Enhancing the Rigor of Common Library Assessment Activities

Craig E. Smith

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries.
Please contact epubs@purdue.edu for additional information.

Enhancing the Rigor of Common Library Assessment Activities

by **Craig E. Smith** (Assessment Specialist and Senior Associate Librarian, University of Michigan Library)
<craigsm@umich.edu>

The University of Michigan (Ann Arbor) Library recently hired me as an assessment specialist, which is a new role in our library system. My background is in psychology and institutional research, rather than librarianship. Many assessment activities were established in our library when I started my job, and my experiences at the **2018 Library Assessment Conference** and the **2019 ACRL Conference** further underscored for me the fact that libraries have embraced assessment. With increasing exposure to library assessment, however, I have observed that embrace of assessment and rigor in assessment are sometimes divorced. This is not universally the case, and it is also wholly understandable. In many cases, those doing library assessment do not have assessment/research backgrounds and are also balancing assessment with other duties.

With this in mind, my goal is to reflect on some basic ways that we can increase the rigor of our assessment work. There are many areas of library assessment. I will, by necessity, limit the scope of this article to assessment involving human data, and to a small number of ways of increasing rigor that are not overly burdensome. My goal in this article is not to be harsh; even seasoned assessment practitioners and researchers make mistakes and continue to learn. Instead, this is an observation of places where relatively easy opportunities for improvement exist.

Perspective Taking

We often ask people to share their library-related experiences and needs. Yet designing surveys, focus groups, tasks, and interviews that yield clear, usable data is genuinely challenging. One common pitfall is a cognitive bias called *the curse of knowledge*: the assumption that others share our knowledge as we make statements and ask questions. For example, imagine this survey question, “How often do you visit the library’s website?” We want information about people’s interactions with the library’s homepage, yet most respondents only think about the online catalog. Or imagine that we conduct intercept interviews to ask how people access and experience library consultation. Many participants indicate that they have never used library consultation services; some of them have, but do not recognize those experiences as forms of consultation.

A basic but important solution to such problems is to vet questions used in all methodologies thoroughly prior to launch. We can first ask a diverse group of colleagues to review questions and prompts with the request that they look for assumptions, biases, and wording issues. A second step should involve piloting with diverse volunteers from the target population. Piloting clarifies pragmatic concerns, such as accessibility and participation time.

Importantly, certain forms of piloting allow us to engage in perspective taking. Imagine a survey or interview that includes the question: “How often do you use the library?” Using a technique called cognitive interviewing allows us to understand how people interpret this question. Cognitive interview questions are things like:

- *Can you tell me in your own words what you think we are trying to ask?*
- *What does “the library” mean to you?*
- *What does “use the library” mean to you?*
- *[If a survey] Do the response options we provide for this question allow you to answer in the way that you would like to?*

We may find that people picture different things (e.g., different buildings) when they think about the library, and have different views of library use (e.g., visiting the café, accessing physical collections). After interviewing a small but diverse group of people, the original question will be more precise, or may become a series of questions, perhaps accompanied by simple framing information. Final vetting of revised questions may lead to additional adjustments, or may affirm that we are ready to launch an effort that will yield data that are interpretable and actionable.

Asking Questions with the Right Focus

There are other ways that problematic question framing can occur. Imagine we want to know whether students who received two workshops, compared to those who received one, better remember how to use Boolean operators. One week later we use a follow-up survey: “Demonstrate how you would use Boolean operators to construct an effective search that addresses: ‘Do violent video games increase aggression in children?’” We find that the two-session instruction approach is associated with better performance. Perhaps the two-session approach is indeed better, but there is a potential confound. The two-session students heard “Boolean operator” far more often; perhaps they were more likely to remember the term and were better positioned to demonstrate knowledge possessed equally by both groups.

Do we care if students need to remember the term Boolean operator in order to demonstrate learning? If not, then our questions should focus on our precise interests. For example, we could ask: “What is an effective way to structure a search for academic literature on the following question: ‘Do violent video games increase aggression in children?’” If there was still a difference between the two

groups — and other potential confounds were also controlled, e.g., via random assignment — then confidence about group differences may be justified, as would using the data to guide future instruction approaches.

In the section on perspective taking above, the focus was on understanding others’ thinking. The takeaway here is that we also need to be reflective and clear about *our own thinking* and goals as we design assessments.

Avoiding Common Problems with Questions

When crafting questions we sometimes make presumptions that can yield inaccurate data. As an example: “When you access our library’s books, do you prefer digital books (eBooks) or physical books?” This question presupposes that a person seeks books and possesses a preference. Someone may provide an inaccurate response because doing so is easier than doing something else. Or perhaps because admitting to not using our library’s books seems socially undesirable. In an interview, for example, a better way to start might be, “We have digital and physical books in this library’s collection. Many people use this library’s books and many others do not. Which is more like you?” If a person reports library usage, a question about preference that includes a no-preference option becomes appropriate. The new approach could also lead the investigator to ask about other ways a participant might access books, and could still lead to a preference question. (Yet beware of other potential assumptions! For example, in some libraries, people logged in on a certain network may be unaware that they have seamlessly accessed a library’s digital collection.)

Another common problem is the “double-barreled” question that asks two different things. Consider a performance-related survey item about a supervisor: “My supervisor encourages and enables collaboration with other work groups.” It is possible that a supervisor could be enthusiastic about collaboration, yet ineffective at enabling it. With such questions, some respondents may provide bad data or simply skip the question. Survey, focus group, and interview questions should be carefully reviewed for common problems such as this. In the example above, the item should be split into two questions if both “encourages” and “enables” are of interest.

It can also be problematic to ask the same types of questions about very different things. For example, in creating a toolkit to measure the impact of spaces, events, instruction, and consultation, questions about a construct such as confidence are not easily asked about — or always relevant to — each type of experience. Yet some libraries use matched sets of items

continued on page 15

about such things. For example: “I feel more confident about my ability to conduct my research” about research consultation, and the companion “Using this space makes me feel more confident about my ability to achieve my goals” about a library space. The first question could yield valuable information about consultation impact, but the latter question will yield data that are nearly impossible to interpret; what about the space did or did not increase confidence, and what were the goals? The examples here are from ACRL’s Project Outcome measures.¹ These measures are described as “designed and tested specifically to be reliable measures of perceived impact.” Well beyond this specific set of measures, it is important to note that measures can perform well in terms of reliability (e.g., internal consistency reliability, test-retest reliability) without being valid or informative in critical ways. There are many good online summaries of the types of reliability and validity one should consider when engaging in various types of measurement.

A final example of a common pitfall in this realm is the uncritical use of existing questions, protocols, coding schemes, and instruments. Simply because a method has been used in the past, or has been published or presented, does not mean that it is the right fit for a new project without adaptation, or that it is of high quality. Both new and existing methods should be viewed with a critical stance. It is too often the case in many disciplines that methods, because they are published or presented, become imbued with a gleam that may or may not be deserved.

There are many useful guides on asking good questions and making effective use of the answers, across a wide variety of methodologies. A small number of examples are listed in the Appendix.

Getting Meaningful Responses

Part of asking good questions involves giving people meaningful ways to respond. Imagine that a faculty member in Biology is asked: “What is your level of satisfaction with the support you receive from your library subject liaison?” There is a 10-point scale and only the poles are anchored (e.g., 1 = low; 10 = high). The faculty member is not sure what kind of support the question is asking about, but she picks a 7 to report feeling mostly satisfied. Her departmental colleague also chooses a 7, but uses it to express slight satisfaction. How can we avoid such problems?

There is an extensive literature on best practices regarding eliciting meaningful responses to questions; a few examples are covered here. First, if you are unsure how to structure a question with response options that include a relatively full range of reasonable possibilities, consider (a) using a non-leading, open-ended question, or (b) using a mixed-methods design that allows you to develop good questions and/or response options by first interviewing and better understanding members of the relevant population. Open-ended questions may require

careful decisions about coding and proper checks on coding reliability. But such questions can yield data that are more meaningful compared to questions with poorly-conceived response options. Second, when using rating scales, mitigate common response biases (e.g., acquiescence bias, response set bias) by using both positively- and negatively-worded items to capture a single construct when possible. Third, when using scales, confirm that multiple participants understand each point on the scale and interpret them in a similar way (e.g., via cognitive interviewing), and try to use scales connected to the construct of interest (e.g., using agreement scales for all questions is not always the best approach).

See the Appendix for guides that address these issues and more.

Seeking Diverse Informants

Assessment participants help guide our work and decisions, yet they do not always represent the diversity of the communities we serve. How well can we understand ourselves and the people we hope to serve if we get marginal input from groups that may already be marginalized? The challenge of achieving true representation in assessment projects is not unique to library assessment. But how can we do better?

If you work at an academic library, consider partnering with stewards of administrative data in your library or on your campus. An office with administrative data access may be willing to support assessment efforts by providing representative lists of campus community members (representative in terms of gender identity, race/ethnicity, first-gen status, area of study, etc.). These lists can be used when recruiting for all sorts of assessment activities. Such a partnership may necessitate training in managing sensitive data and in responsible recruiting practices. For both campus and public libraries, forming mutually-beneficial relationships with organizations that include diverse members of your communities is also an important avenue to increasing the diversity of willing assessment participants. It is also useful to remember that members of campus groups that are small when intersectionality is considered (e.g., female full professors of color in STEM fields) are often disproportionately asked to provide service; assessment participation is a form of service. It is important to ask for help respectfully, to think about the protection of potentially-identifiable data, and to allow people to opt out of future requests.

Achieving genuine diversity in assessment is challenging, but the consequences of falling short are problematic. For example, when there are very small numbers of certain groups in survey samples, a common strategy for ensuring anonymity is to exclude these groups — and the insights they offer — from the results altogether. Another common approach is to lump all responses together, meaning that we may miss important ways that groups might differ. As we commit ourselves to diversity, equity, inclusion, and accessibility, we must also commit to engaging in assessment practices that reflect these values. There are real obstacles that impede success in this area. When

this is the case, we can at least be careful with the claims we make by not generalizing our results to groups who are not well represented in our data.

Being Careful with Claims

We often make claims based on assessment data. For example, someone assessing student-library interactions might make a claim about impact on student retention or GPA. As another example, someone might use data to claim that one group of patrons experiences the library as more welcoming than another group. Sometimes, however, such claims are made without the support of proper study design and/or analyses.

If you do not have people in your library with expertise in study design and data analysis (qualitative and quantitative), consider forming partnerships with people who do. This is easier for academic libraries, and for libraries near colleges and universities. But there are communities of assessment practitioners online who can offer guidance (e.g., the ASSESS email list²). We can also consult with peers in other libraries when we have questions about methods, analyses, and reporting.

Further, I encourage people conducting assessment in libraries to think about some of the following common issues before making claims based on data.

First, correlational data cannot easily support claims about causality. For example, imagine we find that frequency of collections use is positively correlated with GPA. First, even the simple statistic should be subjected to some scrutiny; how large is the association and is it statistically significant? More importantly, the association should be viewed as open to multiple interpretations. It could be, for example, that unmeasured variables (e.g., motivation, self-efficacy) account for both collections use and GPA. To make strong claims about impact, our studies must be designed correctly. Although it seems obvious that correlational data alone cannot support claims about impact or causality, it is common to see presentations of data in which such claims are made subtly or explicitly.

Second, if our goal is to generalize from a sample to make claims about a population, our claims should be supported by both good study design and inferential statistics. It is not uncommon to see presentations of library data in which simple descriptive statistics from a sample (e.g., means, percentages) are used to implicitly make claims about a population (e.g., a campus community). As a hypothetical example, imagine that an assessment with 50 undergraduates finds that 56% of first-year students ($n = 25$) cite social media sources in course assignments, while 44% of second-year students ($n = 25$) report doing so. It would be a mistake to claim that this is a difference between the groups. A more reasonable next step would be to use a statistical analysis — in this simplified case, a chi-square test — to determine whether a generalization is warranted. This test would reveal that the difference of 56% vs. 44% is not large enough, given the sample size, to confidently claim any-

continued on page 16

thing about the larger population of younger undergraduates ($\chi^2 = .72$, $p = .40$, $\phi = .12^3$). Note that if the difference was very large in the hypothetical sample (e.g., 88% vs. 12%), the level of confidence about generalizing to the population would be improved (if the sample was representative), and this should be demonstrated via the use of proper statistics ($\chi^2 = 28.88$, $p < .001$, $\phi = .76$). Or, if the representative sample contained 400 students and the difference was 56% vs. 44%, the level of confidence would also be improved ($\chi^2 = 5.76$, $p = .02$, $\phi = .12$). These last points also speak to proper study design. In this example, with quantitative data, we see that if one wants to investigate a potential difference or effect that is likely to be rather modest though potentially meaningful, planning for the proper sample size is critical. Further, thinking about sample composition is critical; as noted, generalizing to a diverse population from a homogeneous sample is problematic.

Third, effect sizes matter. Consider the hypothetical study with 400 students described above. In that case, one could claim that the result was statistically significant. A claim about practical significance, however, could be scrutinized; an effect size of .12 indicates a very small difference. As more libraries move toward using analytics with large samples, there will be many cases in which statistically significant findings will be obtained. The critical question in such cases is whether the findings convey practical significance. This question can be assessed, in part, by attending to effect sizes when reporting on quantitative results. There are, of course, other problems with common library analytics methods, such as putting too much stock in correlational data that lack the proper controls.

Using Assessment Strategically

We can also fall short if we fail to consider assessment at the outset of a project or endeavor. For example, imagine we conclude that a workshop on website design was successful because participants reported high levels of confidence and self-efficacy in a post-session evaluation. Later, a colleague asks whether we have any way of knowing whether the workshop led to *changes* in confidence and self-efficacy. If we had planned carefully, we could have used pre- and post-session assessments, or at least crafted well-designed post-session questions about *changes* in the constructs of interest. The stakes get even higher as we assess major projects or initiatives without considering assessment as part of the larger planning process. Another advantage of considering assessment at the outset of a project is that we can think across the silos that often exist in our organizations. For example, if the goal is to create an assessment of a website design workshop, there may be real benefits to working on such a project in collaboration with people in your organization who teach or sponsor other types

of workshops (e.g., the benefit of exposure to new instruction and assessment strategies).

When thinking about using assessment strategically, another important question is whether we always need formal assessment to gain insight or inform decisions. Assessment consumes time and resources, so a good question to ask is whether your library has a strategy for how assessment is deployed. There is an emphasis on creating a “culture of assessment” in many libraries. It might be more useful to create a “culture of strategic planning” in which decisions about when and how to use assessment become a standard part of larger conversations about making improvements, starting projects, meeting the needs of users, etc. For example, your library could create a checklist of questions that get asked in the context of new endeavors. One question could be, do we need to use assessment here, or do we need an assessment plan? When might the answer be no? Perhaps if the expertise in the room and the library’s strategic goals give you enough insight and direction to make a decision without collecting new data. Or perhaps you already have access to data that will, if used correctly, illuminate a path forward. A decision to forgo assessment should be made carefully and with people in the room who are willing to ask challenging questions, but such a decision is not always wrong.

Seeking and Providing Critical Feedback

In presentations of data at recent library conferences, I have observed that audience members often provided positive feedback about the studies and findings shared by their peers. Almost absent, however, were kindly-worded comments that probed problematic study designs, analyses, and interpretations. Yet some presentations did indeed have shortcomings. Norms of politeness do not need to be sacrificed in order for us to push each other — and expect each other — to do rigorous work.

The lab meeting model exists in many research disciplines. For example, psychology lab meetings are used to get feedback on pilot data, research ideas, study/instrument design, data interpretation, and presentations/manuscripts. Psychology lab meetings are eye-opening experiences for newcomers. The feedback is abundant and is often more aimed at identifying problems than giving compliments. Yet the investigator leaves with important ideas about how to make their work stronger.

This is a model that we can harness as we plan new assessments, or as we prepare to share findings and interpretations. It is concerning to me that audience members at library conferences may walk away from a presentation thinking something is “true” and actionable when the assessment work has not been properly scrutinized and contains design, analysis, or interpretation problems. I encourage those doing assessment in libraries to create communities of practice in which there is safe space for offering supportive critique. If you do not have people in your library who can offer informed

critique, an alternative could be to partner with people on a college/university campus who are willing to share their time, or to collaborate with an online assessment community.

Relatedly, if you are a reviewer for a publication or a conference and are considering a submitted assessment project, set a high bar. Be very kind, but ask critical questions. If statistics should be reported, ask for them; this is relevant for many types of assessment, including many forms of qualitative research. If you are reviewing work where methods cannot support claims, say so. If you are reviewing work you don’t feel qualified to evaluate, admit it. Analyses of assessment and research data (e.g., regression models, mixed-methods designs, interview coding) can be done very well or very poorly, and there should be at least one person familiar with the relevant methods reviewing a piece of work. Setting a high bar does not necessarily involve rejecting flawed assessment; in some cases it may simply involve asking investigators to adjust their claims. For example, you may end up recommending that claims of “success” regarding an intervention lacking proper controls be tempered, with the results instead described as promising and justifying additional investigation. These can be enlightening moments when we think of ways to conduct more solid assessment, thereby building solid guideposts for our library work.

Appendix: Helpful Resources

Bradburn, N. M., Stern, M. J., Johnson, T. P., & Wansink, B. (2020). *Asking questions: The definitive guide to questionnaire design* (3rd ed.). New York, NY: Wiley.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks, CA: Sage.

Fowler, F. J., Jr. (2013). *Survey research methods* (5th ed.). Thousand Oaks, CA: Sage.

Saldaña, J. M. (2015). *The coding manual for qualitative researchers* (3rd ed.). Thousand Oaks, CA: Sage.

Seidman, I. E. (2019). *Interviewing as qualitative research: A guide for researchers in education and the social sciences* (5th ed.). New York, NY: Teachers College Press.

Tracy, S. J. (2019). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact* (2nd ed.). Hoboken, NJ: Wiley-Blackwell.

Endnotes

1. Project Outcome: Measuring the True Impact of Public Libraries. Retrieved June 10, 2019, from <https://acrl.projectoutcome.org>.
2. ASSESS is managed by the University of Kentucky College of Education in collaboration with the Association for the Assessment of Learning in Higher Education.
3. The typical standard in social science research for statistical significance is $p < .05$. The symbol ϕ represents effect size for a chi-square test (effect sizes are discussed briefly in this section).