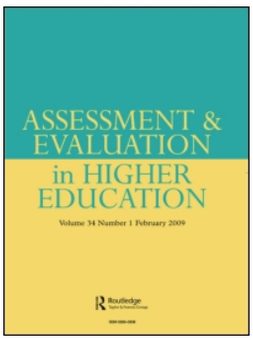

Routledge
Taylor & Francis Group

Assessment & Evaluation in Higher Education

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Student learning in higher education: a longitudinal analysis and faculty discussion

Catherine E. Mathers, Sara J. Finney & John D. Hathcoat

To cite this article: Catherine E. Mathers, Sara J. Finney & John D. Hathcoat (2018) Student learning in higher education: a longitudinal analysis and faculty discussion, Assessment & Evaluation in Higher Education, 43:8, 1211-1227, DOI: 10.1080/02602938.2018.1443202

To link to this article: https://doi.org/10.1080/02602938.2018.1443202

ASSESSMENT & EVALUATION IN HIGHER EDUCATION, 2018
VOL. 43, NO. 8, 1211–1227
https://doi.org/10.1080/02602938.2018.1443202

Routledge
Taylor & Francis Group

Student learning in higher education: a longitudinal analysis and faculty discussion

Catherine E. Mathers [1], Sara J. Finney and John D. Hathcoat

Center for Assessment & Research Studies, James Madison University, Harrisonburg, VA, USA

ABSTRACT

Answering a call put forth decades ago by the higher education community and the federal government, we investigated the impact of US college coursework on student learning gains. Students gained, on average, 3.72 points on a 66-item test of quantitative and scientific reasoning after experiencing 1.5 years of college. Gain scores were unrelated to the number of quantitative and scientific reasoning courses completed when controlling and not controlling for students' personal characteristics. Unexpectedly, yet fortunately, gain scores showed no discernable difference when corrected for low test-taking effort, which indicated test-taking effort did not compromise the validity of the test scores. When gain scores were disaggregated by amount of completed coursework, the estimated gain scores of students with quantitative and scientific reasoning coursework were smaller than what quantitative and scientific reasoning faculty expected or desired. In sum, although students appear on average to be making gains in quantitative and scientific reasoning, there is not a strong relationship between learning gains and students' quantitative and scientific reasoning coursework, and the gains are less than desired by faculty. We discuss implications of these findings for student learning assessment and learning improvement processes.

KEYWORDS

Higher education assessment; learning improvement; examinee motivation

The need to assess student learning in higher education

Given the purpose of higher education, students, faculty and administrators typically assume university curricula lead to gains in knowledge and skill. Yet globally, 'Key questions include whether, how, and to what extent academic competencies can be taught and acquired in various fields of study and types of higher education institutions, such as universities, universities of applied sciences, technical colleges and so on.' (Zlatkin-Troitschanskaia, Pant, and Coates 2016, 656). In the United States, scant data exist to support the influence of college coursework on learning gains. Educational researchers (e.g. Ewell 1983, 1985) and the U.S. Department of Education (2006) have been calling for the collection of student learning data for decades. As the American Association for Higher Education (1992) noted in the early nineties, 'As educators, we have a responsibility to the publics that support or depend on us to provide information about the ways in which our students meet goals and expectations' (3).

If faculty know how much or little students are learning, they may be motivated to make improvements to curricula and pedagogy (Fulcher et al. 2014). Understandably, estimates of learning must be of high psychometric quality to accurately inform curriculum modifications (Coates 2014). Unfortunately, few US institutions collect the type of data that allow faculty to understand how much students are

CONTACT Catherine E. Mathers catherine-mathers@uiowa.edu
[1]Present address: Department of Psychological and Quantitative Foundations, University of Iowa, Iowa City, IA, USA.

learning, what factors contribute to academic growth, and whether gains align with faculty expectations. For example, many institutions collect information about experiences that may contribute to academic growth (e.g. the National Survey of Student Engagement) without examining how much students learn over time and the extent to which such gains align with faculty expectations (Kuh 2009). In this study, we estimated student learning gains across several cohorts of college students, and examined how an institution's curriculum related to learning gains after controlling for personal characteristics (i.e. ability, gender, test-taking motivation). Additionally, faculty discussed their expectations and desires for learning gains which were then compared to empirically estimated gains. Results from this study facilitate greater understanding of learning in college and encourage a culture of learning improvement.

Conceptualising and measuring student learning

Institutions often simply assess *student competency,* or the knowledge and skills students have at the time of assessment (e.g. students' mathematics skills during spring semester of their first year; U.S. Department of Education 2006). Institutions often attempt to infer student learning, or *change* in knowledge and skills within individuals, from data collected using cross-sectional designs (Liu 2011). In these designs, the competency estimate for a group of first-year students is typically compared to that from an independent group of upper-class students (sophomore, junior or senior level students) who may have completed particular coursework. These designs can be problematic because the two samples likely differ in demographic, motivation and academic variables that influence competency, thus compromising inferences about student learning.

Longitudinal designs are more appropriate because they allow faculty to track students over time and thus obtain an estimate of learning gain (Castellano and Ho 2013). A positive change in competency is a learning *gain*. Thus, faculty must collect data on students' prior competency as well as current competency (e.g. students' mathematics skills during spring semesters of their first and second years). Students complete the same test, or psychometrically equivalent tests, both before (pretest) and after (posttest) completing coursework. To determine whether learning gains are due to particular coursework or due to increases in general cognitive development, the estimated learning gains of students who have completed the particular coursework can be compared to the estimated learning gains of those students who have not. Estimates of competency and estimates of learning are closely intertwined – the difference in a student's competency across multiple assessments is the student's estimated learning gain.

Longitudinal designs are also critical for determining *learning improvement*, which is an increase in student learning gains between a cohort that experienced a modified programme/curriculum and a cohort that experienced the original programme/curriculum (Fulcher et al. 2014). Modifications to improve the programme are informed by previous student learning assessment results associated with the original programme/curriculum. The programme/curriculum is then reassessed to determine if the modifications increased learning gains. Thus, the term 'learning improvement' applies to programmes/curricula that have experienced effective modifications. Learning improvement serves as the motivating reason for engaging in higher education outcomes assessment (Borden and Peters 2014). However, few institutions estimate learning improvement (Banta and Blaich 2011; Fulcher et al. 2014). One reason may be that relatively few institutions assess student learning gains.

Student learning gain studies

Only a few research teams have investigated student learning gains in the US using longitudinal methodologies. In their book *Academically Adrift* (2011), Arum and Roksa presented longitudinal Collegiate Learning Assessment (CLA) data (2322 students from 24 four-year institutions were assessed in Fall 2005 and Spring 2007). The CLA is purported to assess general skills in critical thinking, complex reasoning and writing. Students gained .18 standard deviations (computed using the standard deviation of the pretest scores), on average, after three semesters in college (34.32-point gain on a scale from 400 to

1800). In their follow-up study (Arum and Roksa 2014), 1666 of the students initially tested as first-year students were re-assessed four years later. After seven semesters in college, the learning gain estimates were .47 standard deviations (86-point gain).

Blaich and Wise (2011), lead researchers on the Wabash National Study, collected student learning data over a span of four years from 49 American colleges and universities. Their results, similar to those of *Academically Adrift*, indicated that after four years of college coursework, students gained almost half a standard deviation in critical thinking ($d = 0.44$, computed using the standard deviation of the pretest scores) compared to only a .11 standard deviation gain after one year in college as measured by the Collegiate Assessment of Academic Proficiency Critical Thinking Test (Pascarella et al. 2011).

Roohr, Liu, and Liu (2016) investigated learning gains across three cohorts of college students using the Educational Testing Service (ETS) Proficiency Profile. They found no significant learning gains in critical thinking, reading, writing or mathematics after one or two years of college. After three years of college, students gained the most in mathematics ($d = 0.42$, computed using the standard deviation of the gain scores, or 2.72 points on a scale from 100 to 130) and reading ($d = 0.46$, computed using the standard deviation of the gain scores, or 2.64 points on a scale from 100 to 130). Gains were similar after four or five years in college (mathematics: $d = 0.41$, computed using the standard deviation of the gain scores, or 2.70 points; reading: $d = 0.41$ or 2.85 points).

Unfortunately, Arum and Roksa (2011), Blaich and Wise (2011) and Roohr, Liu, and Liu (2016) did not link the estimated learning gains to completion of coursework intentionally designed to impact these specific skills and knowledge. Gains were aggregated across students who varied in exposure to domain-specific coursework (e.g. some students may have completed no mathematics courses, whereas others may have completed several courses).Thus, inferences regarding the impact of intentionally designed curriculum on student learning are extremely limited from these results and, in turn, evidence-based curriculum modifications are nearly impossible. When discussing reactions to the learning gain estimates from the Wabash Study, Blaich and Wise (2011) noted: 'Despite the abundant information they receive from the study, most Wabash Study institutions have had difficulty identifying and implementing changes in response to study data.' (3).

With the goal of linking learning gains to curriculum exposure to inform learning improvement efforts, Pastor, Kaliski, and Weiss (2007) estimated history and political science learning gains after students completed none, one or two courses in that domain of study. A year and a half after beginning college, students who completed one history or political science course gained about half a standard deviation ($d = 0.41$ or 0.54, computed using the standard deviation of the pretest scores; 4 points on 81-item test). Students who completed both courses achieved larger gains ($d = 0.90$; 7 points).

Using two cohorts of students, Hathcoat, Sundre, and Johnston (2015) investigated learning gains in quantitative and scientific reasoning. They disaggregated these estimates by those students who completed the required 10 credit hours in the quantitative domain and those students yet to complete the requirement. After 1.5 years of exposure to college coursework, students who completed the 10 credit hour requirement had moderate estimated standardised gains ($d = 0.46$ and 0.52 for cohort 1 and 2; 3.49 and 2.97 points on 66-item test). However, students who had not completed all 10 credit hours also made moderate gains during the same period of time ($d = 0.42$ and 0.67, unspecified metric, for cohort 1 and 2; 3.13 to 3.23 points). Thus, completing 10 credit hours of quantitative and scientific reasoning coursework did not appear to increase students' learning gains relative to completing fewer credit hours.

Student characteristics may influence learning gains

Arum and Roksa (2011) encouraged educational researchers to measure learning longitudinally *and* to investigate the effects of both curriculum and personal characteristics on learning gains. Informing the need for our study, the authors remarked how few US researchers were conducting such studies. A review of the literature seems to support this statement. Most studies investigating the impact of curriculum *and* personal characteristics examine competency rather than learning gains.

Longitudinal studies estimating learning gains and examining personal characteristics yield contradictory results. Arum and Roksa (2011) examined high school characteristics, ethnicity, gender, academic preparation, ability and parents' education, and found that only ethnicity moderated learning gains. The Wabash National Study found ability and gender interacted with some high-impact practices to influence student learning gains (Pascarella and Blaich 2013). Roohr and colleagues (2016) found no personal characteristics (i.e. gender, race/ethnicity, STEM major status, SAT/ACT scores (standardised test scores typically used for college admissions) and first-year grade point average (GPA)) predicted mathematics gains.

Students' personal reactions to the test can also impact learning gain estimates (Swerdzewski, Harmes, and Finney 2009). For example, low-stakes tests are regularly used for institutional accountability mandates and learning improvement initiatives (Ewell 2004). Students may not expend effort on low-stakes assessments because there are no personal consequences attached to poor test scores (e.g. Finney, Myers, and Mathers forthcoming; Musekamp and Pearce 2016; Wise and Smith 2016), which may attenuate learning gain estimates (Finney et al. 2016; Wise and DeMars 2010). Consequently, faculty may erroneously conclude that students are not learning if they fail to correct for low motivation on low-stakes tests.

Purpose of the current study and hypotheses

Given limited study of student learning gains, the purpose of the current study was to: (1) estimate learning gains by employing a longitudinal design, (2) evaluate if domain-specific curriculum impacted gains as intended, and (3) document faculty reactions to the magnitude of the gains. We employed a mixed methods explanatory sequential design (Creswell and Plano Clark 2011); qualitative data obtained from faculty interviews were collected to inform the results of a larger quantitative study where student learning gains were estimated from multiple cohorts of college students.

For the quantitative strand, students within each cohort were randomly assigned to complete a quantitative and scientific reasoning test at the beginning of their first year of college, and again after completing three semesters of college coursework. Thus, the random samples for each cohort represent the university population. We computed two learning gain estimates: Cohen's *d* and raw gain score. Cohen's *d* estimates from this study were compared to the standardised gain estimates from other learning gain studies with similar quasi-experimental designs (i.e. Pastor, Kaliski, and Weiss 2007) or domains of interest (i.e. Roohr, Liu, and Liu 2016). Four hypotheses based on national trends in college learning and which align with the goals of higher education were tested using the quantitative data:

(1) Moderate learning gains will be observed when collapsing data across completed courses.
(2) Gains will increase with increased domain-specific coursework.
(3) Removing unmotivated students will result in larger learning gains.
(4) Coursework will predict gains after controlling for gender and ability.

Unlike the quantitative phase, the qualitative strand of the study was largely exploratory. In this phase of the study, faculty members who taught courses designed to enhance quantitative and scientific reasoning were interviewed regarding learning gains. More specifically, the qualitative data were used to explore the following questions:

(1) What are faculty members' expectations and desires for student learning gains?
(2) How do these expectations and desires align with the learning gain estimates obtained during the quantitative phase of the study?

Answers to these questions put the learning gains in context and begin to give them meaning necessary for learning improvement efforts. As noted by Pascarella and colleagues (2011, 23):

> As far as we know, however, no one has come up with an operational definition of just how much change we should expect on such instruments during college if we are to conclude that postsecondary education is doing the job it

> claims it is. Some human traits are simply less changeable than others, and that needs to be considered. Until we can come up with standards of expected change during college, the meaning of average gain scores like the ones reported above will be largely in the eye of the beholder. One person's 'trivial' may be another person's 'important'.

Pairing the empirical learning gains with the expectations for learning from faculty who designed both the assessment and the courses begins to shed light on this issue.

Methods

Participants and procedures for estimating and predicting learning gains

At the US public university where this study was conducted, the effectiveness of the general education curriculum has been assessed for over twenty years during the biannual Assessment Day that is held once before the start of the fall semester and again several weeks into the spring semester. All first-year students are tested during the fall. Upper-class students are tested during the spring once they have accumulated between 45 and 70 credit hours. These longitudinal data allow for the computation of gain scores, which can be used for accountability purposes and improvement of general education curriculum.

Each student does not complete all tests administered on Assessment Day. Students are randomly assigned to a testing room based on the last few digits of their ID number. Each testing room corresponds to a specific battery of tests comprised of cognitive and non-cognitive measures, which takes approximately two hours to complete. Assigning students to test configurations by their ID enables university assessment experts to assign students to the same battery as first-year students and 1.5 years later as upperclassmen. Performance on the tests does not affect graduation or course grades; hence, the tests are low stakes for students.

Assessment Day data used in this study were collected from five cohorts: 2007–2009, 2008–2010, 2013–2015, 2014–2016, and 2015–2017. Differences in gain scores across cohorts failed to be practically meaningful $F(4, 1549) = 5.851$, $p < .001$, $\eta^2 = 0.02$); thus, cohorts were combined to produce more stable learning gain estimates.

Measures for estimating and predicting learning gains

Quantitative and scientific reasoning test

Quantitative and scientific reasoning was assessed using a 66-item quantitative and scientific reasoning test developed by faculty and university assessment consultants to align with the general education quantitative and scientific reasoning learning objectives. Psychometric study of the scores supports the computation of one total quantitative and scientific reasoning score (Sundre, Thelk, and Wigtil 2008). Total scores evidenced good reliability at both testing occasions (pre-test $\alpha = 0.74$; post-test $\alpha = 0.81$; $N = 1554$).

Number of courses completed

University faculty designed a set of general education courses intended to increase quantitative and scientific reasoning. This mathematics and science curriculum covers three topics: 'Quantitative Reasoning', 'Physical Principles', and 'Natural Systems'. We gathered data on the number of relevant courses students completed upon the second testing occasion, which ranged from zero to seven.

Academic ability

Academic ability estimates, as reflected via total SAT or ACT scores, were gathered from university records to estimate the effect of ability on learning gains. For students who did not have SAT data but completed the ACT ($n = 25$), ACT scores were converted to the SAT metric using concordance tables (Dorans 1999).

Gender

Gender information was gathered from university records to determine how gender relates to learning gains and if gender moderates relationships between learning gains and other predictors (i.e. number of courses, prior ability).

Test-taking effort

Test-taking effort was assessed via the five-item effort subscale of the Student Opinion Scale (SOS; Sessoms and Finney 2015; Thelk et al. 2009). Two versions of the SOS are available: a *test session-specific* measure and a *test-specific* measure. The test session-specific SOS is administered at the end of a test battery to assess student motivation across *all tests in the session*. The test-specific SOS is administered at the end of a test to assess student motivation on *that particular test*. Instructions for these measures differ slightly to distinguish the context (session or test) but the items are essentially identical (e.g. 'I engaged in good effort throughout these tests' versus 'I engaged in good effort throughout this test'). The test session-specific SOS (Thelk et al. 2009) and the test-specific SOS (Finney, Mathers, and Myers 2016) have been shown to have adequate reliability. Across cohorts ($N = 1554$), reliability estimates were of acceptable magnitudes at pretest (test session-specific $\alpha = 0.82$; test-specific $\alpha = 0.79$) and posttest (test session-specific $\alpha = 0.79$; test-specific $\alpha = 0.81$).

Analyses for estimating growth and predicting growth

Unfiltered learning gains (i.e. raw gain scores) were computed by subtracting pretest scores from posttest scores to estimate individual learning gain on the metric of the points gained ($N = 1554$; see Table 1). We then recomputed the learning gains after filtering, or removing, examinees with low motivation, using the test-session and test-specific effort scores to evaluate if both provided similar estimates. The first cohort did not complete either effort subscale; therefore, their data were not used to investigate the impact of low effort on learning gains. Some students in the 2008–2010 cohort only completed the test-session specific SOS; other students in this cohort only completed the test-specific SOS (see Table 1). To ensure we did not inadvertently remove students of low ability, we compared the SAT scores of the filtered sample to the unfiltered sample (Wise, Wise, and Bhola 2006). A cut score of 15 points was used on both measures for all cohorts but Cohort Four. Applying a cut score of 15 on Cohort Four removed too many low-ability students; different cut scores of 12 on the test-session specific SOS and 13 on the test-specific SOS were used for this cohort.

We consider a 3-point gain on the test metric, on average, to be moderate. We based this unstandardized average learning gain value on prior quantitative and scientific reasoning studies (e.g. Hathcoat,

Table 1. Ethnicity, age, gender, and SAT data for students collapsing across cohorts.

	Unfiltered	Test- specific filtered	Test session-filtered
American Indian	2.38%	0.68%	0.98%
Asian	4.31%	6.38%	7.44%
Black	2.96%	5.02%	5.87%
Hispanic	3.22%	3.39%	4.70%
Not specified	5.08%	5.43%	2.94%
Pacific Islander	0.39%	1.09%	0.98%
White	84.49%	84.40%	85.91%
Age at pretest	18.44	18.44	18.46
Age at posttest	19.91	19.91	19.92
Female	67.83%	66.49%	65.76%
Male	32.11%	33.51%	34.24%
SAT	1132.42	1132.47	1136.91
N	1554	737	511

Notes: Collapsing across the five cohorts, learning gains were estimated from unfiltered data from 1554 students. Collapsing across four cohorts, prior to filtering, 828 students had complete data on the test-specific SOS and 564 students had complete data on the test session-specific SOS. After filtering for low test-specific motivation, learning gains were computed based on 737 motivated students. After filtering for low test session-specific motivation, learning gains were computed for 511 motivated students.

Sundre, and Johnston 2015), where 3-point average gains on this test were associated with moderate standardised gains (*d* values of approximately 0.40 standard deviations).

We standardised the average unstandardized gains (i.e. Cohen's *d* estimate) using both the standard deviation of pretest scores and the standard deviation of gain scores to compare results to previous studies that used these approaches. Conforming to Cohen's benchmarks and findings from Pastor, Kaliski, and Weiss (2007), we consider a standardised gain of 0.50 on the standardised pretest metric a moderate standardised learning gain. Roohr, Liu, and Liu (2016) considered their standardised mathematics gain estimate of $d = 0.41$ on the standardised gain metric to be moderate; thus, we considered a gain of 0.40 SDs on the standardised gain metric to be moderate. We computed unstandardized and standardised learning gain estimates for each number of courses to assess if gains increased with increased coursework (see Table 2).

Table 2. Descriptive statistics regarding learning gain estimates collapsing across cohort.

Number of courses	0	1	2	3	4	5	6	7	Overall
Unfiltered test scores aggregated across 5 cohorts ($N = 1554$)									
Mean									
Gain score	2.69	3.85	3.51	3.78	4.28	4.38	2.95	2.00	3.72
SD_{gain}	5.58	5.73	5.66	5.58	4.90	4.43	3.03	0.00	5.57
Pretest	44.79	44.66	45.26	44.93	45.65	42.76	42.91	40.00	44.95
$SD_{pretest}$	6.36	6.63	6.57	6.87	6.80	4.39	5.77	0.00	6.67
Posttest	47.48	48.51	48.76	48.71	49.94	47.15	45.86	42.00	48.66
$SD_{posttest}$	7.88	6.99	6.74	6.87	7.04	5.09	4.69	0.00	6.96
Cohen's *d*									
d_{gain}	0.48	0.67	0.62	0.68	0.87	0.99	0.98		0.67
$d_{pretest}$	0.42	0.58	0.53	0.55	0.63	1.00	0.51		0.56
N	62	460	509	310	156	34	22	1	1554
Session-filtered test scores aggregated across 4 cohorts with test-session motivation scores ($N = 511$)									
Mean									
Gain score	1.20	3.27	3.54	3.64	2.82	4.80	3.25		3.35
SD_{gain}	4.73	5.57	4.99	5.41	5.11	4.72	1.77		5.28
Pretest	45.70	46.29	46.68	44.78	46.18	44.00	44.25		46.03
$SD_{pretest}$	5.55	6.31	6.18	6.63	6.23	4.72	5.30		6.37
Posttest	46.90	49.56	50.21	48.42	49.00	48.80	47.50		49.38
$SD_{posttest}$	7.38	6.75	5.94	7.04	6.79	4.77	3.54		6.61
Cohen's *d*									
d_{gain}	0.25	0.59	0.72	0.68	0.55	0.88	2.12		0.63
$d_{pretest}$	0.24	0.52	0.57	0.55	0.45	0.92	0.59		0.53
N	20	140	185	103	49	10	4	0	511
Test-specific filtered test scores aggregated across 4 cohorts with test-specific motivation scores ($N = 737$)									
Mean									
Gain Score	2.14	3.19	3.74	3.23	3.99	4.71	2.75		3.47
SD_{gain}	5.37	5.56	5.32	5.09	4.95	4.58	2.76		5.28
Pretest	44.43	46.48	46.18	45.64	45.94	43.43	43.63		46.01
$SD_{pretest}$	7.45	6.05	6.37	6.48	6.51	4.93	6.09		6.33
Posttest	46.57	49.67	49.92	48.88	49.92	48.14	46.38		49.48
$SD_{posttest}$	10.08	6.17	6.04	6.93	6.32	4.99	6.59		6.43
Cohen's d									
d_{gain}	0.40	0.57	0.70	0.63	0.81	1.03	0.99		0.66
$d_{pretest}$	0.29	0.53	0.59	0.50	0.61	0.96	0.45		0.55
N	21	212	266	138	78	14	8	0	737

Notes: 'SD' indicates standard deviation. 'Gain Score' indicates the difference between the posttest and pretest scores. 'd_{gain}' indicates that Cohen's *d* estimates were computed using the standard deviation of the difference scores; '$d_{pretest}$' indicates that Cohen's *d* estimates were computed using the standard deviation of the pretest scores. '*N*' indicates the number of students who completed the particular coursework. 'Overall' indicates that the values were computed collapsing across quantitative and scientific reasoning coursework. Students could score at most 66 points on the test.

Collapsing across the five cohorts, learning gains were estimated from unfiltered data from 1554 students. Collapsing across four cohorts, prior to filtering, 828 students had complete data on the test-specific SOS and 564 students had complete data on the test session-specific SOS. After filtering for low test-specific motivation, learning gains were computed based on 737 motivated students. After filtering for low test session-specific motivation, learning gains were computed for 511 motivated students.

Using multiple regression, we predicted unfiltered and filtered gain scores to assess if coursework predicted gains after controlling for gender and ability. Gain scores were regressed on number of courses, SAT scores, gender (coded male = 0, female = 1) and their interactions. We mean-centred ability to reduce multicollinearity between ability and interaction terms computed from ability (Aiken, West, and Reno 1991). Assumptions of linearity, normality and homoscedasticity were tested and met.

Participants for faculty discussions of learning gains

Three male and one female quantitative and scientific reasoning general education faculty members participated in this study. To recruit faculty, the first author sent a request for participants who had taught at least one quantitative and scientific reasoning general education course within the past 10 years.

Procedures and materials for faculty discussions of learning gains

The first author interviewed each faculty member in his/her office; interviews lasted no more than 45 min. Prior to the interview, faculty members sat through a five-minute presentation that included example test questions and information regarding how the test was developed to align with quantitative and scientific reasoning student learning outcomes. After this presentation, the first author gave the faculty member a form with several questions aimed at investigating expected learning gains when students completed zero, one, two or three courses (e.g. 'How many points do you *expect* students who have completed one quantitative and scientific reasoning course to gain on the test?'). The faculty then noted desired learning gains for each number of courses completed (e.g. 'How many points *would you like* students who have completed one quantitative and scientific reasoning course to gain on the test?). Faculty were then asked to 'Please *explain why* your *expected* learning gain estimates match or do not match your *desired* learning gain estimates for each of the above questions'. Upon completion of the form, a discussion was held with faculty members about their responses. Faculty were then shown estimated learning gains and asked for their reactions.

Analyses of faculty discussions

We employed an inductive content analysis to analyse interview responses. The inductive approach to content analysis strives to be non-directive in that themes are allowed to evolve from our interaction with the data without forcing them to fit within existing theoretical categories (Hsieh and Shannon 2005). Notes were taken during each interview to record faculty responses. After repeatedly reading faculty responses, codes (brief descriptive categories) were assigned to each line of text. Codes judged as similar were then combined into themes that could be compared across each faculty member. Meetings were held throughout this process to discuss the meaning of participant statements, definitions assigned to each code, and the extent to which assigned codes could be combined to create meaningful themes. Member-checks were conducted by asking interviewees to provide us with feedback about our interpretation of their responses. No faculty member asked us to change our interpretation of their responses.

Results

Hypothesis 1: collapsing across courses, students should have moderate gains

Collapsing across number of courses, students, on average, gained 3.72 points on the 66-item test ($N = 1554$; see Table 2). This gain was statistically significant ($F(1, 1153) = 682.86$, $p < 0.001$) and 31% of the variance in scores could be explained by testing time point. Students gained 0.67 SDs on the

standardised gain metric and 0.56 SDs on the standardised pretest metric. Thus, results supported that, on average, students have moderate gains after experiencing 1.5 years of college.

Hypothesis 2: gains will increase with increased coursework

Contrary to expectations, unfiltered gain scores did not increase with each additional course completed in the domain. Gain scores increased after students completed one quantitative and scientific reasoning course but then levelled off after multiple courses were completed. Specifically, when examining the unfiltered data, students who did not complete any quantitative and scientific reasoning courses gained 2.69 points on the test; students who completed 1 course gained 3.85 points; students who completed 2 courses gained 3.51 points; and students who completed 3 courses gained 3.78 points. Standardised learning gain estimates suggest the same conclusion: there is a gain associated with completing one course, but additional courses in the domain are not associated with a systematic increase.

Hypothesis 3: removing unmotivated students will increase learning gains

After motivation filtering, gain scores did not increase in magnitude as expected. Students in the motivated samples scored higher at pretest than students in the total sample (differences between posttest scores were less pronounced), which led to a minimal decrease in gains. After removing students who were unmotivated during the test battery, the estimated learning gain collapsing across coursework decreased (minimally) to 3.35 points ($N = 511$). Likewise, when we removed students who were unmotivated on the quantitative and scientific reasoning test, this estimate decreased (minimally) to 3.47 points ($N = 737$).

The standardised estimates filtered for low test session-specific motivation (0.63 $SD_{\text{standardized gain metric}}$; 0.53 $SD_{\text{standardized pretest metric}}$) and low test-specific motivation (0.66 $SD_{\text{standardized gain metric}}$; 0.55 $SD_{\text{standardized pretest metric}}$) were essentially identical to the unfiltered standardised estimates (0.67 $SD_{\text{standardized gain metric}}$; 0.56 $SD_{\text{standardized pretest metric}}$). Moreover, as with the unfiltered data, there was an increase in gains after completing one course but additional courses did not produce similar increases in gains.

Hypothesis 4: coursework will predict gains, controlling for personal characteristics

Descriptive statistics discussed thus far suggest coursework is not related to learning gains. To formally test that hypothesis, we predicted learning gains from number of courses, gender, ability and their interactions. First, bivariate correlations indicated that gain scores (filtered and unfiltered) were not significantly or practically correlated with coursework, ability or gender (see Table 3). Second, the predictors as a set did not explain a statistical or practical amount of variance in gain scores (filtered or unfiltered; see Table 4). Similar to the findings of Roohr, Liu, and Liu (2016), personal characteristics did not predict gains. Unfortunately, neither did intentional domain-specific coursework.

Table 3. Correlations among gain scores and potential predictors in the unfiltered and test-specific filtered samples.

	# of Courses		Gender		SAT	
	UF	F	UF	F	UF	F
Gain score	.03	.04	.01	.05	−.03	−.08*
Course			.10*	.11*	−.06	−.11*
Gender					−.21*	−.23*

Notes: To simplify the analyses, we focused on the unfiltered ('UF', $N = 1001$) and test-specific filtered ('F', $N = 689$) data aggregated across the four cohorts with test-specific effort scores (2008–2010, 2013–2015, 2014–2016, 2015–2017).
*indicates significance at $p < 0.05$.

Table 4. Regression results for both the unfiltered and test-specific filtered samples.

	F	df	*p*	R^2	*b*	SE	*t*	*p*	95.0% CI for *b*		*sr*
									LB	UB	
Unfiltered data (*N* = 1001)											
Reduced model	0.61	(3, 997)	0.61	0.002							
Intercept					3.03	0.42	7.22	<0.001	2.21	3.85	
Ability					−0.001	0.001	0.66	0.51	−0.004	0.002	−.02
Gender					0.18	0.38	0.48	0.63	−0.57	0.93	.02
# of courses					0.13	0.14	0.90	0.37	−0.15	0.41	.03
Full model	0.47	(6, 994)	0.83	0.003							
Intercept					3.39	0.64	5.25	<0.001	2.12	4.65	
Ability					0.001	0.004	0.19	0.85	−0.01	0.01	.01
Gender					−0.31	0.78	0.39	0.69	−1.83	1.22	−.01
# of Courses					−0.06	0.29	0.21	0.84	−0.62	0.50	−.01
Gender x # of courses interaction					0.25	0.33	0.73	0.46	−0.41	0.90	.02
Gender x Prior ability interaction					<0.001	0.003	0.05	0.95	−0.01	0.01	−.002
Course x Prior ability interaction					−0.001	0.001	0.56	0.57	−0.003	0.002	−.02
Filtered Data (*N* = 689)											
Reduced model	2.15	(3, 685)	0.09	0.01							
Intercept					2.93	0.50	5.88	<0.001	1.95	3.91	
Prior ability					−0.003	0.002	−1.78	0.08	0.01	0.00	−.07
Gender					0.43	0.44	0.98	0.33	−0.44	1.30	.04
# of courses					0.14	0.17	0.81	0.42	−0.20	0.48	.03
Full model	1.69	(6, 682)	0.12	0.02							
Intercept					3.91	0.77	5.07	<0.001	2.39	5.43	
Ability					−0.001	0.004	−0.24	0.81	−0.01	0.01	−.01
Gender					−0.97	0.94	−1.03	0.30	−2.81	0.87	−.04
# of courses					−0.38	0.34	−1.11	0.27	−1.04	0.29	−.04
Gender x # of courses interaction					0.70	0.40	1.75	0.08	−0.09	1.48	.07
Gender x Prior ability interaction					−0.001	0.004	−0.31	0.76	−0.01	0.01	−.01
Course x Prior ability interaction					−0.001	0.001	−0.37	0.71	−0.003	0.002	−.01

Notes: 'LB': lower bound. 'UB': upper bound. 'sr': semipartial correlation. To simplify analyses, we focused on unfiltered (*N* = 1001) and test-specific filtered (*N* = 689) data aggregated across the four cohorts with test-specific effort scores.

Faculty's expected or desired gains

Prior to presenting the results of the inductive, qualitative content analysis, we first present the differences between faculty member's expected gains and desired gains. Expectations were defined as the number of points on the quantitative and scientific reasoning test that faculty believed students would gain. Desired gains were defined as the number of points on the quantitative and scientific reasoning test faculty would like students to gain.

Faculty members had similar expectations of student learning gains, and their responses suggested an expected relationship between the number of relevant courses completed by the student and learning gains (see Table 5). Faculty expected students to gain approximately 4 points on the test after

Table 5. Empirical learning gain estimates filtered for low test-specific motivation compared to faculty-based estimates, and alignment of expected estimates and desired estimates.

	Faculty One				Faculty Two			Faculty Three			Faculty Four		
	Actual	Expect	Desire	Aligned	Expect	Desire	Aligned	Expect	Desire	Aligned	Expect	Desire	Aligned
0 courses	2.14	2	2	Not aligned	3	21	Not aligned	4	4	Aligned	2–3	?	Aligned
1 course	3.19	4	5		4	21		7	7		3–5	5	
2 courses	3.74	6	9		5	21		10	10		5–7	7	
3 courses	3.23	7	14		5	21		15	15		7–10	10	
Overall	3.47	4	5		4	21		4	4		–	–	

Notes: Values refer to the point-gain on the test for each number of quantitative and scientific reasoning courses. For example, students who did not complete any quantitative and scientific reasoning courses, on average, gained 2.14 points on the 66-item test (after removing students with low test-specific motivation) and students who completed three quantitative and scientific reasoning courses, on average, gained 3.23 points on the 66-item test (after removing those with low test-specific motivation). 'Overall' indicates the average learning gain collapsing across number of courses completed (i.e. after 1.5 years of any college coursework). We asked faculty their expected and desired gains collapsing across number of courses given this is often how assessment data are presented – they are not linked to course completion. Specifically, we asked faculty to answer questions such as 'How many points *would you expect* students who have completed two quantitative and scientific reasoning courses from the general education curriculum to gain on the test?' and 'How many points *would you like* students who have completed two quantitative and scientific reasoning courses from the general education curriculum to gain on the test?' Faculty Four did not provide estimates collapsing across courses. As he explained, it was difficult to produce these estimates without knowing how much relevant coursework students had completed. Faculty Four also did not provide written estimates for students with zero courses because he did not have an opinion on how much these students should gain. 'Aligned' refers to the alignment between faculty's expected and desired gain scores.

1.5 years of any college coursework (see 'Overall' in Table 5). In other words, faculty members expected students to gain 4 points on the test irrespective of the courses that student happened to complete after 1.5 years of college. Gains were expected to increase with each additional course. For example, Faculty One expected students to gain 4 points after completing one course, 6 points after two courses, and 7 points after completing three courses. Faculty One expected *a change* in student learning gains of 2 points from 0 to1 course completed, a change of 2 points from 1 to 2 courses completed, and a change of 1 point from 2 to 3 courses completed. When interpreted this way, on average, faculty reported that each subsequent course should result in approximately 2 additional points on the test (i.e. 0–1 courses: $M_{change} = 1.87$, $SD_{change} = 0.85$; 1–2 courses: $M_{change} = 2.00$, $SD_{change} = 0.82$; 3–4 courses: $M_{change} = 2.13$, $SD_{change} = 2.21$).

Unlike expected gains, faculty members tended to depart in their desired learning gains. Most faculty members reported a relationship between desired learning gains and the completion of coursework. Faculty were divided, however, according to the extent to which their expected gains aligned with desired gains. Two faculty reported their expected gains were lower than what they desired from students; the other two faculty had expected gains that aligned with what they desired from students.

Faculty one interview

When analysing the responses from Faculty One, we derived two themes: Tempered Expectations of Learning and Inconsistent Pedagogical Practices across Courses. This faculty member indicated during his interview that each domain-specific course should contribute to student learning. He explained that he viewed his expected increase of roughly 2 points on the test with each additional course as a low expectation, especially when compared to his desired gains of 3 to 5 points with each additional course (see Table 5). When explaining this discrepancy, he stated that student learning across each course may differ due to inconsistent pedagogical practices. He went on to say, 'I believe the gains will vary across the courses' since instructors have 'different expectations that may affect the progress students make'. That is, students taking a class from Professor A may have greater gains than students taking the same class from Professor B. After reviewing the empirical learning gain estimates, Faculty One wrote that he 'felt the values were fairly small, but not too surprising… very much under-estimated my desired estimations, except those students who had taken 0 courses… possibly due to overlap of these reasoning gains among different courses'.

Faculty two interview

The following two themes were created from the interview with Faculty Two: Unrealized Ideals and Disappointment with Student Learning. Faculty Two, similarly to Faculty One, had what he deemed as low expectations. However, he had high desires for student learning (i.e. desired students to gain 21 points on the test irrespective of coursework). Faculty Two positioned himself as an 'idealist', saying he would 'like for all of the students to answer each item correctly' on the test. He elaborated on this point during the interview when stating, 'I keep the bar high because I think that is where it belongs'. During the interview, he expressed disappointment with how little he perceived students were learning in their courses. When discussing this issue, he provided an anecdote about a statistics course in which students 'cannot explain a *p*-value' after completing the class. Nevertheless, he believed there should be a relationship between gains on the test and coursework within the domain, even if students were not gaining much. After reviewing the empirical gains, Faculty Two reacted similarly to Faculty One, saying that he found 'these changes are too small to be interesting'. When unpacking his feelings regarding the empirical gains, he stated, 'How do I feel? I wish it [courses] makes a difference. I don't consider that a difference'.

Faculty three interview

The following two themes were derived from the interview with Faculty Three: Attainable Gains and Learning from Non-Domain Coursework. During the interview, Faculty Three found it challenging to estimate students' gain scores. When asked about expected learning gains after students completed one

or two courses, she exclaimed 'It's so hard!' She could not explain what made this task difficult. Instead, she discussed how her expectations and desires for student learning gains were aligned given that her estimated gain scores were, in her opinion, 'reasonable to obtain'. Student learning, however, was viewed as a function of non-domain coursework given that quantitative and scientific reasoning skills are 'taught in other general education courses' (e.g. 'Economics'). Despite such factors and the belief that some students may 'simply mature' during college, or have general cognitive gains, Faculty Three stated that there should be a relationship between relevant coursework and the subject-specific learning gains. After reviewing the empirical gains, she declared that she 'must have super high expectations' given the discrepancy between her expectations and the empirical gains. She clarified that she was particularly surprised by the gains of students who completed one or two courses.

Faculty four interview

The following two themes were derived from the interview with Faculty Four: Realistic Expectations Informed by Student Interactions and Uncertain Learning Gains. Faculty Four also expressed difficulty estimating overall learning gains without knowing how many courses students had completed. Despite these difficulties, he explained that his expectations resulted from personal experiences with students: 'My expectations have become more reasonable over time', he said '[but they] would have been higher when I started [teaching]'. He emphasised that students 'do not learn everything they are taught' so it was 'not realistic for students to gain 20 points' on the test. He also indicated that some students may show gains due to development or maturity, though he did indicate that gain scores should increase with relevant coursework. After reviewing the empirical learning gains, Faculty Four said the gains were 'smaller than what I expected. I believed each additional course added gains. Smaller than I desired. Probably comes from perhaps lack of attention to learning objectives in some of the courses'.

Discussion

Given limited research on student learning gains in the US (e.g. Arum and Roksa 2011; Blaich and Wise 2011; Hathcoat, Sundre, and Johnston 2015; Pastor, Kaliski, and Weiss 2007; Roohr, Liu, and Liu 2016), we estimated gains in quantitative and scientific reasoning. Students demonstrated moderate gains after experiencing 1.5 years of college coursework. There was insufficient evidence to suggest such gains can be attributed to intentional coursework designed to increase quantitative and scientific reasoning. Although students experienced some gains after completing a single relevant course, subsequent coursework did not increment learning gains, which was contrary to the views of faculty members who expected larger empirical gain estimates. Importantly, results were consistent when removing unmotivated students from the sample and when controlling for gender and prior ability.

The magnitude of learning gain estimates in higher education within the US

Students at this institution demonstrated greater aggregate learning gains after 1.5 years of college ($d = .56$) than found in prior studies in mathematics (Roohr, Liu, and Liu 2016; $d = .42$ after three years of college) and critical thinking (Arum and Roksa 2011; $d = .18$ after three semesters in college; Blaich and Wise 2011; $d = .44$ after four years of college). The efficacy of coursework completed within the first two years of college had been called into question by Roohr, Liu, and Liu (2016). These researchers found that students with one or two years of college coursework failed to achieve statistically or practically significant learning gains in mathematics. They believed students' lack of acclimation to college after a year or two may have led to this small effect. However, results from this study indicate that moderate learning gain estimates can be obtained irrespective of whether students are acclimated to the college culture.

Improved sampling techniques in the current study may account for the incongruity in findings between our study and those by Roohr, Liu, and Liu (2016), Arum and Roksa (2011) and Blaich and Wise (2011). At this institution, a large number of students were randomly assigned to complete the quantitative and scientific reasoning test. In comparison, Roohr, Liu, and Liu (2016) obtained estimates

from a small, conveniently sampled group of students. Arum and Roksa (2011) and Blaich and Wise (2011) gathered data from institutions that employed different sampling and retention strategies, with an overall rate of retention being less than 50% (Arum and Roksa 2011) or 70% (2006 cohort; Blaich and Wise 2011) across the many institutions.

In sum, students appear to be learning in college, though this learning cannot be attributed to intentional coursework designed to increase their knowledge and skills. Our results simply indicate the extent students are gaining, which can be evaluated against faculty expectations and desires. Aggregate gain estimates alone provide limited information for learning improvement initiatives.

Understanding the nonexistent relationship between learning gains and coursework

Because learning gains did not increase with additional coursework, one may question the quality of the data. To investigate this concern, we removed students with low test-taking motivation, re-estimated learning gains, and re-predicted these gains from theorised variables. Gain estimates were unaffected by motivation filtering; coursework continued to be unrelated to gains. Faculty and administrators should not assume this nonexistent relationship between learning gains and coursework was biased due to low test-taking motivation or confounding variables such as gender and ability; these predictors did not moderate the learning gains, which replicates the findings of Roohr and colleagues (2016).

Obviously, the quasi-experimental nature of any study examining the relationship between coursework and learning gain must be acknowledged. Students decided to either complete or not complete the quantitative and scientific reasoning courses based on interests or academic schedules during the first 1.5 years in college. As students were not randomly assigned to the number of quantitative courses completed, causal statements regarding curriculum impact must be avoided. The quasi-experimental design reflects the realities of much educational research on learning gains. Given these limitations, our results do not necessarily imply that college fails to add value. As noted by Pascarella et al. (2011, 24):

> Conversely (and this seems counter-intuitive), little or no gain during college does not mean that college is failing to add value. On some traits, such as quantitative skills, students do not always appear to progress much during college, but their counterparts who do not attend college actually retrogress substantially over the same period of time (Pascarella and Terenzini 1991).

However, we cannot test this hypothesis without random assignment, which is impractical.

Instead, we turned to faculty who teach these courses to uncover additional explanations regarding the learning gains. Some faculty perceived that students learn in college from experiences outside of intentional coursework. Faculty Three indicated that students may also gain quantitative and scientific reasoning skills from classes that do not fulfil their general education requirements. Although students might learn quantitative and scientific reasoning skills in such courses, it is questionable whether these gains would be of sufficient magnitude to influence the relationship between student learning gains and number of quantitative and scientific reasoning courses. Faculty Three also suggested these gains may be due to maturation, which seems plausible. Students without relevant coursework demonstrated some quantitative and scientific reasoning gains. However, students who completed at least one quantitative or scientific reasoning course demonstrated greater gains than these students. If the gains were solely due to maturation, we would expect that learning gain estimates would be more similar across the two groups.

Faculty One and Four mentioned a second, related explanation for the unexpectedly low learning gains and nonexistent relationship between gains and coursework: *implementation fidelity*, or whether the curriculum aligned with the objectives and test is actually taught and received in the intended manner (Gerstner and Finney 2013). If implementation fidelity is low, students may have varying learning gains due to differences in quality of instruction and curriculum (Finney and Smith 2016). Evaluating instruction is critical for higher education assessment reformation, as Coates (2016, 669) notes:

> …inasmuch as academic autonomy, in its various encapsulations, provides faculty with a presumption of private ownership over academic work it can be a significant impediment to change. Research proposals and papers undergo peer review, and there is no reason why teaching, engagement and leadership should not as well.

Thus, assessing implementation fidelity may be necessary to establish if students are receiving the intended curriculum, to pinpoint areas of weakness, and to (re)train faculty in the domains of cognition and learning (Gerstner and Finney 2013).

Finally, the coursework may simply not be effective in meeting the specified student learning objectives. Faculty Two alluded to this hypothesis during his interview, describing how students he encountered could not 'explain a *p*-value' after successfully completing statistics coursework intentionally designed to enhance statistics learning outcomes. If students are not learning after they receive the intended curriculum, then it may be time to change the curriculum and/or pedagogy to foster students meeting these outcomes deemed important to the university. However, modifications to the general education curriculum require a large-scale learning improvement initiative.

Implications for learning improvement

Given that few exemplars of learning improvement exist, how can the faculty in this study begin the process of demonstrating learning improvement? First, faculty should come to a consensus on the magnitude of desired learning gains. This is not an easy task. This study has increased our concerns about how learning gain estimates are reported and interpreted in the literature. Most researchers report and interpret standardised estimates for ease of comparisons with other studies as well as convention. Solely interpreting standardised estimates does not provide a clear or accurate depiction of student learning gains. We aligned our unstandardized gain score benchmark (i.e. 3-point gain) with Cohen's arbitrary but widely-used effect sizes (Cohen 1992). Without interviewing faculty, we concluded that students demonstrated 'moderate' learning gains. Faculty, however, were disappointed with the magnitude of learning gains; learning gains deemed by faculty as 'low' or 'attainable' were not being attained after students completed coursework. Therefore, interpreting results on the *test* (i.e. unstandardized) metric provides stakeholders with a clearer understanding of student learning gain and facilitates comparisons with their expectations of learning. Without interpretable gains to compare to expectations, faculty may be unlikely to engage in curriculum enhancement to improve student learning.

Second, intentional curricula that should result in expected (and desired) learning gains should be (re)designed by the faculty. On-campus support from experts in cognition and learning may be necessary (e.g. Lewis 2010). Through use of implementation fidelity assessment, the delivered curriculum should then be examined to evaluate if it aligns with the designed curriculum (Smith, Finney, and Fulcher 2017). Poor implementation fidelity should be acknowledged and remedied to accurately assess the designed curriculum. Learning gains associated with this modified curriculum should then be estimated and disaggregated by completed coursework. The faculty can compare the learning gains computed from this study to those learning gains from the cohort who experiences the modified curriculum.

These procedures align with what may be considered best practices for student learning outcomes assessment. Banta and Blaich (2011) explicitly discussed the importance of involving faculty during student learning outcomes assessment when stating:

> If faculty do not participate in making sense of and interpreting assessment evidence, they are much more likely to focus solely on finding fault with the conclusions than on considering ways that the evidence might be related to their teaching. (24)

However, faculty often do not receive assistance on how to use assessment results to improve student learning (Fulcher et al. 2014). At the most basic level, using assessment results requires faculty to implement modifications to pedagogy or curriculum. Thus, as noted by Brown (2017, 3): 'Testing and measurement need to integrate with classroom teaching, learning, and curriculum if it is to support schooling…'.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Catherine E. Mathers is a doctoral student in the Educational Measurement and Statistics program at the University of Iowa. While pursuing her master's at James Madison University under the supervision of Dr. Sara Finney, her work focused on higher education assessment, as well as measurement of test-taking motivation and its effects on test performance. Currently, her research interests include test development and applications of item response theory.

Sara J. Finney is a professor of Graduate Psychology at James Madison University where she teaches courses in the areas of Multivariate Statistics and Structural Equation Modeling, and she advises graduate students in the Assessment & Measurement PhD program and the Quantitative Psychology Concentration within the Psychological Sciences MA program. Sara's research involves the application of latent variable modelling techniques to investigate questions related to college student development, examinee motivation and the functioning of self-report instruments. Sara is also the associate director of the Center for Assessment and Research Studies at JMU. In this role, she works with faculty and staff to design assessment efforts that contribute to making empirically based decisions about programme effectiveness and ultimately student learning and development. Sara serves on the editorial boards of the Journal of Educational Psychology, Educational and Psychological Measurement, Educational Assessment, International Journal of Testing and the Journal of Experimental Education.

John D. Hathcoat is an assistant professor in Graduate Psychology and associate director of University Learning Outcomes Assessment in the Center for Assessment and Research Studies at James Madison University. John has taught graduate-level courses in educational statistics, research methods, measurement theory and performance assessment. His research focuses on validity theory, instrument development and performance-based assessment in higher education.

ORCID

Catherine E. Mathers http://orcid.org/0000-0002-8959-0013

References

Aiken, L. S., S. G. West, and R. R. Reno. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Thousand Oaks, CA: Sage Publications.

American Association for Higher Education. 1992. *Nine Principles of Good Practice for Assessing Student Learning*. http://www.learningoutcomesassessment.org/PrinciplesofAssessment.html#AAHE

Arum, R., and J. Roksa. 2011. *Academically Adrift: Limited Learning on College Campuses*. Chicago, IL: University of Chicago Press.

Arum, R., and J. Roksa. 2014. *Aspiring Adults Adrift: Tentative Transitions of College Graduates*. Chicago, IL: University of Chicago Press.

Banta, T. W., and C. Blaich. 2011. "Closing the Assessment Loop." *Change: The Magazine of Higher Learning* 43: 22–27.

Blaich, C., and K. Wise. 2011. "The Wabash National Study: The Impact of Teaching Practices and Institutional Conditions on Student Growth." Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April.

Borden, V. M. H., and S. R. Peters. 2014. "Faculty Engagement in Learning Outcomes Assessment." In *Higher Education Learning Outcomes Assessment: International Perspectives for Quality Improvement*, edited by H. Coates, 201–212. Frankfurt: Peter Lang.

Brown, G. T. L. 2017. "The Future of Assessment as a Human and Social Endeavor: Addressing the Inconvenient Truth of Error." *Frontiers in Education* 2(3). https://www.frontiersin.org/articles/10.3389/feduc.2017.00003/full

Castellano, K. E., and A. D. Ho. 2013. *A Practitioner's Guide to Growth Models*. Washington, DC: Council of Chief State School Officers.

Coates, H., ed. 2014. *Higher Education Learning Outcomes Assessment: International Perspectives*. Frankfurt: Peter Lang.

Coates, H. 2016. "Assessing Student Learning Outcomes Internationally: Insights and Frontiers." *Assessment & Evaluation in Higher Education* 41: 662–676.

Cohen, J. 1992. "A Power Primer." *Psychological Bulletin* 112: 155–159.

Creswell, J. W., and V. L. Plano Clark. 2011. *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage.

Dorans, N. J. 1999. *Correspondence between ACT and SAT I Scores* (College Board Report 99-1). New York: College Entrance Examination Board.

Ewell, P. 1983. *Information on Student Outcomes: How to Get It and How to Use It*. Boulder, CO: National Center for Higher Education Management Systems.

Ewell, P. 1985. *Transformation Leadership for Improving Student Outcomes*. NCHEMS Monograph No. 6. Boulder, CO: National Center for Higher Education Management Systems.

Ewell, P. 2004. *General Education and the Assessment Reform Agenda*. Washington, DC: Association of American Colleges and Universities.
Finney, S. J., and K. L. Smith. 2016. "Ignorance is Not Bliss: Implementation Fidelity and Learning Improvement." *National Institute for Learning Outcomes Assessment: Guest Viewpoints*, January 12, 2016. https://illinois.edu/blog/view/915/309716.
Finney, S. J., C. E. Mathers, and A. J. Myers. 2016. "Investigating the Dimensionality of Examinee Motivation across Instruction Conditions in Low-stakes Testing Contexts." *Research & Practice in Assessment* 11: 5–17 http://www.rpajournal.com/dev/wpcontent/uploads/2016/07/A1_Corrected.pdf.
Finney, S. J., D. L. Sundre, M. S. Swain, and L. M. Williams. 2016. "The Validity of Value-added Estimates from Low-stakes Testing Contexts: The Impact of Change in Test-taking Motivation and Test Consequences." *Educational Assessment* 21: 60–87.
Finney, S. J., A. J. Myers, and C. E. Mathers. Forthcoming. Test Instructions Do Not Moderate the Indirect Effect of Perceived Test Importance on Test Performance in Low-stakes Testing Contexts. *International Journal of Testing*.
Fulcher, K. H., M. R. Good, C. M. Coleman, and K. L. Smith. 2014. *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, December.
Gerstner, J. J., and S. J. Finney. 2013. "Measuring the Implementation Fidelity of Student Affairs Programs: A Critical Component of the Outcomes Assessment Cycle." *Research & Practice in Assessment* 8: 15–28.
Hathcoat, J. D., D. L. Sundre, and M. M. Johnston. 2015. "Assessing College Students' Quantitative and Scientific Reasoning: The James Madison University Story." *Numeracy* 18: 1–19.
Hsieh, H.-F., and S. E. Shannon. 2005. "Three Approaches to Qualitative Content Analysis." *Qualitative Health Research* 15: 1277–1288.
Kuh, G. D. 2009. "The National Survey of Student Engagement: Conceptual and Empirical Foundations." *New Directions for Institutional Research* 141: 5–20.
Lewis, K. G. 2010. "Pathways toward Improving Teaching and Learning in Higher Education: International Context and Background." *New Directions for Teaching and Learning* 122: 13–23.
Liu, O. L. 2011. "Value-added Assessment in Higher Education: A Comparison of Two Methods." *Higher Education* 61: 445–461.
Musekamp, F., and J. Pearce. 2016. "Student Motivation in Low-stakes Assessment Contexts: An Exploratory Analysis in Engineering Mechanics." *Assessment & Evaluation in Higher Education* 41: 750–769.
Pascarella, E. T., and C. Blaich. 2013. "Lessons from the Wabash National Study of Liberal Arts Education." *Change: The Magazine of Higher Learning* 45: 6–15.
Pascarella, E. T., and P. Terenzini. 1991. *How College Affects Students*. San Francisco, CA: Jossey-Bass.
Pascarella, E. T., C. Blaich, G. L. Martin, and J. M. Hanson. 2011. "How Robust Are the Findings of Academically Adrift?" *Change: The Magazine of Higher Learning* 43: 20–24.
Pastor, D. A., P. K. Kaliski, and B. A. Weiss. 2007. "Examining College Students' Gains in General Education." *Research & Practice in Assessment* 2: 4–17.
Roohr, K. C., H. Liu, and O. L. Liu. 2016. "Investigating Student Learning Gains in College: A Longitudinal Study." *Studies in Higher Education* 42: 2284–2300.
Sessoms, J. C., and S. J. Finney. 2015. "Measuring and Modeling Change in Examinee Effort on Low-stakes Tests across Testing Occasions." *International Journal of Testing* 15: 356–388.
Smith, K. L., S. J. Finney, and K. H. Fulcher. 2017. "Actionable Steps for Engaging Assessment Practitioners and Faculty in Implementation Fidelity Research." *Research & Practice in Assessment* 12: 71–86.
Sundre, D. L., A. Thelk, and C. Wigtil. 2008. *The Natural World Test, Version 9: A Measure of Quantitative and Scientific Reasoning, Test Manual*. Harrisonburg, VA: James Madison University.
Swerdzewski, P. J., J. C. Harmes, and S. J. Finney. 2009. "Skipping the Test: Using Empirical Evidence to Inform Policy Related to Students Who Avoid Taking Low-stakes Assessments in College." *Journal of General Education* 58: 167–195.
Thelk, A. D., D. L. Sundre, S. J. Horst, and S. J. Finney. 2009. "Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences about Student Performance." *Journal of General Education* 58: 129–151.
U.S. Department of Education. 2006. *A Test of Leadership: Charting the Future of U.S. Higher Education*. Washington, DC.
Wise, S. L., and C. E. DeMars. 2010. "Examinee Noneffort and the Validity of Program Assessment Results." *Educational Assessment* 15: 27–41.
Wise, S. L., and L. F. Smith. 2016. "The Validity of Assessment When Students Don't Give Good Effort." In *Handbook of Human and Social Conditions in Assessment*, edited by G. T. L. Brown and L. R. Harris, 204–220. New York, NY: Routledge.
Wise, V. L., S. L. Wise, and D. S. Bhola. 2006. "The Generalizability of Motivation Filtering in Improving Test Score Validity." *Educational Assessment* 11: 65–83.
Zlatkin-Troitschanskaia, O., H. A. Pant, and H. Coates. 2016. "Assessing Student Learning Outcomes in Higher Education: Challenges and International Perspectives." *Assessment & Evaluation in Higher Education* 41: 655–661.