2016

Using a “Messy” Problem as a Departmental Assessment of Undergraduates' Ability to Think Like Psychologists

Patricia Marten DiBartolo
Smith College, pdibarto@smith.edu

Lauren E. Duncan
Smith College

Minh Ly

Alan N. Rudnitsky

Using a 'Messy' Problem as a Departmental Assessment of Undergraduates' Ability to Think Like Psychologists

Abstract: This paper presents a case study of faculty in a psychology department whose shared questions about pedagogy and learning informed a data-driven curricular review and revision using an open-ended assessment that privileged deep learning. This paper describes the development of this assessment and how its results across the arc of the major led to a revision of the department's curriculum, including the creation of new courses that focused on developing students' abilities to 'think like psychologists.' The study indicates that faculty intuitions of potential problems in student learning can be successfully assessed and then addressed through curricular changes.

Keywords: assessment; deep learning; psychology; interpretive knowledge

Teachers want their students to learn in ways that are meaningful, useful, and lasting. Terms like deep learning, adaptive problem solving, self-regulated learning, critical thinking, and 21st century skills are all descriptive of the important qualities characterizing this kind of learning (Sawyer, 2014). In academic disciplines, a closely related learning goal is expressed as being able to think like an expert (e.g., mathematician, anthropologist, historian, engineer, biologist). The hallmark of this kind of thinking is its reliance on disciplinary frameworks for understanding so that students can work on the solution of novel, ill-defined problems.

Each discipline defines its foundational courses based on the critical body of knowledge upon which advanced study is based. In psychology, an essential foundation in psychology requires study of the systematic empiricism of the discipline, with a scientific methods course as one essential feature (American Psychological Association, 2013; Dunn, McCarthy, Baker, Halonen, & Hill, 2007). In order to think like psychologists, students have to understand where psychological knowledge comes from, what counts as 'knowing,' and how knowledge is appropriately generalized and applied. Acquiring this core knowledge is the fundamental reason for studying research methods. Teaching in ways that help students progress toward long-range mastery of this material is challenging.

This paper presents a case study of faculty in the psychology department of a small liberal arts institution, whose shared questions about pedagogy and student learning helped to inform a data-driven review and revision of the curriculum. The work reported here grows out of questions of understanding articulated by faculty who teach the research methods course. What we describe are the open stages of design research (Collins, Joseph & Bielaczyc, 2004) in which we used a theoretical framework from the learning sciences to engage in an evidence-based direct assessment of student learning.

Our research emerged from a year-long series of sustained conversations within and beyond the methods instructors' group about how best to teach our students so they would learn deeply within the context of the department's required foundational methods course. In these conversations, faculty instructors of the course articulated a shared concern: students in advanced courses within the major often evidenced limited transfer of the important ideas and skills of research methods, despite evidence that most students generally did well in our required methods

course (cf. *citation deleted for the purposes of blind review*). Faculty agreed that any proposed changes in instruction designed to address this problem would need to be informed by a better understanding about what students were (and were not) learning and retaining after the course ended. Over a year, a working group of methods faculty developed an assessment designed to shed some light on students' ability to think like psychologists. These discussions dovetailed with a set of broader psychology faculty conversations related to a college-mandated departmental review.

Ultimately, and consistent with the intent of design research (Collins et al., 2004), these efforts were focused on actions that the faculty could take to adjust both curriculum and pedagogies to improve student transfer of methods knowledge. In the subsequent sections of this paper, we explain the major ideas about learning and transfer that influenced our thinking about assessment; we describe the assessment task presented to students and explain how it was scored; we present findings on students and their learning; and we conclude by describing the curricular and pedagogical impacts of the findings.

Theoretical Grounding of the Assessment

Being able to think like an expert, in this case a psychologist, is a complex learning goal that is more than a collection of discrete skills (Bransford, Brown, & Cocking, 2000). Undergraduate psychology students, especially after completing a research methods course, seem well able to answer closed-ended questions about (for example) statistical significance, the inferences one can validly draw from different research designs, and the importance of construct operationalization in the interpretation of research results. Thinking like a psychologist involves applying these ideas flexibly where the weight or even applicability of any particular idea is a matter of context. Assessing this kind of student learning needs to take these complexities into account.

Traditional assessments of student learning typically ask students to recall and apply facts and concepts within contexts similar to the classroom. Rarely are testing contexts markedly different from the learning context in order to measure transfer of learning. Common forms of assessment are often close-ended tests (multiple choice, short answer) that are time limited and constrained by format in the type of learning they value. Bransford and colleagues (Bransford & Schwartz, 1999; Schwartz, Bransford, & Sears, 2005), building on the work of Broudy (1977), saw these types of assessments as measuring replicative and applicative uses of knowledge. Replicative knowledge, 'knowing that,' is operationalized on tests as memory-based, recall type questions. Tests of applicative knowledge, 'knowing how,' ask students to use what they know. Students might, for example, be asked to identify examples of a phenomenon or solve problems that call for a known formula in a different context. College students tend to be very strategic learners when preparing for assessments of replicative and applicative knowledge. They may perform well at the time of testing but tend to forget replicative knowledge and not be able to sustain or expand transfer of applicative knowledge to novel situations.

Broudy (1977) described a third type of knowledge, interpretive knowledge or 'knowing with.' Schwartz et al. (2005), explaining the importance of interpretive knowledge, noted that 'for many new situations, people do not have sufficient memories, schemas or procedures to solve a problem, but they do have interpretations that shape how they begin to make sense of the situation.' (p.9) There is considerable evidence that how one defines or frames problems has major effects on subsequent thinking (e.g. Bassok & Holyoak, 1989; Bransford & Stein, 1993; Chi, Feltovich &

Glaser, 1981; De Groot, 1965; Gibson & Gibson, 1995; Greeno, Smith, & Moore, 1993; Marton & Booth, 1997; Schuyler, 2003).

Interpretive knowledge is at the heart of 'thinking like a psychologist' and yet is mostly overlooked by traditional assessments. Interpretive knowledge should allow students who have taken a methods course to move on in the psychology curriculum, approaching new problems with a set of skills that are needed to make sense of messy problems the way a fully trained psychologist would. The students may not be able to label all concepts with precise terminology (that would show replicative knowledge), but should be able to approach the problem knowing what the core issues are, how the problem should be framed, and what information is relevant to devising a solution. We set about determined to assess interpretive knowledge in our students.

Scope of the Assessment

We began asking: what is the best way to measure how well students transfer their knowledge from our introductory research methods course to more advanced and real-world contexts? Thinking like a psychologist means in part that one is a highly skilled consumer of research who can: think critically about behavior, brain, and mental processes; understand the relations among theories, observations, and conclusions; and weigh evidence in evaluating particular theories or approaches. Although most of our majors would not become research psychologists, we wanted them all to have the ability to interpret data and studies presented in ordinary contexts (e.g., newspapers). Psychologists and other educated citizens often need to engage in this process as they wrestle with important issues. When personal narratives or persuasive anecdotes trump scientific reasoning in real-world decision-making (Lilienfeld, Ammirati, & David; 2012), serious negative public policy and health outcomes can follow (cf. Gross, 2009).

We undertook a review of the scientific literature to understand how others in the field measured deep learning across a departmental curriculum. We found few options; extant measures focused on replicative or applicative, rather than interpretive, knowledge. One (cf. Dolinsky & Kelley, 2010) was the multiple choice Major Field Test in Psychology (taken by over 6,300 undergraduates across over 200 colleges/universities between September 2014 and June 2015; Educational Testing Service, 2015). Another was the Psychological Critical Thinking Exam (PCTE; Lawson, 1999), a short answer test that probes for students' application of specific critical thinking concepts (e.g., falsifiability) based on brief descriptions of research studies and their conclusions.

Even at this early stage, we discussed how the interpretive learning we most wanted to impart (and assess) was not captured by existing measures because they relied on sequestered problem-solving rather than the kinds of complex, critical, and integrative thinking that we hoped our students would bring into far transfer contexts outside the classroom. Discovering that 'we need new measures of transfer' (Schwartz et al., 2005; p. 13), we decided to create our own, one that would be open-ended and 'messy,' requiring students to reconcile their own preconceptions with contradictory data using discipline-based ideas.

Just as our faculty group began its work, the local media was giving a great deal of coverage to a contentious community issue. Namely, families of high school students were advocating for a later school start time citing its positive impacts on adolescent health and learning based on what they claimed was overwhelming scientific evidence. Yet, when faculty examined this evidence,

we realized that much of the public discourse on the topic was not well-grounded in a nuanced interpretation of the studies based on scientific literacy and methods expertise. We saw school start time as a multifaceted issue that would be an engaging and relevant topic for our students and decided to place it at the center of our measure.

Based on information we provided to them, we developed an open-ended assessment demanding that students offer their best advice on whether the local high school should delay its school start time. In our assessment packet, we included a fact-based newspaper article about the topic as well as an op-ed against the delay from the same newspaper. We also included one-page synopses of two empirical scientific papers, one correlational and the other quasi-experimental, derived from real-life studies in the literature (cf. Owens, Belon, & Moss, 2010; Wahlstrom, 2002; Wolfson & Carskadon, 1998). Consistent with the conclusions of researchers in these actual peer-reviewed published studies, we crafted the article synopses so that they overstated the strength of their data in support of delayed school time. We also modified some of the specific findings in order to provide ambiguous and contradictory evidence of the benefits of late school start, making the task a more complex undertaking.

We set out directions for the students on the very first page of the packet. They were told to imagine that they were serving as an intern working for a local psychologist, Dr. Post, who was a member of the local school committee. Because the other members of the committee expected Dr. Post to share her professional opinion on the topic at their next meeting and Dr. Post was too busy to do a careful reading of the resources she collected on the issue, she asked the student interns in her office to help. The packet instructed students to use a *scientific perspective* to examine the evidence that the resources in the packet provide about school start time, further noting:

> Your job is to write an essay in which you state and explain your conclusions about the evidence, articulating its strengths and limitations. Be sure to include your best recommendation to Dr. Post about keeping versus delaying the school start time. Recall that Dr. Post is a trained psychologist who is familiar with scientific vocabulary within the discipline of psychology.

Involved faculty understood as we were crafting the assessment that it was a difficult task requiring students to sort through multiple sources, evaluate the quality of the varied knowledge claims in the packet, recognize contradictory evidence and the limits to 'expert' claims, and integrate their methods knowledge with the evidence and arguments in the packet in order to come to a well-informed conclusion. Indeed, the faculty in this department adopted these learning goals for the students in the major long before the start of this assessment project (http://www.smith.edu/ir/assessment/psychologymethods.php, accessed September 11, 2015). Thus, we felt that our open-ended task provided just the kind of challenging but important assessment of critical scientific thinking we target for our students. In order to compare our students' ability to apply their methods knowledge using applicative versus interpretative assessments, we also included 8 multiple choice questions at the end of our packet that drew on the most central concepts we teach. These questions were intended to give us an indication of how students perform on more traditional measures of learning.

After a brief pilot, we administered our assessment to students at various stages of our major, ranging from students in their first psychology class to seniors in the major who were

nearing the end of their studies in the discipline. A small group of faculty invested a great deal of time discussing how best to quantify and evaluate the essay responses. Focused on our initial pilot essays as well as a small random sub-group of essays from students at all stages of their college careers, we read student essays and began to articulate our shared assumptions about what constituted evidence that students could use methods concepts interpretively.

Just as we crafted multiple choice questions focused upon important methods concepts, we decided upon a coding rubric that reflected our learning goals. We coded for methods knowledge as well as overall problem-solving and critical thinking skills using 16 codes (for a full list, see Table 1).

Table 1

Rubric items

<table>
<tr><th colspan="2">Methods Knowledge</th><th rowspan="2">Overall Problem-Solving</th></tr>
<tr><th>Applied to a Correlational Study</th><th>Applied to a Quasi-Experimental Study</th></tr>
<tr><td colspan="2">Operationalization/measurement
Interpretation of data provided in study table
Recognition of evidence contradictory to researchers' claims
Strengths and limitations of the chosen research design
Statistical conclusion validity
External validity</td><td>Formulation of problem
Evidence used to make final recommendation
Final recommendation
Overall quality of writing</td></tr>
</table>

The rubric scoring scale was based on that developed for the AAC&U's Value rubric (www.aacu.org/value/rubrics, accessed September 8, 2015). Each item was scored on a 1 (poor), 2 (mixed) to 3 (good) scale, resulting in scores ranging from 16 to 48. Table 2 provides an exemplar rubric item from our assessment. Three coders met weekly until they were able to establish adequate inter-rater reliability. With a workable rubric, we were ready to code and assess student responses to the open-ended task.

Table 2

Exemplar rubric item from the correlational study

Category	3	2	1

	Good	Mixed	Poor
Interpretation of correlational data based on table	Provides accurate explanations of table data and makes correct inferences from that data; must specify directions of correlations; must mention at least 2 correlations. *E.g., accurately explains and interprets correlational data, and consistently makes correct inferences from these data*	Provides mostly accurate explanations of table data, but occasionally makes inferential errors or neglects to specify direction of correlation; mentions no more than 1 correlation. *E.g., accurately explains correlation data, but draws inconsistently correct inferences from these data.*	Either does not mention OR attempts to explain table data but does so at least partly incorrectly. *E.g., attempts to explain correlation data, but misinterprets the nature of that data, perhaps by labeling it as causal, misreading the direction of r, or interpreting non-significant results.*

We wanted to compare our advanced students in the major to students just beginning their study of psychology. We also wanted to look closely at students when they completed the research methods course. We were especially interested in comparing student performance on the applicative versus interpretive assessment format. In general, we anticipated stronger scores for more experienced students, both on the multiple choice replicative assessment and the open-ended interpretive assessment. The multiple choice questions resembled typical replicative assessments, in that only relevant information was presented, and there were limited options from which to choose.

On the other hand, we acknowledged early on that the open-ended assessment set a high bar for student understanding. Performing well on this interpretive assessment would be much more of a challenge. If our major was effectively preparing students to handle this sort of problem, we would find that students taking their first psychology course would perform worse than those who had just completed a research methods course. Seniors, having taken methods and additional courses in the major, ideally would outperform all other groups of students. However, as mentioned earlier, our anecdotal experiences had suggested that this might not be the case, and that graduates of the methods course might not possess the interpretive knowledge necessary for future successful transfer.

Insights on Student Learning from the Assessment

In order to understand the arc of student learning across our major, we gathered essays from the following groups:

- Students in our general introductory psychology course (N=8);
- Students in our introductory research methods course, at its start (for one section with an N=15) and end (for this same section plus another for a total N=29);
- Students in our elective advanced research methods course (N=15);
- Students nominated by at least one faculty member as one of that year's 'best' seniors (N=8).

All essays were anonymized and coded by faculty raters, blind to student course status.[1] Through quantitative and qualitative analysis of the essays, we achieved a number of important insights about student learning.

The conclusions we drew about student learning depended on how we measured understanding so it was important to assess what we were interested in knowing.

As hypothesized, our seniors outperformed Introductory Psychology students on the multiple choice test, providing evidence consistent with our classroom experiences indicating that our majors were able to apply their knowledge on traditional measures of applicative learning. We were quite surprised, however, to find that the seniors did not score significantly higher on the open-ended interpretive assessment than students who were just beginning their study of psychology (see Figure 1). Hence, seniors 'looked smarter' than introductory students on an applicative test, but looked equivalent to introductory students on a test of interpretive knowledge. Senior majors demonstrated mastery of the methods material that emerged when using a sequestered problem-solving assessment, allowing students to focus singularly on the narrow frame of the question at hand. In essence, senior majors' research methods learning was evident when they were prompted to access it but this learning did not necessarily express itself when they were presented with a messy, less structured, and more real-world-like problem.

[1] Interrater reliability analyses across pairs of three different faculty coders examining a subset of 10 essays indicated excellent reliability, ranging from .92 to .99.

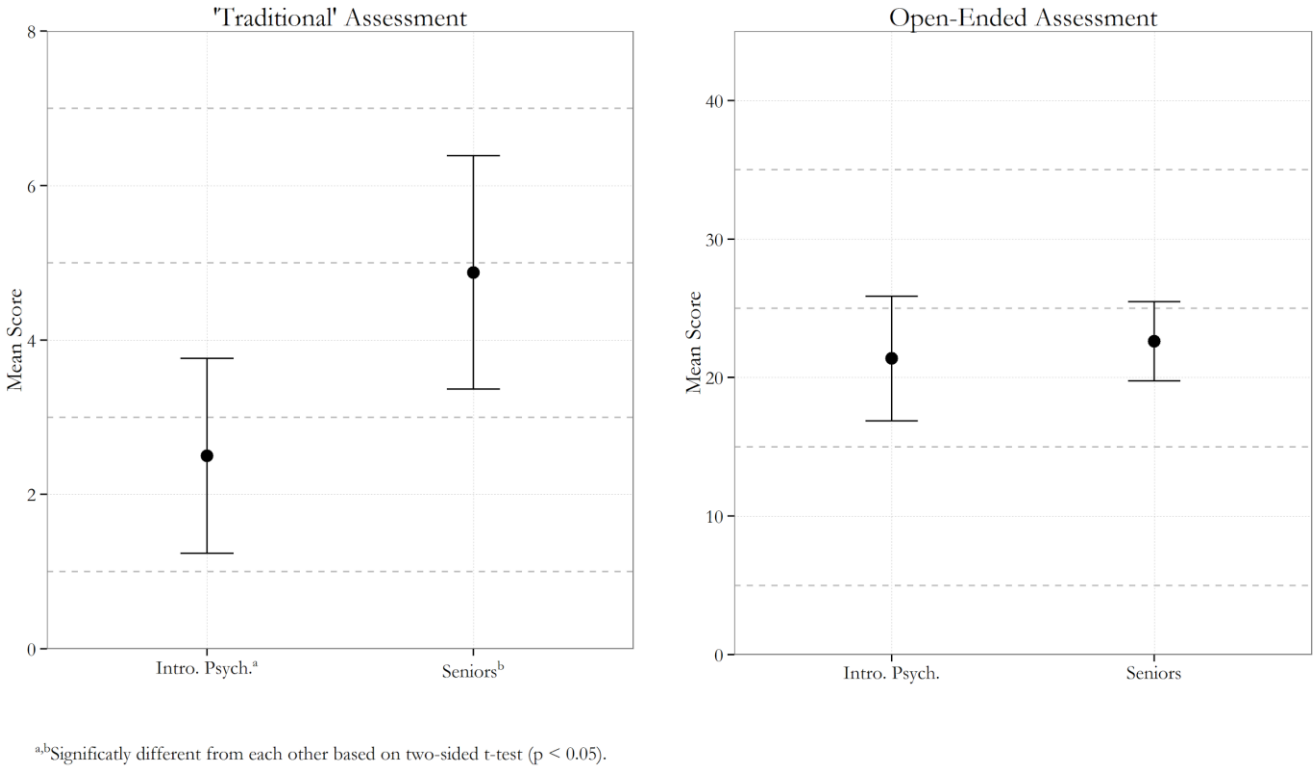


Figure 1. Mean scores and 95% confidence intervals for students in Introductory Psychology and senior Psychology majors on the 'traditional' multiple choice applicative learning assessment versus the open-ended essay interpretive learning assessment.

This was not the expected outcome. Although the traditional assessment of student learning indicated that our curriculum imparted applicative and replicative knowledge, student performance on the school start time problem indicated that our senior students were not able to organize, interpret, and think critically about the world in the way that we hoped and anticipated they would. Were we to base our judgments solely on student performance on the traditional assessment, we might be reassured about how much students had learned as they neared completion of their psychology major. As Schwartz et al. (2005) made clear, this kind of narrowly focused and sequestered assessment would have ignored important learning outcomes and misled us about the effectiveness of our instruction.

The faculty involved in the project began to discuss what sort of pedagogical approaches would best prepare our students for complicated, open-ended tasks and decided to dive more deeply into understanding student learning. We examined student responses on our interpretive essay task across the curriculum and engaged in a closer reading of the student essays themselves in order to observe patterns of learning. This led faculty to other important insights.

Students showed evidence of interpretive research methods knowledge in near, but not far, transfer contexts at various points in the arc of the major.

We anticipated that our students would score higher on the essay assessment as they moved along our major, after encountering ideas in different, presumably more complex contexts that would increase the depth of their knowledge over time. In our cross-sectional analysis, the performance of our students did not bear this out (see Figure 2). We wondered how, in comparison, our students would perform immediately after taking one of the methods courses. The assessment was given to students at the start and end of our introductory research methods course and also to a group of junior and senior students who elected to take an advanced research methods course in which they conducted their own original group research projects.

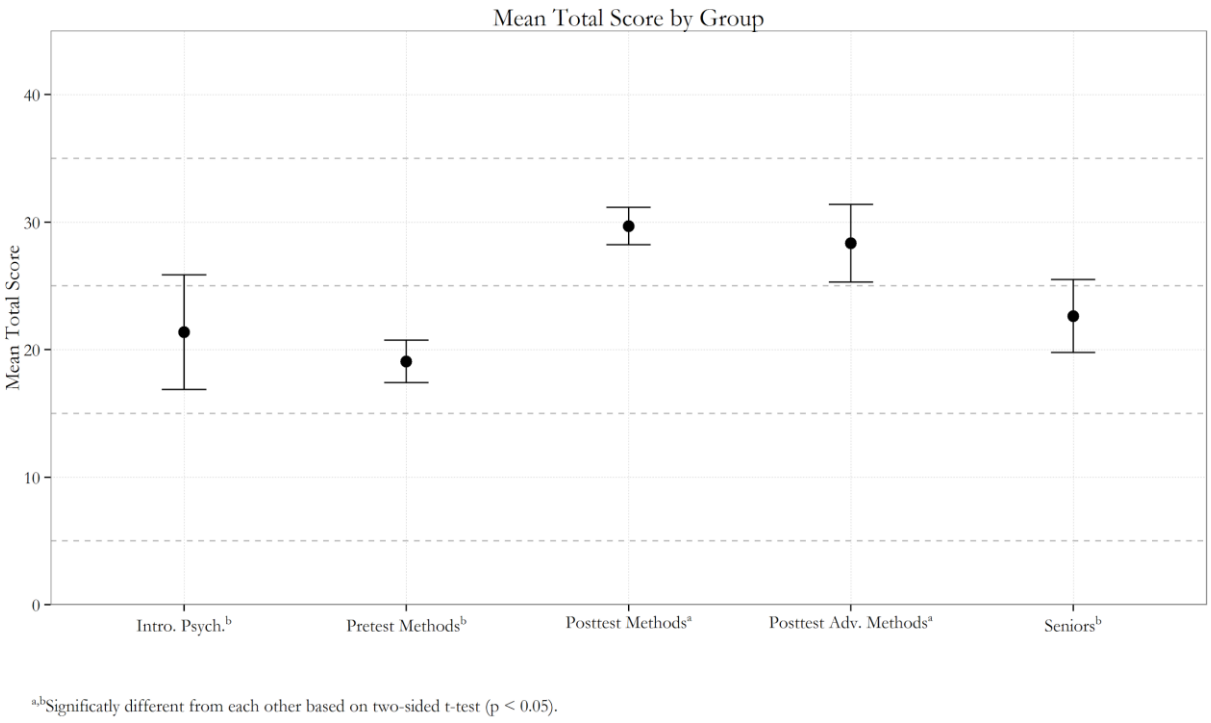


Figure 2. Mean scores and 95% confidence intervals on open-ended assessment at points along the curricular path of Psychology majors.

Students at the start of our introductory research methods course scored no better on the open-ended assessment than students at the end of their first psychology course. This was not surprising given that these courses are often taken close to one another in a student's course of study. At the end of the introductory research methods class, on the other hand, students evidenced significantly higher essay scores than when they began the course. Our advanced research methods students who took the assessment at the course's conclusion also outperformed our 'best' seniors, introductory students, and students just beginning the methods course. They did not have significantly higher scores than students at the conclusion of the introductory methods course.

These results suggest that when fresh in their minds, students are able to use methods concepts interpretively. Whether students' knowledge changes as a result of the advanced methods course is not something that the test or rubric scoring could discriminate.

We gained useful insights into student understanding by examining discourse closely, an exercise made possible by our narrative assessment approach.

The rubric for scoring the assessment awards points when specific methods concepts are invoked. Faculty wondered whether this obscured some bigger picture or overall understanding. To get beyond total essay rubric scores, we engaged in a closer reading of the essays themselves, gaining insight into several interesting features of student thinking. First, there was sometimes a lack of transfer of knowledge within student essays. For example, there were times when students seemed to understand the material (e.g., correctly interpreting results in a table, something the rubric valued) but their ability to interpret data did not transfer to inform their final recommendation. Or, students would note that the correlational study did not allow causal interpretations but relied on those same findings to make causal arguments. In some instances, students correctly noted the packet's contradictory evidence in an early part of their responses but ignored their own (accurate) interpretations in reaching a final recommendation. These students reverted to cherry-picking evidence that led them to an emphatic one-sided conclusion, without any attempt to explain these contradictory findings.

Over the course of reading dozens of essays, faculty agreed that the place where students' methods knowledge and critical thinking converged was the part of the rubric assessing the quality of their final recommendation. The best student answer was that the school start time should not be changed because the design of the studies in the packet precluded causal claims and their data in support of school start delay were not compelling (e.g., GPA was unrelated to increased sleep/school start delay). The strongest student essays recognized the contradictory and null evidence about the benefits for delayed start:

> It is true, looking at both studies, that delayed start times result in students sleeping more and being less sleepy. However, there is no evidence that this increase in sleep has any impact on academic performance, when measured in GPA. Additionally, [one study] suggests that later start times might result in increased tardiness. Therefore, there is no evidence that delayed start times are beneficial beyond allowing students to get more sleep. Based on this research, Northampton should not delay the start time of it's [sic] high school.

In contrast, weak student essays either relied on the opinion pieces and/or overstated the impact of delayed school start benefits based on the data. Faculty agreed that unquestioningly accepting the researchers' conclusions and using them as a framework for interpreting the data as consistently in favor of delayed start raised doubts about whether a student possessed a deep understanding of research methods. Although almost all students purported to base recommendations on the scientific data (rather than the opinion-based pieces), they often based their recommendations on the researchers' overly neat and inaccurate descriptions of the data in the article summaries. Many errors were based on students' inappropriate description of the findings:

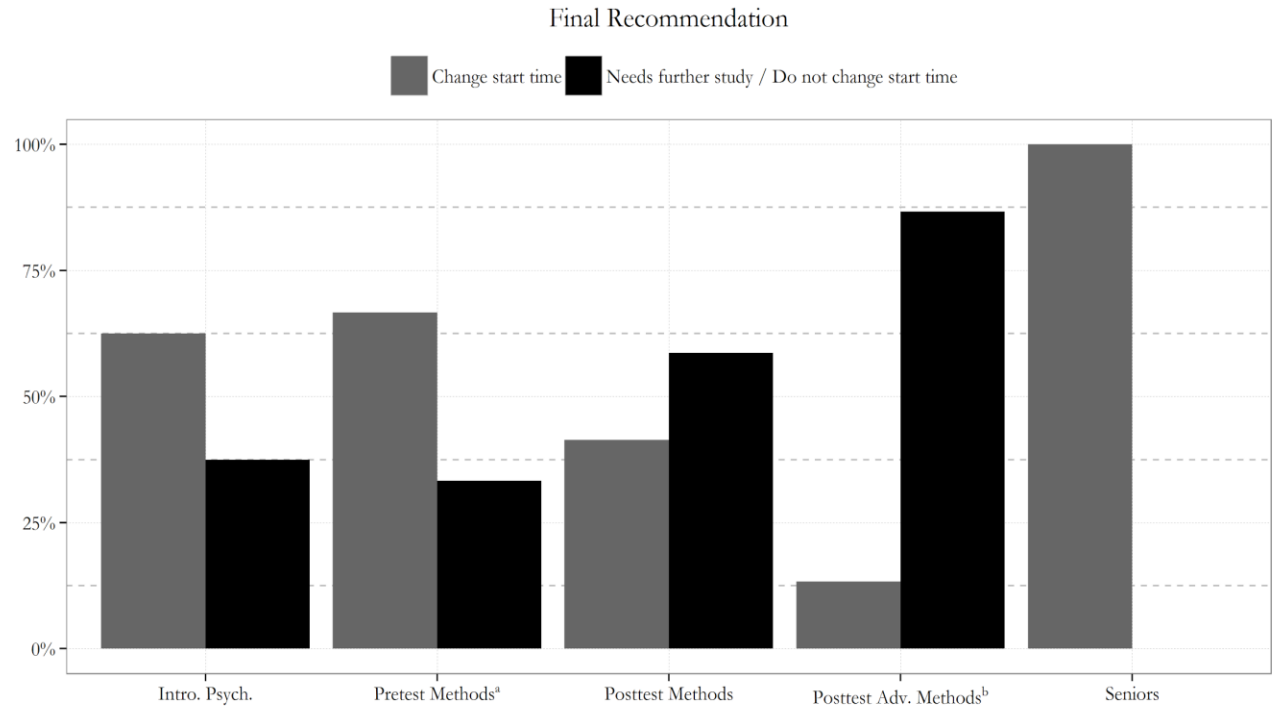
> The proof found in both scientific research articles shows positive outcomes to creating a later start time. I feel that the most significant outcome to these studies was the rise in GPA.

> When looking into this situation, one should keep in mind that improving our student's [sic] academic performance is our prime focus; if research shows this improvement to be true, there is no need in fighting the evidence.

Occasionally, some weak student essays resorted to hyperbole and emotional reasoning (Lilienfeld et al., 2012) in order to justify a final recommendation, but these were rarer:

> Not taking the money and one-time transition for the students is simply cruel to them and the future children they may educate. Not to mention, it is lazy. This one change could positively impact an entire community for decades.

Our more fine-grained look at final recommendations shows students completing the introductory and advanced methods courses performing best (see Figure 3). These students were more likely to make a recommendation to Dr. Post that she should testify against delaying school start (with either a clear vote against changing or a recommendation that there was need for further study before investing in the time and money required logistically by such a change because of the conflicting evidence). Final recommendation was an area where our seniors performed most poorly, with every student stating that school start should be delayed. This is something that we need to better understand.



[a,b]Significatly different from each other based on two-sided z-test ($p < 0.05$).

Figure 3. Proportion of students at points along the curricular path of Psychology majors making different final recommendations.

Translating Insight to Action

According to Walvoord (2010; p. 4), 'the end of assessment is action,' and indeed, the design research our faculty undertook provided insights about student learning and inspired change within our department. We outline these benefits below.

Thinking like a psychologist' is a challenging intellectual framework to develop, requiring intensive and explicit attention from faculty through pedagogy and curriculum.

Faculty involved in the project discussed the department's learning expectations and environments and worked to understand the pattern of assessment results observed. Our majors showed higher levels of interpretive knowledge when assessed in proximity to a methods course (whether at the introductory or advanced levels). Indeed, faculty members who teach methods in the department all anecdotally reported that students make significant gains in their scientific thinking over the course of the term. But our seniors, those nominated as some of the 'best' by at least one faculty member in our department, failed to outperform beginning students on our interpretive assessment measures. These data forced us to question whether our curriculum fully prepared students to become critical consumers of knowledge in a world in which the rise of the internet created a proliferation of 'evidence' that requires, now more than ever, citizens who are transliterate (Liu, 2004). Many of our courses had relied on assessments that were carefully scaffolded with clear direction and void of any distracting information, rather than requiring that students sort and organize information validly based on their understanding of epistemology and methods in order to come to well-reasoned and empirically-based decisions. We questioned whether our assessments and curriculum fully prepared our students to participate in the sustained knowledge advancement seen as essential for social progress of all kinds and for the solution of societal problems (Scardamalia & Bereiter, 2006). In our interpretive assessments, faculty found that student thinking sometimes tilted away from awarding appropriate weight to empirical evidence and inappropriately relied on what passed as accepted knowledge or the wisdom of common sense (Stanovich, 2010).

We spent a great deal of time considering the trajectory of student learning and how the kinds of thinking that students bring into their course of study are powerful, but not always valid, organizers of their understanding of how the world works. According to Lilienfeld and colleagues (2012), 'scientific thinking...is unnatural: it must be learned, practiced, and maintained as a habit of mind' (p. 13). This kind of thinking is difficult to master because of the power of human cognitive biases and heuristics (Lilienfeld et al., 2012) in creating shortcuts for predicting and understanding the world. These biases 'transfer in' to classrooms with powerful impacts on how knowledge develops, persisting even for undergraduate students who are emerging experts within a discipline (cf. Coley & Tanner, 2015). We realized that students do not develop deep understanding all at once and we needed to pay more explicit attention to students' intuitive ways of thinking as we teach them about scientific thinking.

Assessment helped us to improve and evolve our collective ideas about student learning and our pedagogies. These ideas had resonance beyond the faculty involved in the project.

The group of faculty working most closely on the assessment project were members of the department who taught sections of our introductory research methods course. During the time this group began to analyze data, the psychology department undertook a year-long review of the curriculum and student learning. In those department-wide conversations, faculty discussed a concern that emerged from advising students within the major: some of them did not share the defining view held by faculty that psychology is a science (e.g., when students told advisers that they wanted to get methods/statistics 'out of the way'). Through our review discussions, departmental faculty as a whole came to the conclusion that these student views were not successfully challenged by the structure of our major. Our students were able to move through our requirements without consistently being asked to think in ways that paid explicit and critical attention to the strength of knowledge claims.

Through these conversations, faculty in the department realized that we had few, if any, prerequisites for our intermediate courses. Partly this was an attempt to make our classes widely accessible, based on our shared belief that liberal arts students would benefit from exposure to the discipline of psychology. The department undertook an internal review of syllabi that revealed that our intermediate classes, particularly those with higher enrollments, had fewer rigorous learning opportunities focused explicitly on honing scientific thinking, including fewer pages of total writing and fewer empirical primary source reading assignments (in favor of more textbook reading assignments).

As a direct consequence of these discussions, we introduced a set of limited enrollment intermediate colloquia that scaffolded disciplinary skills focused on psychology as a science (through close reading of primary source journal articles; careful review of study tables and graphs; more class time devoted to discussion of studies). We envisioned these colloquia as opportunities for intensive student discourse devoted to the development of scientific and critical thinking skills, through collaborative projects, substantive writing projects, and discussion-heavy class meetings.

At the time faculty conversations began, each of the sections of the research methods course were taught differently, without a common text or instructional approach. The conversations seeded by the assessment project and related curricular review helped to create a variety of commonalities across sections. A good example is that almost all of us adopted the same text (Morling, 2015) because of its central organization around 'Three Claims, Four Validities,' a framework that provides a clearer schema around which students could hang and retain their methods knowledge and ability to interrogate knowledge claims. Many of us now also share the on-line quiz resource from this text to 'flip' our classrooms, making space for more engagement with our students during class time on tasks that require interpretive knowledge.[2] We are also paying fuller attention to teaching students about the tendency and pitfalls of biases and heuristics.

In general, our conversations and assessment data convinced us as a faculty to devote more of our teaching to messy interpretive kinds of active learning. These changes are consistent with findings in the learning sciences that students learn best when they are given opportunities to

[2] Quizzes are untimed and open book, focusing on applicative knowledge questions. Students can take a chapter quiz as often as is necessary to reach mastery defined as a score ≥ 80%, (with a rotating bank of questions).

engage in problems that call for deepening their understanding by articulating their ideas via collaborative and sustained discourse (Barnes & Todd, 1995; Sawyer, 2014).

Conclusion

Developing the assessment and its rubric and examining these data collectively led to productive, ongoing conversations among our faculty about learning in the discipline of psychology. The essay task gave us the opportunity to look at our students' capabilities and thinking in ways not allowed by more traditional assessments. It has left a number of questions for us. At this point, our new curriculum has been in place for over 2 years and we need to review student outcomes more systematically to understand where our teaching and student learning stand within our department. We are also piloting work to expand the rubric in order to get at additional interpretive knowledge features and to explore what other kinds of messy problems might require different and important ideas to solve.

Like Lawson (1999), we do not intend to present our assessment as a standardized measure for the field. Rather, as we see it, the power of this assessment is its connection to authentic local faculty questions about complex, interpretive student learning. As Walvoord (2010) notes, assessing certain 'ineffable qualities' (p. 6) expected from learning can be daunting. Nonetheless, this attempt to understand our students helped hone faculty thinking about teaching and learning both within and outside of our department (cf. Rowen et al., 2015). This, in turn, fostered faculty ownership over the curriculum and led to meaningful change, outcomes in stark contrast to some notions of assessment as superfluous exercises required by external accreditors.

Acknowledgements

This work was supported by the Davis Educational Foundation as well as the Smith College Sherrerd Center for Teaching and Learning.

We would like to thank Annaliese Beery, Phil Peake, Bill Peterson, Beth Powell, and Cate Rowen for their working group participation in assessment and/or rubric development. Thanks to Kathryn Aloisio, Emily Pagano, Dana Sherwood, Kevin Shea, and Christine Pelletier for their logistical support. Special thanks as well to Bill Peterson for his contributions to essay coding.

Portions of this paper were presented at the January 2015 Association of American Colleges and University meeting in Washington, DC.

References

American Psychological Association. (2013). *APA guidelines for the undergraduate Psychology major: Version 2.0.* Retrieved from http://www.apa.org/ed/precollege/undergrad/index.aspx

Barnes, D., & Todd, F. (1995). *Communication and Learning Revisited: Making Meaning Through Talk.* Portsmouth, NH: Boynton/Cook Publishers.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology, 15*(5), 153-166. doi: 10.1037/0278-7393.15.1.153

Bransford, J. D., Brown, A. L., & Cocking, R.R. (eds. 2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, D.C.: National Academy Press.

Bransford, J. D., & Schwartz D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61-100.

Bransford, J. D., & Stein,B. S. (1993). *The Ideal Problem Solver.* New York: Freeman.

Broudy, H.S. (1977). Types of knowledge and purposes of education, *Schooling and the Acquisition of Knowledge*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152. doi:10.1207/s15516709cog0502_2

Coley, J. D., and Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE--Life Sciences Education, 14*(1), 1-19. doi:10.1187/cbe.14-06-0094

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences, 3*(1), 15-42. doi:10.1207/s15327809jls1301_2

De Groot, A. D. (1965). *Thought and Choice in Chess*. The Hague: Mouton.

Dolinsky, B. & Kelley, J. M. (2010). For better or for worse: Using an objective program assessment measure to enhance an undergraduate psychology program. *Teaching of Psychology, 37*(4), 252-256. doi:10.1080/00986283.2010.510978

Dunn, D. S., McCarthy, M. A., Baker, S., Halonen, J. S., & Hill, G. W. (2007). Quality benchmarks in undergraduate psychology programs. *American Psychologist, 62*(7), 650-670. doi:10.1037/0003-066X.62.7.650

Educational Testing Service. (2015). 2015 Major field test comparative data guide: Major field test for psychology. *Educational Testing Service.* Accessed September 8, 2015, from

http://www.ets.org/s/mft/pdf/acdg_psychology.pdf

Gibson, J. J. & Gibson, E. J. (1995). Perceptual learning: Differentiation or enrichment? *Psychological Review*, *62*(1), 32-51.

Greeno, J. G., Smith, D. R., & Moore, J. L. (1993). Transfer of situated learning. *Transfer on Trial: Intelligence, Cognition, and Instruction*, 99-167. Norwood, NJ: Ablex.

Gross, L. (2009). A broken trust: Lessons from the vaccine-autism wars. *PLOS Biology, 7*(5), 1-7. doi:10.1371/journal.pbio.1000114

Lawson, T.J. (1999). Assessing psychological critical thinking as a learning outcome for psychology majors. *Teaching of Psychology, 26*(3), 207-209. doi:10.1207/S15328023TOP260311

Lilienfeld, S.O., Ammirati, R., & David, M. (2012). Commissioned Article: Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology,* 507-36. doi:10.1016/j.jsp.2011.09.006

Liu, A. (2004). *The Laws of Cool: Knowledge Work and the Culture of Information*. Chicago: University of Chicago Press.

Marton, F. & Booth, S. (1997). *Learning and Awareness*. Mahwah, NJ: Erlbaum.

Morling, B. (2015). *Research Methods in Psychology: Evaluating a World of Information.* New York: W.W. Norton.

Owens, J. A., Belon, K. & Moss, P. (2010). Impact of delaying school start time on adolescent sleep, mood, and behavior. *Archives of Pediatric and Adolescent Medicine, 164*(7): 608-614. doi:10.1001/archpediatrics.2010.96

Rowen, C., DiBartolo, P. M., & Jamieson, E. R. (2015). "The pHunger Games and Think Like a Scientist: How One College Used Direct Assessment of Student Learning to Strengthen Faculty Governance of the Curriculum." Roundtable discussion at the meeting of the Association of American Colleges and University meeting, Washington, DC, January 21-24.

Sawyer, R.K. (2014). Introduction: The new science of learning. *Cambridge Handbook of the Learning Sciences*, 1-16. New York: Cambridge University Press.

Scardamalia, M. & Bereiter, C. (2006). Knowledge building: theory, pedagogy, and technology. *Cambridge Handbook of the Learning Sciences*, 97–115. New York: Cambridge University Press.

Schuyler, D. (2003). *Cognitive Therapy*. W. W. Norton & Company.

Schwartz, D.L., Bransford, J. D. & Sears, D. (2005). Efficiency and innovation in transfer. *Transfer of Learning: Research and Perspectives*, 1-51. Charlotte, NC: Information Age Publishing.

Stanovich, K.E. (2010). *How to Think Straight About Psychology*. Boston: Allyn & Bacon.

Wahlstrom, K. (2002). Changing times: Findings from the First Longitudinal Study of Later High School Start Times. *NASSP Bulletin, 86*(633): 3-21.

Walvoord, B.E. (2010). *Assessment Clear and Simple: A Practical Guide for Institutions, Departments, and General Education.* San Francisco: Jossey-Bass.

Wolfson, A.R., & Carskadon, M.A. (1998). Sleep Schedules and Daytime Functioning in Adolescents. *Child Development, 69*(4): 875-887. doi:10.2307/1132351