average, it is possible for students to rank their self-rank their abilities relative to other students as exposed by actual performance. And the bar we need to reach is not absolute knowledge of a student’s ability (which is probably unachievable by any means), but having enough information to make improvements. Clearly self-reports should not be arbitrarily ruled out.

The dismissal of student self-evaluation as meaningless is also at odds with our educational mission. If a student graduates with a wildly wrong estimate of his own abilities to bridge, shouldn’t we do something about that? Such a case implies that our feedback mechanisms are so poor that students have no sense of their own achievements—certainly an opportunity for us to improve.

The prohibition against surveys is at least understandable when we talk about skills like math or writing. Even if we ask the students, we’ll want another source in order to validate their responses, as I showed with course grades. But for attitudes, behaviors, beliefs, goals, and other questions for which the student is the unchallenged expert, we also prohibit surveys from being the measure. The usual advice for measurement in this case is to have the student write a paper about the thing you care about, then rate the paper with a rubric.

How would you feel if the waiter at a restaurant asked you to write a personal reflection about the menu so that it could be rated with a rubric to see what you want to eat?

One has to wonder how all the accumulated knowledge from behavioral science about the use of surveys has gone unnoticed in the creation of this rule. I know a lot of assessment people who routinely use factor analysis and reliability measures like Cronbach’s alpha to look at survey data. I’m not sure who is responsible for the essay-rating idea, but that research needs to be exposed to critical inquiry.

As with grades, the survey prohibition has a chilling effect on inquiry. Instead of starting with “our surveys aren’t giving us good information” and proceeding to “how can we improve them?” the economic incentives from the report culture push practitioners away from spending time on surveys as measures of anything.

See Appendix 2 for more on the practical usefulness of survey data.
This is concise version of an assessment report that I picked up from friends at Reinhardt University, but I don’t know the original source. I have mocked up entries in the boxes, which cover all the required elements.

It’s never certain that any one reviewer will like or not like a report, but I think that you’d have to be pretty picky not to accept this one. I did hear an objection from participants in the session that VALUE rubrics (i.e. LEAP) are not supposed to be summed up across all dimensions of scoring, so that’s a possible objection along validity lines—although if we start talking about validity all is lost.

If we take the report at face value, it makes a nice story about slipping scores that were partially fixed by an intervention—not only seeking improvement, but demonstrating it.

But as many participants in the session noted, the sample sizes are too small to believe that story. The program doesn’t have very many majors, and even in large programs the effort to collect data often restricts the overall N, so we may end up in this situation.

Note that the action for improvement is a top-up solution: if we don’t have enough of something, let’s add more of it.
Suppose you’ve regraded 10 randomly chosen senior papers with a rubric and averaged the sub-scores for analysis, evidence, format, and style. The graph shows a simulation of such scores from our 10 papers over five years. The underlying statistical model for the simulation is based on real characteristics of our writing assessment data, and programmed here to have no change over the five years. The variation you see in the graph does not represent an actual change in the average student outcome, but rather it represents variation due to other things, like varying student characteristics and experiences, as well as measurement error (e.g. inconsistencies in rating).

Because we are visual creatures, the graph is a poor way to represent the data. The dots representing individual averages suggest that these are precise numbers, and that the differences between them are meaningful. As we know this, is not the case.

Due to pressures to reach conclusions from the data, and sometimes the unfamiliarity with this sort of statistical estimation, it is tempting to create a narrative out of the graphs. For example, the purple line representing Style seems to have started low and zoomed up over two years, perhaps because of an intervention. But that wore off for some reason (maybe that new faculty member we hired?). And so on. None of this is valid, since the variation is purely due to factors unrelated to student learning in the model. Yet this is a common pattern that you’ll see in many assessment reports. It behooves us to be able to distinguish variation in learning from variation due to other reasons.

---

1 One objection from a participant that is averaging an ordinal 1-2-3-4-5 scale as if they are scalar values makes unwise assumptions about the thing we are trying to measure. This is true, and if you want to read about it and a possible fix, see Liddell & Krushke (2018). Other approaches include proportional-odds models and item-response theory. But with 10 samples, this is futile.
A better way to summarize results like this (averages or proportions) is to not just show the point, but also a reasonable estimate of how smeared out that point should be due to unrelated variation. This is standard practice in reporting statistics in most disciplines, but I rarely see it in assessment reports. Of course, there is little motivation to find ways not to believe the data are meaningful—to do so diminishes our opportunity to seize upon some plausible top-up action for improvement.

During the presentation I gave a live demonstration of these effects to illustrate the interplay between sample size, effect size, and types of variation. For this write-up I chose to include that in Appendix 1 so as not to slow down the narrative. You can also find a link to the simulation app there, so you can try it yourself.
Evaluate the Plan

A university plans to measure the change in global awareness for students in study abroad classes, with about 20 students in each. The proposed method is:

1. Ask students to write essays related to the outcome before the trip.
2. Ask students to write essays with the same prompt after the trip.
3. Score the essays with a suitable rubric.
4. Subtract average Pre- and Post- scores to determine the effect.

I asked the workshop participants to work together to evaluate this assessment plan. Then we discussed it. It’s an actual example from my work at Furman.

There were plenty of requests for more information about the rubric and essay prompt and so on. One important question is whether each experience is unique; yes—the travel courses are each their own thing. One of the intended uses is to be able to compare experiences—to rate them and rank them if possible, in order to identify particularly good practices.

In the actual event, a consultant was on campus who claimed expertise in this area and suggested a plan similar to the one outlined above, which you’ll note is the recommended replacement for using surveys. I asked the natural question: what should we expect the standard deviation of the results to be? The answer would tell me what effect sizes I could likely detect. Based on previous work with rubric ratings of student writing, I was dubious that the actual gains would be perceivable through the other types of variation.

The consultant seemed stumped, and didn’t have any references where this had been used, nor seemingly any familiarity with that kind of question. So I asked around on the ASSESS email list, and a nice person actually sent me summary data from just such an exercise. It showed a slight negative difference over the pre/post assessments, and generally confirmed my suspicion that the sample size was never going to overpower the combination of natural variation, error, and small effect size.

Note that improving inter-rater agreement only minimizes measurement, error, not natural variation. See my work on agreement statistics (Eubanks, 2017) for more on that subject.
I ran a test on our writing ratings to see the effect of sample size on detecting real differences in measured learning (Eubanks, 2019). You may recall from before that a .4 difference is about a year’s gain for our writing ratings, so the numbers in the .4 column of the table are the chance of detecting a year’s development in writing with samples of the size listed at left.

With 100 samples in each, the test gives the right answer 76% of the time. These numbers could be increased by setting alpha to a higher value. But generally, this confirms the rule of thumb (see Appendix 1) that around 200 samples is a good minimum threshold for this kind of data. In statistics this is called a power analysis. It is used before a study is conducted, to determine how large a sample size you will probably need in order for the test to be trustworthy.

---

2 It worked like this: first take a random sample of, say 30 and average them. Then take another sample of 30, add the effect size of say .4, and average those. We know that there’s a .4 gain built in that way. Run a t-test to see if we can detect the difference (alpha = .05); that is does the test reject the null hypothesis that the two samples have the same mean. The decimal number is the fraction of tests that successfully could tell the difference.
“Analyses presented here demonstrate that assessment efforts by smaller schools may never provide the statistical power required to obtain valid results in a reasonable time frame.”


This paper is a power analysis similar to the previous slide, and reaches generally the same conclusions. The authors argue for research to be done only when it's likely to yield trustable results, and look for general results that help everyone, the way medical research is done.
I wanted to see how big academic programs are in general, at the undergraduate level. So I pulled down three years of IPEDS completions data and filtered it to four-year degrees. For each institution I counted a unique CIP code as a unique program and summed up the number of graduates in that program to get the distribution above. I have left out all the zeros (programs with no graduates in a given year), but that was the most frequent category.

A possible flaw with this analysis is that a program might include more than one CIP code. Even so, it seems unlikely to raise the median too dramatically. Most programs seem to be small in terms of the amount of data they can gather from each graduating class. This finding aligns with what I see in assessment reports—lots of small samples. In conjunction with the last few slides, we should conclude that the statistical conclusions of most assessment reports are not trustable.

Source: three years of IPEDS completions data, omitting programs with zero graduates.
Here are some ways we can improve current practice, based on what we have seen so far.

- Start applying statistics and measurement theory where applicable;

- If such theory tells us our measurement isn’t going to be very good, then do something else;

- In practice, with so many small programs, the “something else” could include:
  - Getting more data, e.g. grades and surveys
  - Relying more on professional judgment
  - Create credible large-scale projects that can improve many programs through general changes, which extends to
  - Stop trying to assess every program as a stand-alone entity and adopt an integrated approach to student success.

Consider these numbers. The ten rubric ratings aren’t going to tell you much, but what might tell you something is, say, three years of all the grades assigned in all the courses in the program, augmented with survey data (with matched student IDs) and all the institutional data you want (admissions, financial aid, student life, etc.). Even a small program will generate thousands of data points pretty quickly that way. Standardize the analysis so you can just run the thing for all programs at one time. Now you have a platform for researching student success. And it’s much easier to maintain than the current system.
This is a great exercise to try with your students, colleagues, and random people you meet in the airport. We started doing this as a complement to the more formal research done on high impact practices. The definition of impact in those studies is determined by the researchers. We thought it would be good—before we get too sure of ourselves—to ask students what they think is impactful.

Some of the participants shared their stories, and I wish that I'd been able to take notes. One was validation from a teacher that echoed a story from one of the plenary sessions—the idea that mentorship can happen anywhere.
At Furman, we crudely quantify these impact assessments by asking students to choose among experiences for the most impactful one (if it exists). Such data would be dismissed as “indirect evidence” by the rules of assessment (see Appendix 2 for more on that), but the data turned out to be valuable.

The results you see on the graph are typical of what we’ve found from surveying students and alumni. Research, study away, and internships are the most frequently-cited experiences, and are in agreement with other research on high impact practices. And jobs and work-study always appear at the bottom. This seemed like an opportunity to us—why can’t we make work study more like an internship, so that it will be more impactful? We’re working on that now.

This example shows that trusting students as experts on their own lives can help us make improvements—without asking them to write essays about every topic.
In addition to the small sample problem, assessment report-writing has a problem with cause and effect. The top-up method of improvement assumes that the course is the sole cause of learning. Because the standards of statistical inference are usually ignored (small sample sizes being part of the problem), usual report-writing implicitly assumes that there is no variability among students and no measurement error. This set of assumptions is far too simple to solve most real problems.

With learning, and more generally student success, there are many factors that can influence the outcome. This is best demonstrated with an example.
This is the data set I used in *A Guide for the Perplexed*. It represents a general education outcome, and comprises several hundred ratings over two years. The results show that faculty are not entirely satisfied with the outcome, since about 17% of students didn't meet their lower threshold for language proficiency.

This is the kind of data that a program usually starts with—although in this case, the sample size is actually large enough to work with. I asked the conference participants to put their heads together and make a recommendation for to the department—what action for improvement is called for?

The usual response I get is that we need to look at the papers that were unsatisfactory and see if we can spot a pattern there. I didn’t do that, but I can tell you what you’d find: a higher rate of fundamental language errors. Since the top-up method of improvement always assumes it’s something about the course that needs improvement, one suggestion is to add a week of remedial work at the beginning, to let those students catch up. Or maybe encourage more use of supplementary instruction. This is the top-up solution.
I gave the groups more information. Above are two graphs that show average scores (on a 0, 1, 2 scale with 2 = exceeds expectations). The bottom axis denotes the year in school when the student took the language class. You can see that for both the languages shown, proficiency drops after the first year.

The graphs suggest that students generally should not wait to take the class, and that we should see improvements through different course scheduling. It’s actually a little more complicated than that. A regression analysis suggests that student grades and the timing of when they take the course interact, so that less academic preparation (via high school grades) and waiting to take the course both pose learning barriers, and that lower-GPA students are more likely to wait. That means we can be selective in advising students to take the class right away (low-GPA students for sure), while allowing students who can still do well after waiting a year to do so.

This example illustrates how complicated it can be to understand the outcomes we want, and why an integrated success model is desirable. It shows that the top-up approach to improvement is too simple for real problems, especially if we want to promote equity in education.
I promised to show that we can easily use grades to improve learning, so here it is. You can try this at your institution, and my guess is that you’ll find a way to make a significant improvement to student success rates.

The data from the previous two slides, on foreign language proficiency, comes from faculty ratings of student work. After I found that result, I tried the same thing with grades by computing averages for each introductory course, disaggregated by what year of school the student took the course—just like the previous slide except with GPA.

That method easily rediscovered the foreign language issue, and also turned up several more—probably the one’s you’d expect. There’s a penalty, on average, for waiting a year or two to take a course like Calculus I or introductory chemistry.

Some might object that I haven’t specified the learning outcomes, but I think it’s safe to assume that they are learning less calculus or chemistry and leave it at that. It’s no less specific that foreign language proficiency.

Note that we can’t do this kind of analysis unless we collect student IDs on the assessment ratings. Grades come with them already attached.
Here are some of our writing ratings, averaged by years in college (each data point is a term of attendance), disaggregated by high school grade averages divided into thirds.

It appears as though the lowest ability group (red) finishes college a year and a half behind the highest ability group (green). We have to treat any claim like this with some skepticism, and you can read more about the details in Eubanks (2019). However, this pattern pervades most of our assessment data, so we have created some program-level projects to see what can be done to get the low-GPA group off to a better start, and maybe a higher learning trajectory.

If we want to address equity issues in student learning or success, we need more data than we can get with the usual means of assessment, and we need it by student ID so that we can disaggregate or (more usefully) create regression models.
I had cut this slide from the presentation for the sake of time. It shows gaps in academic progress related to socioeconomic factors, do demonstrate the importance of these contributing factors. Note that the cluster of high school GPAs is pretty tight, but gets spread out for college GPAs, probably partly because of high school quality. I've found that number of AP courses taken can be a reasonable proxy for high school quality. I assume that “better” high schools offer more of those courses.

Notice that the lower SES groups tend to lag in credits earned. The study that produced this graph was part of a program to offer grants to lower-SES students to make up credits in the summer, so they can stay on track. Does this have anything directly to do with SLOs? No. Does it have anything to do with learning and student success? Of course it does.
This graph comes from some research I did using the Equality of Opportunity Project’s bonanza of data, to estimate how much institution type contributes to graduates’ salaries. You can find more information, including data and a report you can run for your own institution, on my code-sharing site: [https://github.com/stanislavzza/GraduateSalaries](https://github.com/stanislavzza/GraduateSalaries).

The graphs show the association between the listed variables and graduate income one to twelve years after graduation, using national data (very large sample size). It supports other findings that parental income is hugely important in the graduates’ salary. The graduation rate of the college attended is also important, but it is conflated with parental income since high-income parents tend to send kids to high-graduation rate schools. Marriage seems to have a negative effect on salary the first few years, but then becomes decisively positive. The red line reiterates other findings that women’s salaries are depressed compared to men.

The point of this slide in the presentation is to illustrate how big the equity gap is. We can’t apply a top-up method by just giving the student rich parents. If we’re going to level the field for them, we need an integrated model and good data systems to back it up.
This table comes from Sullivan & McConnell (2018). It shows a summary of a good-sized data set of rubric ratings that were scored by raters trained on the rubrics as part of the VALUE Institute at AAC&U. The left column shows time-in-college, from first year to fourth, and the other two columns name the respective learning outcomes being assessed using ratings of student papers.

We would hope that students learn at college and that the measurements would show a steady increase in learning (as with the Furman writing ratings shown earlier). However, this table doesn’t show that pattern. Only when the researchers added a third variable—an assessment of the difficulty of the assignment—did the expected pattern of growth appear. This work is being tested more explicitly with a follow-up to validate the findings, and Furman is part of that project.

So first, this is another example of needing a more complicated model to understand the world than a simple cause-in-the-middle one, just like Astin said in 1984. What’s remarkable about this study, however, is what didn’t happen. The findings challenge the validity of using VALUE rubric scores as a simple pre/post (subtract the difference) effect size. Since there are a lot of people using the rubrics, we could have expected the assessment world to have erupted in debate about what this result means for them. Do we need to go back and recalculate our gain scores now? As far I can tell from reading the ASSESS email list, the study made no impact at all on practice. It’s another sign that most assessment work isn’t serious about getting the right answer; it’s more about just following rules.
A program was unsatisfied with the performance of some of its graduating seniors, so they increased rigor across the program.

Following this, there were no more weak performances among graduating seniors.

Discuss the effectiveness of the change and draw diagrams of cause and effect.

This is another to show that we need to get beyond simple top-up models. In this case, the simple model is more rigor = more learning, which is one explanation of the result. But as participants pointed out, it could also be that more rigor = more attrition. If the less able students are dropping out, it will naturally only leave the best behind. We always need to look for selection effects when trying to tease out causes.
Part of the reason small samples of assessment data is indecipherable is because of natural variation in students and their experiences—this is good variation, because by doing more work we can start to tease out subtle relationships. (The bad kind of variation is measurement error.)

The point is that more sophisticated models can partially account for sample size limitations. As noted earlier, this depends on intentional efforts to collect student identifiers with as much data as we can, so we can link it to other sources of variation like demographics. All of this points once again to the need for integrated thinking about student success, not just a narrow focus on learning outcomes.

What We Learned

The variety of individuals and the uniqueness of their experiences is the source of much variation in scores.

Student success is affected by many factors, not just how a course is taught. Everything is connected.

To understand assessment data we need to take student characteristics into account. Grades are especially important.
The preceding demonstrates, I hope, that most of our assessment report-writing efforts are empirical in name only. They lack the raw material needed (enough usable data) and the tools (useful interaction models). It would be surprising if it were otherwise—even a small institution might have a hundred or more of these data projects going on (one for each outcome). We have quantity rather than quality, and this shows in the results: minor tweaks that are based on random fluctuations in averages.

But assessment reports still don the guise of measurement, and lay claim to its authority. For example, actions seeking improvement are supposed to stem from the data, no matter how small the sample, or how poor the measurement. An unfortunate consequence of that is to diminish the role of faculty expertise. Faculty knowledge counts for nothing unless it is “validated” by the numbers. As the NILOA director indicated, faculty distain for assessment is earned.

It’s really hard to imagine how we could design a system that would be more likely to produce reams of irrelevant paperwork. Unsurprisingly, more and more of it is being automated by software.

We need a reboot—a ground-up review of what we are trying to achieve and how we might likely achieve it. The left-brain/right-brain functions of research and evaluation, respectively (the top boxes on the diagram) are powerful means of improvement. They just need the right conditions to flourish.
I've taken data for this slide from the Chronicle article shown in the link above. The English department at Lehigh wanted to improve the career outcomes of graduates. They used their professional experience and some data to try to re-engineer the program, and seem to have made a difference. There is a selection effect here, too, which you can read about in the article. I didn’t include error bars because the samples are very small. That’s the point—we can achieve positive outcomes for individual students without large sample sizes, because there’s another half to the decision apparatus: professional evaluation. One benefit of working with an outcome like job placement is that we have little if any measurement error, so all the remaining variability is interesting.
Innovations from data science are all around us; the left side of the assessment brain has never been more capable. We can use that in combination with decision systems that trust faculty to be experts to work toward goals many institutions share. Perhaps topping this list is demonstrating the value that a college education has, and doing so in personalized ways. Net price calculators individualize the cost already; what’s the range of likely futures for a student like me? How do we make it easier to find the path, and does that mean a yellow brick road or a pith helmet and a compass?

The charge to integrate student success systems is timely and much needed. To do that we need to know lots of things about students that only they can tell us. Treating surveys as an important research tool is the most efficient way I know to do that. See Appendix 2 for more.

With the right leadership and the right infrastructure we can hope to breach equity barriers by understanding how to help student succeed.

---

**Wish List**

- Demonstrating to students and parents the value of college despite the cost
  - Demonstrate pathways to careers

- Integration of student success efforts
  - Learning, belonging, engaging, succeeding – all one network

- Equity—emphasis on individual student success

- Benefit from 21st century data science
Appendix 1. Simulations of Rubric Ratings

Introduction
In the session I showed a simulation of rubric ratings in order to demonstrate the effects of sample sizes and other parameters that affect the average ratings. As a starting place, I used results from my validity work on writing ratings.\(^3\)

You can find the simulation on my github (code-sharing) site: https://github.com/stanislavzza/assessment_sims. To run it, you’ll need R and RStudio installed, and the two packages “tidyverse” and “shiny”. All of that is free to download.

Demonstration

Pre-Post and Sample Size

\(^3\) A reasonable model of that data comes from a linear regression with time and grade averages as predictors. It gives \(R^2 = .3\). The standard deviation of scores is about 1. So I calculated reliability (in the classic sense) as the true score variance divided by total variance, or \(T^2/(E^2 + T^2)\), where \(T\) is the standard deviation of the true score and \(E\) is the standard deviation of the measurement error, which includes any variation that’s not the true score itself. These are parameters in the simulation that you can move around. So \(T + E = 1\), and we have two equations with two unknowns. With some algebra I got \(T = .4\) and \(E = .6\), which seems reasonable. These became the default settings. This method is crude and probably underestimates the actual reliability by some amount, but it’s a good start. A one-year difference in average writing scores is about .4, so I used that as the starting effect size, to ask how large does the sample size need to be in order to detect a one-year gain in writing skill?
The default settings show two bell curves representing the likelihood of getting a particular answer when we randomly select 10 students and average their writing scores. Imagine that the “pre” group is ten sophomores and the “post” group is ten juniors. If the students were matched pre/post, we’d need a somewhat different model that I’m showing here—so this doesn’t cover that case. It does cover the usual case where we look at a program’s assessment data over time, although simplified here to assume that there are the same number of students in each group. You can modify the app if you want to add more parameters; I tried to keep it simple.

The overlap in the two bell curves shows that a lot of the time the two averages will be close together. Ideally, we’d like those distributions to be separated more. By pressing the simulation button you can see what happens when we take the samples 1000 times, subtracting each time to get the change score.

You’ll get somewhat different results each time, since it’s random. The height of each gray bar represents the number of times we got a difference in that range of values. The correct answer is the dark blue line, and zero (no difference) is the red line. Notice that a significant portion of the time we get answers to the left of the red line. That means that by taking our two sets of ten, averaging and subtracting, the answer can easily be that the junior know less about writing than the sophomores. I have shaded in blue what I suggest is a reasonable range of values that we could consider reasonably good measurement. Clearly, with ten samples in each, we’re not close to that standard.

If we have a really large effect size, we might be able to detect the difference with reasonable accuracy, even with only ten samples in each. If I approximate a difference between freshmen and seniors (hypothetically a .4 times 3 = 1.2 gap in average score), we get something like this:
My code for the blue area can’t handle this case—the acceptable region should be wider than what’s shown unless we have a very high standard for accuracy. If we triple its width, we’re still fine. It’s reasonable to think that most of the time we could get a fuzzy estimate of the gap between freshmen and seniors with a total of 20 samples. Note the separation of the distributions at the top—they are pushed apart by the large effect size (difference in where they are centered).

But most of the time we aren’t looking for effects that large. Even a year’s gap is pretty large. For certain we’d want to know a group of seniors was a whole year behind the historical reference group!
If we crank up the sample size to 100 in each group, we get a reasonable measurement of the one-year gap. For this reason, my rule of thumb is that about 200 samples for typical data and effect sizes is about where to start looking at the statistics seriously. Below that, it’s likely to be such poor measurement that it isn’t worth the time. For sure, it doesn’t seem wise to make major changes based solely on small samples in these conditions.

As a final note, sometimes we have nearly zero measurement error. If we want to know the fraction of students who have attended another college or school after leaving our institution, we can get highly accurate data from the National Student Clearinghouse. Most assessment data will have a lot of measurement error, however.

Why is Rubric Data so Noisy?
Rubric data often has a lot of measurement error, so the combination of small samples (due to cost of rubric-rating) with low reliability is a recipe for random results. A particular problem with rubric ratings of student writing is that there is a hidden dimension to it. John Hathcoat addressed this topic at the 2019 AAC&U meeting in Atlanta.
The hidden dimension is that a given piece of student writing may not have enough information to tell you about the outcome you’re trying to rate. When we use rubrics this way, we design them to distinguish between the two: (1) how much evidence present, and (2) what does it say? John’s presentation convincingly showed (with lots of data analysis) that you shouldn’t ignore this issue.
Appendix 2. Surveys

These two questions seem to ask the same thing, but have very different distributions. We discovered this in our pathways-oriented research, as part of the PANDA project (the data are not KOALA-tative\(^4\)). The survey randomly administered one of these two items to students, and then we looked to see which one worked better. We want to see variance, because that’s how we distinguish between cases, and we particularly want to see variance that we understand. So in assessing item quality, we want to see a nice, centered spread (like the right graph, which is not skewed like the left one is), and that it distinguishes between at least one important demographic. You can see that there’s a gender difference, a credit hour difference, and maybe a GPA difference on the responses to the right question. So we dropped the left one and use the right one on most surveys.

One summer project for our IR/assessment team is to organize survey data so that it’s more useful. Even at Furman, which only has 2,700 undergraduates, a single survey can generate half a million data elements, and these can quickly become unmanageable— it takes too long to make discoveries. We now have a systematic way to warehouse the data.

---

\(^4\) You can thank Michelle Horhota for that.
This image shows part of the survey warehouse interface that lets me search for a prompt. You can see it at the top—the PANDA item we just saw the distribution for. The slider allows for some variation in the prompt, like if the period is left off, or an extra word is added. In this case, there are a total of twenty surveys where the prompt has been used (the image only shows the first four).

The data engine combines all this data, and since most of the surveys have student IDs, it attaches institutional data like demographic and grades information to the survey data. I can download all that with a click.

The next tab over shows the distribution, with options to disaggregate. Here I’m showing the responses by gender—one of the data elements automatically attached via student ID. We can see again the gender difference—it looks like women are more stressed out on average than men, about career transitions.
The next tab lets me look at the data as a time series, showing either means (using either the naïve ordinal scale or a latent model mean—see the Liddell & Kruschke paper for details) or proportions. Here I’m using year in school as the measure of time (can also use cohort or the time the survey was administered). It clearly shows the gender gap and suggests that stress peaks in the junior year.

The Survey tab on the interface lets us compare items within the chosen survey (PANDA).

I've sorted the items by their covariance with our chosen item (the one about the transition stress). The statistics here are used to judge the quality of information we get from an item. If it's not good enough, we drop it or fix it. Notice that the two closest items by covariances are both about stress or being overwhelmed, and all three show higher stress values for women (the negative values in the Gender column). They are also skewed somewhat toward lower GPA students, and toward the more senior years of attendance, except for the question about social life, which shows more stress for underclassmen.

All of this can be done virtually instantaneously, allowing for a free exploration of a very large data set. This is the kind of tool we need to understand the range of factors that influence overall student success, including success after graduation. We will add more features to this soon, including the ability to build predictive models and create “construct” clusters in order to reduce the number of things we have to think about (there are thousands of prompts, and we’d like a short list of about fifty constructs or outcomes).
Using this tool, I can quickly find out that sense of belonging is decreased by low grades and lower socioeconomic status, and that it predicts affinity at graduation (e.g. “was it worth the cost?”). This kind of knowledge is essential as we build an integrated platform for student success. If we treat surveys as a strategic asset and apply research methods we can get good results.

Direct Versus Indirect Measures

The prejudice against self-reports is usually hidden with technical-sounding terms, describing evidence as “direct” or “indirect.” This is not a terrible idea on the face of it, but once it became codified as a rule, it became a way to arbitrarily devalue data based on superficial qualities, instead of actually examining the usefulness of the data. As with the ban on grades, this empowers the assessment office (and peer reviewers) and disempowers other voices like faculty. It also strongly discourages good research on survey data, which is very unfortunate.

I am indebted to John Hathcoat, who posted the following to the public ASSESS email list during a conversation about these terms:

Unfortunately, there is a tendency in academic circles to treat these terms as a dichotomy when I think it is better conceived as a continuum. Messick (1994) states, “…the term direct assessment is generally inappropriate, especially in the behavioral sciences, and should not be used. It always claims too much…” (pg. 21)

Here’s more from that paper by Messick (a well-known validity theorist):

The portrayal of performance assessments as authentic and direct has all the earmarks of a validity claim but with little or no evidential grounding. That is, if authenticity is important to consider when evaluating the consequences of assessment for student achievement, it constitutes a tacit validity standard, as does the closely related concept of directness of assessment. We need to address what the labels authentic and direct might mean in validity terms. We also need to determine what kinds of evidence might legitimize both their use as validity standards and their nefarious implication that other forms of assessment are not only indirect, but inauthentic. (pg. 14)

Indeed, “direct assessment” doesn’t even appear in the index of the standard reference Educational Measurement from the National Council on Measurement in Education (Brennan, 2006). That’s probably because the concept of measurement is based on how to assess the quality of the tools and data with multiple perspectives: reliability and validity methods. The terms “direct” and “indirect” add no information by themselves; the value of a method for gathering data and the results depend on what it’s being used for, and cannot be generalized with sweeping classifications like “self-reported data can’t be trusted because it’s an indirect measure.”

Of course you can find examples of where such data really can’t be trusted, but that in no way proves some universal truth. In good data science we should let the data speak for itself.

I’d like to point out that the original research on high impact practices, a topic that now pervades higher education, was based on NSSE data—all self-reported by students, e.g. telling us about learning gains. Subsequent studies like Kilgo, Sheets, & Pascarella (2015) have shown effects using longitudinal standardized measures (some of which are self-reports), and our own work highlighted in this presentation also reinforces the idea that some practices really are highly
impactful. If George Kuh and colleagues had dismissed the original self-reports as “indirect” and therefore unusable, we wouldn’t be where we are now. But how many useful discoveries have been left uncovered because of the direct/indirect mindset?

It’s not just that self-reports are dismissed as unusable—that’s bad enough, but the direct/indirect prejudice by contrast gives anything that looks like a “direct” measure a free pass. For example, rubric ratings of student works are seen as direct measures, regardless of their statistical properties; that is, regardless of whether or not the ratings have reliability or validity for the purpose.

The total effect of the direct/indirect mindset is to severely diminish the ability to know what’s true. Sometimes a discussion will pop up on the ASSESS email list about whether or not a particular data source is direct or indirect. In the cases I have observed, these conversations turn on philosophical or rhetorical elements. I’ve never seen the quality of data raised as an issue, aside from the quote above referencing Messick. The tone is ideological, not empirical, and the role of compliance is always there with the question “am I allowed to do this?” Since reviewers could arbitrarily declare some new data source to be indirect, you might find yourself in hot water because of the distinction.
References


