Program Review and Assessment Committee

Thursday, October 25, 2018 - 1:30-3:00 pm, AD 1006

Minutes

---

1. Welcome, Review & Approval of Minutes (3 minutes)
   Congrats to Stephen Hundley and Susan Kahn for a successful Assessment Institute. Show of hands of those who attended the conference indicate a lot of people were in attendance.

2. Stephen Hundley introduced our speakers. "Fixing Assessment." Guests: David Eubanks, Assistant Vice President for Assessment and Institutional Effectiveness, Furman University & Josie Welsh, Director of Institutional Effectiveness, Missouri Southern State University (30 minutes)

Josie – most recent conversation with David about problems they see with assessment with the goal of making assessment more meaningful. What are we doing that is meaningful – we are doing a lot of activity and we talk a lot about data, but we don't have much to show for it. The Chronicle article had over 540 comments and the good news is that the assessment community responded and engaged. We are pretty excited to engage IUPUI's PRAC in this dialogue as well.

Much like what PRAC does we have faculty attend a Board of Governors meetings to discuss reports, findings. (Showed example report in WEAVE). These are then reviewed again in 6 months or 1 year. Example indicated that the expectation (80% of students meet the objective) was not met. Faculty reflections were documented (and shared). 5 of the 7 were scored at 4 or 5. The other 2 were scored at 3.8 and 3.16.

David – even at large institutions, you have a lot of programs that inherently don't have a lot of data. But, that doesn't excuse us from having to report to our accreditations. Showed a slide about "writing assessment data". "in my reviews, I've seen hundreds of reports like this. This is the kind of things programs are presented with and the assessment office asks them to look at it and determine what to do to improve. At best it's like a whack-a-mole. What can we do to improve analysis and writing?" The problem is that these are just random numbers and even in random numbers we try to find patterns.

Small scale studies we are forcing ourselves to interpret from noisy data. The interesting thing is that it's intellectually dishonest. The central limit theorem applies to assessment as well. How did we get here??? Why do we ask people to turn in reports when we know the reports contain randomness or junk? Is it a good thing that faculty want to get together to talk about their teaching?

K. Alfrey – part of the purpose of the assessment process is to stimulate those conversations and we hope that the assessment plan results in useful information. My argument is – there is value in the conversation, but you also need to have a dialogue about improving the assessment process.

David E. - it's important to note how hard it is to talk about assessment because it means different things to different people. 3 things: faculty teaching & learning (goals of the course, how they are doing), program evaluation (take into account everything we know about the program – budget, capstone), but the way accreditors think about it – explicitly based research project. Empirical project without any curiosity of whether the numbers are random.

Josie – these data do not support that conclusion. These numbers may well just be random. You do not have enough information to conclude that your students are not successful in this area. So, a dialogue is fine, but you don't have grounds to revamp the curriculum.

Why are we making faculty do all of this activity and forcing them to use these numbers when we aren't sure there is enough information to inform decision making. It seems like we've developed an orthodoxy of activity.

T. Roberson – at conference at Salt Lake City – D. Eubanks said, "the standard methodology of doing assessment where you count and fill out rubrics is a year-long process to dismiss the central limit theorem" "Deep faculty expertise – I don't need to count all of these things to have impressions about what we need to do". Impressions in the classroom in a 1 page report of a couple of problems and a few solutions. We can count, but we should turn to expertise.

D. Eubanks – the standard assessment practice has a bunch of rules (e.g., can't use grades, can't use standardized testing), which means we've narrowed what type of information we can use. So, throw out the rules and be more creative. Ex. – ask faculty to tell me what they think students are subjectively learning.

S. Hundley – concluding thought.

Josie – *What's broken in assessment*? Too many, too small projects. A culture that talks the talk about data, but doesn't walk the walk. Lack of attention to data science fundamentals. A generation of assessment directors who ignore statistics.

Why are faculty upset with us? We say one thing, but are doing another.

Measure learning---- ignore measurement theory

Look for learning gaps --- ignore the statistics of differences

Grades don't measure learning ---- make them regrade papers

3. Update on IUPUI Faculty Survey Results—Robbie Janik, Assistant Director of Survey Research and Evaluation (20 minutes)

   49% response rate! Even better than previous years. The overall summary report will be coming out very soon and it will be in Tableau. Included FT and PT faculty. No Medicine.

   Why did you come to IUPUI and the importance of that in your decision to come to IUPUI? Top 3: Climate and supportive atmosphere; Support for teaching. Support for research/creative work and research quality; competence of colleagues.

   Would you do it again? Pretty similar across faculty type.
   Job satisfaction – See slide for the highest (there were 50 options). Women tenure track faculty are less likely to be satisfied with service load. PT and Assoc are happy teaching here (not their salary).

   Longitudinal (2015 used a 4 point scale; 2018 was 5pt, so cannot compare): The good news is that Tenure and tenure track faculty are much higher in terms of overall job satisfaction, autonomy and independence.

   Satisfaction – faculty development and mentoring. Mentoring is low within their department. FD opportunities during teaching not great.
   In addition to preparing you to succeed, see specific satisfaction items. Those 0-3years (more recently on-boarded) are more pleased with on-boarding for research and service). It's a small n, but comparing men to women, men indicate being more satisfied.

   Confidence in going up for P&T – men are more likely to be confident. 75% indicate they have a mentor for the P&T process. For those that don't have one, they said it's because their unit didn't help to facilitate the relationship. And those that have a mentor, it was an informal relationship.

   Part-time/Assoc – What would help you? They want to be included more in their department (meetings, what's going on in the school, more money).

   Career goals at IUPUI – most everyone feels that what they do is valuable and worthwhile. Clear sense of purpose.

   K. Murtadha – appreciated the distinctions in the questions between SL and CE.
   L. Bozeman – questions about global learning? R. Janik will make sure you get them in the full report.

4. PRAC Grantee Report—Michael Golub, Purdue School of Engineering and Technology (20 minutes)
   Dept. of Mechanical and Energy Engineering

Students in the ME305 course conduct experiments, then do a survey of the outcomes, which resulted in 2 types of data.
Testing to see if the creation of the new equipment improves student learning.
Improvements – ME 350 had a lot of labs and most took it their senior year and the lecture wasn't intertwined. So, we re-designed the curriculum and spread the lab course across multiple semesters to scaffold the learning throughout the program of study.
Academic competitions – trying to intertwine these (beyond the senior capstone).
Taking 3 courses and making them more hands-on. SERI grant to modify things starting in spring 2020 and if it works, change the curriculum for the future.

5. Update and Discussion on HLC #1—Stephen Hundley, Senior Advisor to the Chancellor & Susan Kahn, Planning & Institutional Improvement (15 minutes)

S. Hundley – we are accredited by the HLC, which is different than discipline specific accreditations. Gearing up to 2022's assurance argument. The HLC have a criteria they use. The "open pathway" we are in (highly functioning institutions; elite) now has a year 4 assurance audit. Does not involve a site visit. The HLC asked us to do it a year later and Susan Kahn led the group on that. As we get closer to the 2022 deadline, we will expand these conversations beyond PRAC.

S. Kahn – There is another component of the 10 year review beyond the assurance argument, which is called the Quality Enhancement Initiative. Proposal due in the Spring and will likely include implementation of the new IUPUI+ outcomes.
Criteria keep changing (2002, 2012, 2017) – now I see incremental changes in 2017 and will be interested in looking at them in 2022. Increased pressure to demonstrate students are learning basic skills, which comes from the growth in the for-profit educ industry. HLC (and others) were expected to put in better checks and balances. That decisions are made that align with the mission of the institution, not shareholders or owners. 2012 criteria 1 was mission and integrity, which are now separated.

Mission has evolved – (see notes).

What kind of artifacts that demonstrate these? – In our case, this is more simple because we aren't for-profit. The current Secretary of Education is trying to loosen some of these because they were enacted during the Obama administration.

S. Hundley- Assurance argument vs self-study. We relied heavily on institutional artifacts (PRAC reports, IUPUI Strategic Plan, State of Campus address).
Keep in mind that HLC cares about teaching and learning, but the process looks at the entire institution as an enterprise (e.g., budget, governance).

5. Announcements (2 minutes)

- PRAC Report Due Date and Submission Process: Reports to be emailed to Linda Durr (ldurr@iupui.edu) and Susan Kahn (skahn@iupui.edu) by October 31

- Reminder – no November meeting and instead, PRAC member attend one of the workshops and invite one of your colleagues. Dates are as follows with more details to follow: 1st, 7th, 19th, 27th.

Adjourn

Future PRAC Meeting Dates:

Thursday, December 13, 2018	University Hall 1006
Thursday, January 17, 2019	University Hall 1006
Thursday, February 21, 2019	University Hall 1006
Thursday, March 21, 2019	University Hall 1006
Thursday, April 11. 2019	University Hall 1006
Thursday, May 9, 2019	University Hall 1006

A GUIDE FOR THE PERPLEXED

By David Eubanks

On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" ... I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.
—Charles Babbage

The difficulty in using assessment results to improve academic programs is a recurring theme at assessment conferences. This topic also puzzled me greatly for years. I read books and attended conference sessions that described how to define learning outcomes, create rubrics, map a curriculum, and so on. The theory was beautifully simple: set a goal, measure the goal, then use the data to make adjustments. It was that last part where everything seemed to fall apart.

If you have wondered why it is so difficult to use the outputs of standard assessment practices into a credible understanding of learning, you may have at times felt—like I have—that you were simply not following the script closely enough. *Surely all those other people are making it work. Why can't I?* The intent of this "guide for the perplexed" (with apologies to Maimonides) is to show why it is not your fault.

My conclusion, after seeing hundreds of real assessment reports from many different institutions, supervising assessment programs at four institutions myself, and talking to many other assessment practitioners, is that *it is difficult to use assessment results because the methods of gathering and analyzing data are very poor*.

Academic assessment is like any other data-driven enterprise: effectiveness stems from careful attention to the details and bearing in mind the prime directive of any research activity. Richard Feynman said it best in his 1974 commencement address: "The first principle is that you must not fool yourself—and you are the easiest person to fool."

The Rules of Assessment

The functions of university-wide assessment programs are driven by regulatory requirements as described by regional accreditors. I will use Middles States standards as a template for regulator-prescribed practice, because they are particularly detailed. The requirements below are condensed versions of the criteria in Standard V, which you can find (in video form) here. Quotes are from the video narration.

1. Inter-related goals between programs and institutions, with emphasis on assessing programs and institution (e.g. not individual courses). Documentation must link the mission to outcomes.
2. Assessment at institution and program level "should be of such quality that they meaningfully evaluate the extent of student achievement. […] Assessment processes should enable faculty and other qualified professionals to identify strengths and weaknesses with regard to the student learning outcomes […]. Assessments used should be defensible, meaning that they involve direct observation of the knowledge, skills, and habits of mind or values that students are expected to achieve. […] In summary, the Commission expects accredited institutions to demonstrate than an organized and systematic assessment has prompted meaningful and useful discussions about strengths and weaknesses with regard to student learning outcomes […]"
3. The assessment results must be used for the improvement of educational effectiveness.

This list describes a research program that depends on good data and well-reasoned analysis. In the second item we find the requirement that assessment data should "meaningfully evaluate the extent of student achievement." This is a measurement task, in other words. The research program is not allowed to fail, for example by yielding inconclusive results, but must be used to guide decision-making.

Trudy Banta and Charlie Blaich described the corresponding reasoning behind the assessment movement in "Closing the Assessment Loop" (2011).

> An internally driven, formative approach to assessment is based on the belief that a key factor inhibiting improvements in student learning or allowing students to graduate without learning enough is that faculty and staff who deal with students lack high-quality information about the experiences and conditions that help students learn. If they had information about how much their students were or were not learning and the practices and conditions that helped them learn, practitioners would put this knowledge to work, and improvement would naturally follow. (pg. 27)

The belief is that the barrier to improving programs is a lack of good information. In this light, the Middle States requirement to measure learning is logical: it is intended to provide this essential ingredient.

However, if the data *do not* "meaningfully evaluate the extent of student achievement," then the requirement is Kafkaesque, requiring institutions to legitimize the use of bad data, and punishing them when they cannot.

Fortunately, there is a mature body of work on how not to fool oneself with educational measurement (Brennen, 2006). Unfortunately, that accumulated knowledge can almost never be applied because of the large number of concurrent assessment projects and consequent lack of attention each can get.

Explaining Failure

Program assessment requirements like the ones quoted above apply to most institutions of higher education in the United States: thousands of institutions and their respective academic programs, each with a handful of outcomes to be assessed, which must number in the hundreds of thousands when taken together. This program has been in place for more than a decade, comprising a huge number of mini-research projects. It is reasonable to expect that if this program of data gathering and analysis were successful that it would have produced a great volume of useful findings about pedagogy, curriculum, and student development.

In the same article quoted above, Banta & Blaich (2011) looked for such examples of successful assessment efforts. What they found surprised them.

> 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. (pg. 22)

As Fulcher, Good, Coleman & Smith (2014) point out, the 6% of submissions that Banta & Blaich found to identify improvements is bound to be an overestimate of the actual case, since these submissions were chosen presumably on their merits, and not at random.

There are two possible conclusions. One is that the faculty are generating good data, but are not using it. This way of thinking extends the diminishment of faculty expertise that began with telling them that grades do not measure learning: we've replaced grades with something better—assessments that *do measure learning*—but they still are not producing the intended results. Therefore (the argument goes) we just need

to work more on our processes, so that when the faculty finally do fully adopt these changes a fountain of good educational research will spring forth.

There is another possible conclusion from the Banta & Blaich article, one that is confirmed by my decade of experience: it is not that the faculty are not trying, but the data and methods in general use are very poor at measuring learning.

Common Sense Isn't Enough

In 1989, Patrick Terenzini published "Assessment with Open Eyes: Pitfalls in Studying Student Outcomes" in the *Journal of Higher Education*, an article that anticipates our current situation. On measurement Terenzini wrote:

> […] though locally developed measures may be more carefully tailored to local purposes and educational objectives [than standardized instruments], they are also likely to be untested (at least in the short run) and, consequently, of unknown reliability and validity. […] Many faculty members will have neither the time, commitment, nor competence to develop local measures. (pg. 657)

This problem is exacerbated because "The certainty implied by statistical testing can mask problems that may lead to the serious misinterpretation of results." A specific problem is that

> Common sense suggests that if one wishes to know whether something changes over time, one should measure it at Time 1 and again at Time 2. The difference between the pre- and post-test scores, the "change" score, presumably reflects the effects of some process. […] In this instance, however, common sense may harm more than help. (pg. 660)

He describes problems with the common-sense approach of naively comparing one set of numbers to another, including unreliability of difference scores, ceiling effects, and regression to the mean. See Bonate (2000) for a book-length treatment on how to analyze pre/post scores. The common-sense oversimplification of measurement is a general symptom, I believe, of the trade-off that favors breadth of assessment efforts over meaningful depth.

For readers looking for an excuse, Terenzini provides one, writing that "Methodological standards for research publishable in scholarly and professional journals can probably be relaxed in the interests of institutional utility and advancement," concluding that "Although the methodological issues reviewed here cannot and should not be ignored, neither should one's concern about them stifle action."

This sentiment may seem palliative, as in: *of course, we can't hold ourselves to a high standard of intellectual rigor. We don't have the resources for that.* The whole assessment process would fall apart if we had to test for reliability and validity and carefully model interactions before making conclusions about cause and effect.

How would we feel if the airline industry took that approach to building, flying, and maintaining aircraft? Should we also revert to a pre-scientific era of medical research because randomized trials are difficult and expensive? Are student outcomes valued so much less than health and safety that we should abandon all but the pretense of rigor for the majority of our work?

The disregard for measurement quality combined with the perils of common-sense inference create problems for innumerable assessment projects. There is a sense in the assessment community that once all the proper processes have been followed, then the data produced are inherently meaningful. As such, all manner of comparisons within the data are called forth to illustrate possible uses. One outcome may have

larger averages (or different distributions) than another, or vary from one year to the next, and meaning is read into these differences with only common sense as a guide. Because of the limitations in time and expertise, measurement and statistical considerations are waived. It becomes a pernicious enthymeme: *we used the proper process, ergo the results are guaranteed to be meaningful and amenable to common sense understanding.* Under these conditions the use of data is akin to a Rorschach test.

Using and Misusing Data

The ability to infer meaning from data requires good data and good models of inference. It may be helpful to illustrate this with a real example.

The common-sense method of using assessment data goes like this: (1) find a number that looks too low in assessment results, and (2) imagine some change that might raise the number. Popular changes include adding a new subject to the syllabus, changing reading or assignments to emphasize some aspect of learning, changing a textbook, or sometimes adding a new course to the curriculum. What is being *replaced* in the process is rarely addressed.

The example below shows real data from assessing basic foreign language proficiency over two academic years. Course instructors rate student performance in the language courses for general education using a three-point scale (does not meet expectations, meets expectations, or exceeds expectations), using a rubric established by the faculty when the general education curriculum was created at my institution.

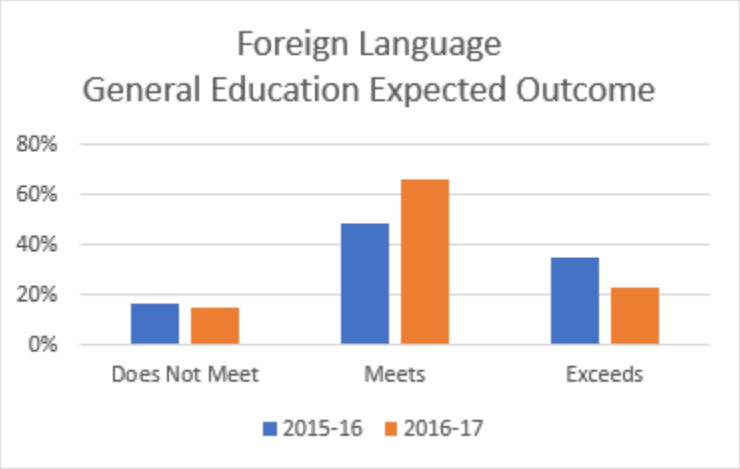

Figure 1. Summary Graphs of Language Proficiency

We can see that about 16% of students are not meeting the expectation. The graph gives us some information, but not much. The requirement to use these results to make improvements typically leads to conclusions such as the following:

> Over the last two years, 16% of students have not met expectations in the general education language proficiency. An analysis of end-of-term papers shows that lack of proficiency is related to problems with basic vocabulary and grammar. Consequently, we will spend an extra week at the beginning of the term reviewing this material.

On the surface, this "solution" addresses the problem, but it has the hidden cost of reducing content for 84% of the students who do not need the review. Worse, it has been known for year that the common-sense solution of remediation can backfire (Hillocks, 1986).

We now take another look at the same data with a more inclusive model of cause and effect. One useful approach is the Astin model (1991), which can be used to categorize interactions as shown in Figure 2.

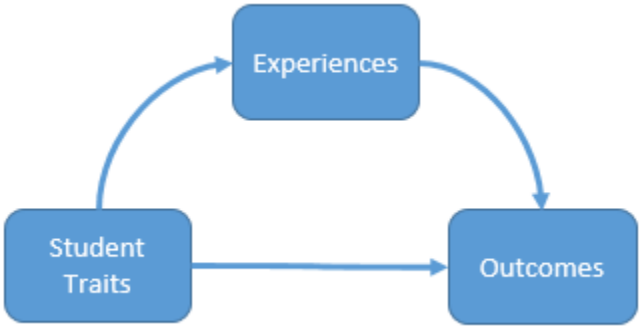

Figure 2. Astin's Input-Environment-Output model

In this case we have:

- **Traits:** Students arrive with varying degrees of academic preparation, which can be partially assessed via their high school transcripts. In particular, a recalculated high school grade average (HSGPA) predicts college GPA, and can be considered a measure of a student's academic preparation, talent, and work habits.
- **Experiences:** Students may wait zero, one, two, or three years before enrolling in the foreign language courses required by the general education requirement. These constitute different learning experiences, since they will begin to forget what they learned in high school.
- **Outcome:** The same ratings found in Figure 1.

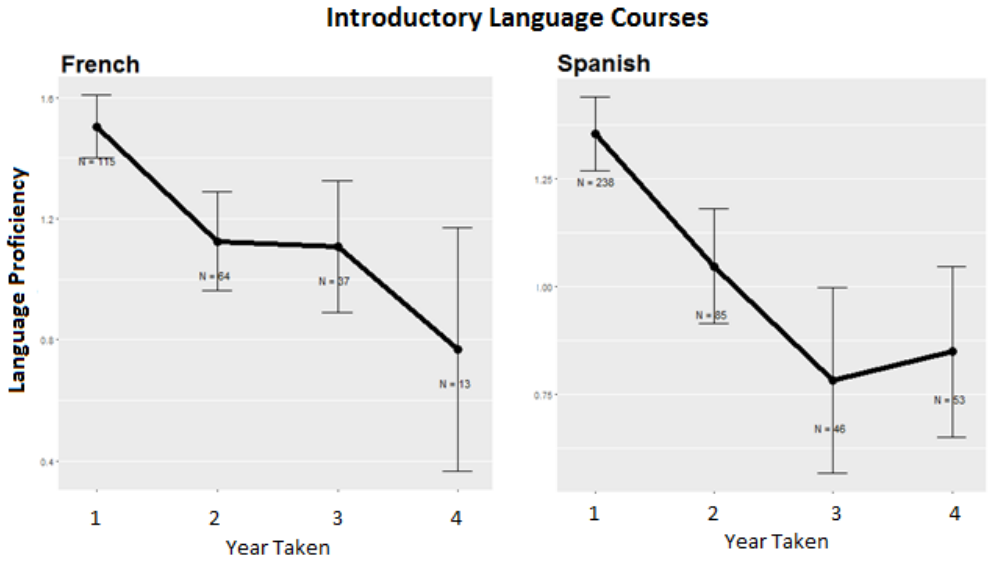

Figure 3. Language Proficiency by Year the First Course Was Taken

The graphs in Figure 3 relate the experience (year taken) to the outcome (rubric rating) and show that for both French and Spanish, students who wait even one year show significantly decreased outcomes on average. This suggests two possibilities. Maybe students who wait to take language courses simply forget what they learned in high school, and their learning suffers. Or it could be that students who are weaker academically avoid the course as long as they can, and it doesn't matter when they take it. To resolve this, we introduced the student characteristic variable (HSGPA), and a regression analysis finds that *both* the wait *and* HSGPA contribute to the decline in scores. Discussions with the language faculty confirmed that this finding was reasonable. The solution is different advising, to prevent students from waiting to take the required language course.

The language program faculty take assessment very seriously, and several of them are certified as instructors for teaching the rating system adopted by the American Council of the Teaching of Foreign Languages (ACTFL). For many years they have been pulling a selection of student essays and rating them according to the ACTFL rubrics. They had gotten accustomed to the 16% unsatisfactory rate, and eventually just assumed that this was the best the program could do. In fact, there was nothing in those essays that would have told them what the actual problem was. They were not surprised by the graphs in Figure 3, and even pointed me to published research that confirmed the finding. With the language faculty's support, the graphs in Figure 3 are powerful in communicating to advisors the danger of letting students wait to take these courses.

The effect in Figure 3 is also detectable using course grades, and a scan of all the 100- and 200-level courses taught at my institution identified other introductory courses (especially mathematics), where it is detrimental for a student to wait a year.

The common-sense use of assessment data illustrated in the discussion of Figure 1, is not complex enough to account for real educational processes, *even when the assessment data are meaningful.*

Unfortunately, data that result from usual academic program assessment activities are inadequate to use a model like the schematic in Figure 2, even if someone has the time to do it. The mandate to use the results leads to a random shuffling of educational practices, or *post-hoc* justification of a change that is desired for other reasons.

Data Problems

Because assessment data must be mass-produced, we typically create dozens or hundreds of shallow pools of data, with small decontextualized samples. There is no time to diagnose, let alone fix, the data problems. This creates insurmountable problems for analysis.

Samples of student work or observation are small (e.g. <100), making it likely that even if measurements are good, we will still get the wrong answer to many of our questions. Small samples also make it impossible to assess reliability and validity. The graphs in Figure 3 are based on hundreds of observations; with small sample sizes (e.g. 30) it would not be possible to detect the effect shown there.

The data are decontextualized, for example by not considering student characteristics. Omitting context leaves out the most powerful means of discovering cause and effect, as in the example above with foreign language proficiency. It is also essential for assessing change. See (Ewell, 1991), "student learning and development is a complex, multifaceted phenomenon unusually resistant to single-factor explanation" (pg. 95). As Ewell notes, this leads to longitudinal studies. See Singer & Willett (2003) for a comprehensive

statistical treatment on using longitudinal data to estimate change. See Kilgo, Sheets, & Pascarella (2015) for an example assessing high impact practices.

Recall that the mandated objective of assessment work is to place "emphasis on assessing programs and institution (e.g. not individual courses)." But "a program" is almost certainly a different experience for every student (refer back the Astin model in Figure 2), with different instructors, different courses in a different sequence, different starting preparations, and many other important variables that can affect "program outcomes." Additionally, each student is different, with unique academic strengths and interests, and so on. Since none of these variables is usually accounted for, only large effects could possibly be detected, and even then we may fool ourselves as to the cause. This is a hopeless situation given the state of the actual data and inferential methods used. *If an effect is large enough to show up under these conditions, the faculty almost certainly already know about it from their experiences with students.*

Even in cases where a curriculum is highly structured (e.g. cohort-based with a fixed course sequence), it is necessary to take into account student traits when trying to understand the cumulative effect of the curriculum.

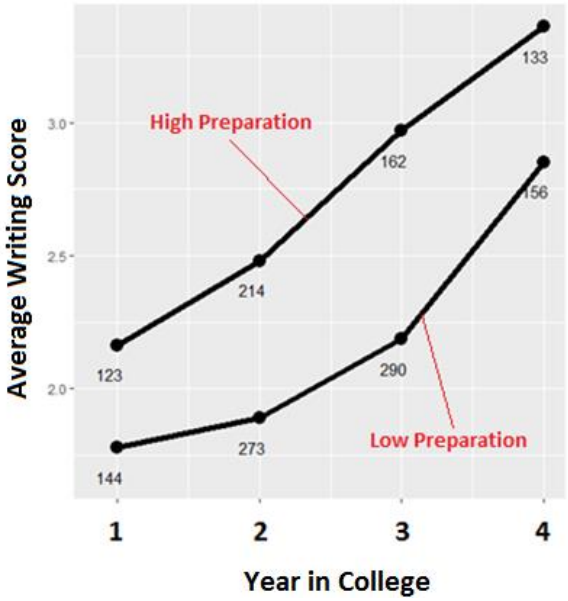

Figure 4. Writing Scores by Academic Preparation, with Numbers of Students

The developmental paths in Figure 4 show two different average trajectories for writing scores at my institution (based on two years of data). The top line shows students who were in the upper half of the standardized high school grade average for their college entering class. The bottom line is the lower half, by high school grades. It is well known that high school grades predict college grades reasonably well, so it is not surprising that writing ratings would show a similar effect. Notice that the lower group appears to lag the upper group by a year or two (survivorship bias is strongest in the lower group, which slightly inflates those scores).

The point of this illustration is that if an analysis does not incorporate levels of student preparation, even if the data are good and the program completely standardized, the results may be driven by varying student qualities and not program effects. Imagine a program so terrible that only the most determined and talented students can survive it. The assessment results will glow with the accomplishments of these talented few,

and since there are no results for all the ones who dropped out immediately, they are invisible to assessments. In other cases, a change in assessment measures may be attributed to program characteristics rather than changing student traits. It is important to understand the difference when trying to make improvements (e.g. more rigorous curriculum versus more tutoring versus better advising).

Language as Camouflage

So why is it not generally accepted that poor data and common-sense inference invalidate the majority of assessment projects? On the contrary, judging from the rhetoric within the assessment community and from accreditors, there is great confidence in the processes that are in place.

Resolving that paradox requires taking a closer look at where confidence is placed, namely in the language and processes of assessment: its bureaucracy. There are rubrics to rate the language/process "correctness" of assessment programs: a list of checkboxes with things like [X] Defined at least three outcomes, [X] Outcomes are measurable, [X] Outcomes relate directly to the next higher level of outcomes at the institution, [X] Outcomes mapped to the curriculum, and so on. There are articles, books, and lectures on how to write outcomes statements for courses, programs, and institutions, how to create curriculum maps, how to create rubrics, and how to organize and evaluate all of this work for each academic program.

The emphasis on form over function extends to the reviews we do of each other's programs in accreditation work. *Did the program have outcomes? Were they assessed? Were the results used for something?* Everything is checked except whether or not the data are any good and the inferences are reasonably justified.

Most institutions probably have a small number of assessment projects, perhaps in general education, that do get the attention they need to be successful as educational research. But the majority can only pass accreditation reviews through attention blindness induced by a box-checking mentality of correctness.

The Future

In the era of "fake news," it is imperative that higher education holds itself to a high standard of intellectual honesty. We should follow the lead of academic psychology in a self-examination of our standards of practice. That field is enduring a "reproduction crisis" that calls into question a large amount of peer-reviewed, published research. Relying on small sample sizes is one of the causes (Simmons, Nelson & Simonsohn, 2011).

Imagine if each town and village were required to research and produce its own drugs, and ignore large-scale science-based medical research. That is our current situation with respect to assessment.

By contrast, research in teaching and learning is booming. Look at the proceedings of *Educational Data Mining* for many examples of the creativity and energy being devoted to this research. There is more data available than ever before, computation is cheap, and new methods for visualization and analysis abound.

We can imagine a future where assessment leaders work closely with institutional researchers and scholars to create and share large sets of high-quality data. These might be organized by discipline or at the institutional level to focus on a manageable number of outcomes—not hundreds of them at once. We would work with faculty members to understand and use research findings instead of cajoling them to do paperwork, re-grade papers, and then stare at bar graphs trying to divine meaning. Assessment conferences can be about what we discovered and how faculty are using that information.

One model of that approach is the English composition program at University of South Florida-Tampa, where my colleague Joe Moxley has turned a two-semester writing requirement into a large-scale research program. His work has garnered grant money, attracted dozens of researchers, launched a journal and a conference, and produced a corpus of hundreds of thousands of student papers, peer reviews, rubric ratings, and survey items that is available to researchers. What makes Moxley's program so outstanding is the constant critical attention to the quality of data. The goal is not perfect measurement; the goal is to not fool ourselves.

Other models are possible that would fit different situations and types of institutions. I am continually impressed with intelligence and dedication of people I meet in the assessment field. It is appalling how much of that talent gets wasted filling out checkboxes. If existing assessment resources were redirected to trying to understand student learning, we could revolutionize education in the next ten years.

The assessment profession is now decades old, and it is time that the standards of practice are defined by the community of practitioners who do the job. In collaboration with other stakeholders, the Association for the Assessment of Learning in Higher Education (AALHE) is the logical choice of a body to lead the creation of such standards.

Acknowledgements

I am indebted to Jeff Barbee, Gray Scott, Jane Marie Souza, and Josie Welsh for their excellent advice and feedback while writing this article.

References

Astin, A.W. (1991). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Washington, DC: American Council on Education/Oryx Press Series on Higher Education.

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning*, *43*(1), 22-27.

Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. CRC Press.

Ewell, P. T. (1991). Chapter 3: To capture the ineffable: New forms of assessment in higher education. *Review of research in education*, *17*(1), 75-125.

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014, December). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. (Occasional Paper No. 23). Urbana, Il: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Hillocks, G. (1986). Research on written composition. *Urbana, IL: National Council of Teachers of English.*

Kilgo, C. A., Sheets, J. K. E., & Pascarella, E. T. (2015). The link between high-impact practices and student learning: Some longitudinal evidence. *Higher Education*, 69(4), 509-525.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Terenzini, P. T. (1989). Assessment with open eyes: Pitfalls in studying student outcomes. *The Journal of Higher Education*, *60*(6), 644-664.

David Eubanks is Assistant Vice President for Institutional Effectiveness at Furman University and a member of the AALHE board. He can be reached at david.eubanks@furman.edu.