Numeracy
Advancing Education in Quantitative Literacy

Volume 12 | Issue 2 Article 8

2019

Using Meta-Analysis to Assess Affective Outcomes in a Multi-Course QR Module Intervention

James Friedrich
Willamette University, jfriedri@willamette.edu
Kelley D. Strawn
Willamette University, kstrawn@willamette.edu

Recommended Citation

Friedrich, James, and Kelley D. Strawn. "Using Meta-Analysis to Assess Affective Outcomes in a Multi-Course QR Module Intervention." *Numeracy* 12, Iss. 2 (2019): Article 8. DOI: https://doi.org/10.5038/1936-4660.12.2.8

Using Meta-Analysis to Assess Affective Outcomes in a Multi-Course QR Module Intervention

Abstract

When quantitative reasoning(QR) interventions share a common hypothesis or goal, a promising approach for evaluation involves integrating separate analyses through the use of meta-analysis. This paper reports an assessment of a module-based QR intervention distributed across 20 courses at a single institution. Topics and participating courses were diverse, including arts & humanities, quantitative behavioral sciences, and natural sciences & mathematics groupings, but all addressed the shared affective goals of reducing student QR self-doubt and increasing appreciation for QR value and utility. With a local framework to guide module development, we assess these outcomes using reliable self-report measures in a pre-post design for each course. Random effects meta-analysis for self-doubt outcomes reveals significant moderation by course grouping, with significant but modest-sized reductions for arts & humanities (M_d = -0.27, $CI_{95\%}$[-0.45, -0.08]) and quantitative behavioral sciences (M_d = -0.24, $CI_{95\%}$ [-0.47, -0.01]) but not for natural sciences & mathematics (M_d = 0.13, $CI_{95\%}$[-0.06, 0.32]). Analysis of perceived utility outcomes reveals a significant overall increase without moderation, but again with a pattern of significant change for the arts & humanities (M_d = 0.47, $CI_{95\%}$[0.11, 0.84]) and quantitative behavioral sciences (M_d = 0.29, $CI_{95\%}$[0.02, 0.55]) but not for natural sciences & mathematics (M_d = 0.12, $CI_{95\%}$ [-0.18, 0.42]). Overall, the meta-analyses reveals expected patterns that would have gone undetected in the underpowered (small *N*) individual course implementations. We discuss strengths and limitations of meta-analytic approaches to QR assessment, along with the potential value of such aggregated information for researchers, individual instructors, and institutions.

Keywords

meta-analysis, modules, self-doubt, perceived utility

Cover Page Footnote

James Friedrich is a Professor of Psychology at Willamette University and co-director of its undergraduate Quantitative Understanding, Analysis, and Design (QUAD) Center. His interests focus on biases in judgment and quantitative reasoning. His work also addresses recent reform recommendations for data analysis, including how they are incorporated into the undergraduate curriculum.

Kelley D. Strawn is an Associate Professor of Sociology and Faculty Associate Dean of Curriculum at Willamette University and co-director of the QUAD Center. His interests focus on collective action and protest in Mexico and Latin America, as well as on trends in religious belief and behavior in the United States.

Introduction

Researchers contributing to this journal and related outlets develop a rich base of information on quantitative reasoning and quantitative literacy (QR/QL) interventions, including data on potential outcome measures and evidence of program effectiveness and impact. Many of these demonstrations necessarily occur in quite specific local environments, with published assessments often based on carefully scripted and controlled interventions implemented in single courses, departments, or institutions. Collecting multi-replication and multi-institutional data from diverse samples (e.g., Sundre and Thelk 2010; Gaze et al. 2014; Follette et al. 2017) is one of several ways in which scholars establish the reliability and robustness of their findings and build confidence among prospective adopters.

Adopters frequently make subtle or even significant modifications when adapting interventions to new circumstances. Local conditions can vary in terms of intervention content, levels of instructor training and commitment, broader institutional requirements and culture, and between-course or between-institution differences in content, student populations and resources. These factors all raise the possibility that the success of evidence-based interventions might not always replicate. As a result, there is an important role for local, "adopter-specific" validation studies, as well as for aggregating information across diverse implementations. Demonstrated local effectiveness can be especially useful for motivating faculty to participate in and support QR/QL curricula. Moreover, such data can contribute to a larger body of evidence when findings are synthesized across multiple programs or institutions.

A methodological and analytical approach that holds particular promise in this regard is the use of meta-analysis (Borenstein et al. 2009; Cumming 2012). This technique allows multiple interventions within a single institution, or replications across multiple institutions, to be evaluated by integrating them into an overall analysis. In doing so, the analysis determines the average size of effects, explores their degree of consistency across replications, and identifies potential moderator variables that might explain variation in outcomes.

Traditional null hypothesis significance testing tends to focus on dichotomous "yes/no" or "works/doesn't work" decisions for isolated studies (Cumming 2014). Yet this approach is particularly problematic for assessing effectiveness when, as is often the case, local adaptations have small sample sizes and thus low power. Meta-analysis instead shifts attention toward the magnitude and consistency of effects while also highlighting the role that factors such as sampling variability and statistical power might play in the apparent success or failure of individual replications (Maxwell et al. 2015).

Meta-analysis is already well-established as a method of retrospectively synthesizing existing literature on a topic. However, increasing attention is being paid to its prospective use — designing and assessing studies with the intention of a final meta-analytic synthesis (e.g., Open Science Collaboration 2015). Our goal in this paper is to introduce such a prospective use of meta-analysis in QR/QL assessment. Specifically, we illustrate a meta-analytic evaluation of a single-institution project consisting of "module" interventions implemented in 20 different courses at our small undergraduate liberal arts college.

We begin by describing our module program in more detail, including two new outcome measures serving as the primary focus of the assessment. Next, we discuss the potential benefits of a meta-analytic approach to data integration, emphasizing how this strategy might address significant research challenges experienced both within and between institutions. Following this step, we report the actual meta-analytic assessment of our QR module data and discuss implications for broader use of this technique in intervention development and program evaluation.

QR Module Interventions and Goals

A common theme throughout the literature on quantitative literacy pedagogy is the importance of distributing quantitative reasoning opportunities across the curriculum (Bressoud 2009; Hillyard 2012). Ideally, these encounters occur both at the introductory level and at more advanced levels through general education programs and through disciplinary experiences within majors. Educators have specifically explored the potential benefits of smaller, targeted QR experiences embedded within diverse offerings. Whether in the form of distributed topics and activities over a term or through more narrowly circumscribed assignments or "modules" (cf. Wenner et al. 2009; Steele and Kilic-Bahi 2010; Vachar and Lardner 2010), these targeted activities are designed to provide highly contextualized opportunities for students to develop QR skills and to appreciate their value in answering questions of interest.

One advantage of module approaches is that they can be adapted to a broad range of course structures and content; they represent critical learning opportunities in which empirical and numerical analysis help illuminate and/or apply the subject matter, whatever it might be. A related advantage is that such flexibility creates a more inviting framework for participating faculty than one might achieve through a specific, fixed curriculum that would have to be adopted by all instructors.

Focusing on this approach, and with grant support from the Teagle Foundation, Willamette University agreed to pilot 20 different course modules to infuse aspects of quantitative reasoning across its undergraduate curriculum. The university's general education program already included a two-course

requirement in the area of quantitative and analytical reasoning—a requirement relying heavily on traditional course offerings in mathematics and statistics. The module offerings were construed as something to complement and augment such experiences by providing additional contextualized, "modest-dosage" experiences across the curriculum.

Particular courses and participating instructors were not selected prior to our grant submission, but the proposal committed to having roughly half of the 20 modules be developed in courses for which QR activities were not typically central aspects of the curriculum. For planning purposes, instructors consulted with the campus Quantitative Understanding, Analysis, and Design (QUAD) Center — a statistics and research-methods support center that serves students and faculty at all levels from general education to senior thesis and publication-oriented work. Participating faculty self-referred and/or were invited by the first author (the QUAD Center director at the time), with an eye toward broad disciplinary and course-level representation. Instructors chose which one of their courses to include, with projects designed to complement and supplement existing syllabi rather than fundamentally alter course structure.

Modules ranged from targeted projects that could be completed over the course of a week or two to projects spread over the entire term.[1] For example, a history course that normally includes original writings of the Roman architect Vitruvius developed a project in which students receive a guest lecture on understanding ratios and proportions. Readings and instruction highlighted their role in Greek and Roman notions of ideal structures. Students then had to construct a "blueprint" for their own temple design that illustrated the application of the Vitruvius readings and the relevant quantitative principles (cf. Diefenderfer 2012). A QUAD Center undergraduate assistant trained in the relevant course readings worked with students to help them understand the quantitative aspects of the readings and the design assignment. In contrast, for an introductory biology course, the QUAD Center helped develop custom software tutorial videos to aid students with analysis and formal reporting skills for several existing lab projects spread over the course of the term.

Considerable debate and overlapping terminology regarding the constructs of quantitative reasoning (QR) and quantitative literacy (QL) are found in the literature (Karaali et al. 2016). Drawing on the long history and diversity of perspectives in the field, the QUAD Center attempted to develop a local framework tied to our institutional concerns that would have sufficient detail to guide discussion of possible projects for modules. The core elements—

[1]A brief description of the 20 interventions is available from the authors.

conceptualizing quantitative literacy as *conducting*, *deploying*, and *embracing* quantitative reasoning—are organized according to what we came to refer to by the "$C^3D^3E^3$" mnemonic. Our working definitions and the associated QR/QL framework are described in Table 1.

Table 1
A "Conducting, Deploying, & Embracing Quantitative Reasoning" ($C^3D^3E^3$) Module Design Framework

Working definitions:

"Quantitative reasoning (QR) is the process of solving problems, drawing valid inferences, and understanding, formulating, and disseminating appropriate arguments based on information subjected to quantitative analysis."

"A quantitatively literate individual is someone who has the necessary skills to successfully engage in QR, is inclined to do so in relevant contexts, and values QR work appropriately vetted by others."

Quantitatively literate individuals should be able to CONDUCT quantitative analyses by:

- COLLECTING valid data (new or from existing sources) amenable to quantitative analysis
- COMPUTING informative quantities based on data
- CRITIQUING and interpreting numerical, graphical, and verbal representations of results

Quantitatively literate individuals should be able to DEPLOY quantitative analyses by:

- DEBATING or defending positions using such evidence in persuasive arguments
- DISSEMINATING quantitative findings, interpretations, and arguments clearly
- DESIGNING better methods and new questions based on quantitative insights

Quantitatively literate individuals should EMBRACE quantitative analysis through:

- EFFICACY or a sense of self-confidence in dealing with quantitative information
- ENGAGEMENT and use of C^3D^3 skills in spontaneous ways when appropriate
- ENDORSEMENT of quantitative reasoning, recognizing its value and its limitations

Note: This framework served as a heuristic tool for meeting with instructors and developing modules. Module elements focused on individual instructor preferences and emphasized only selected aspects of the framework. Assessment measures for the research discussed here focused on the efficacy and endorsement elements.

Modules did not need to address all aspects of the framework; the aim was to target whatever element or elements individual instructors felt best fit with their curricular and disciplinary goals. This decision meant that specific skills and outcomes would vary widely across classes and would be evaluated through normal graded assignments. The common elements across all modules, however, were the affective goals tied to having students embrace quantitative reasoning more broadly (cf. Wilkins 2000; Rheinlander and Wallace 2011). Specifically, regardless of content or skill focus, the goals of each module were to help students (1) develop a stronger sense of subjective comfort or positive affect when contemplating quantitative issues and (2) better appreciate the potential value of

quantitative analysis generally. These goals were specifically linked to the "efficacy" and "endorsement" elements in our $C^3D^3E^3$ framework, with outcomes assessed through a pair of locally validated self-report measures administered early and late in each course as part of a pre-post design.[2] The interventions focused on these belief elements within our notion of *embracing* QR but did not attempt to assess the more behavioral "engagement" element, which—given our framework—would likely have involved novel post-intervention QR performance measures that were somewhat idiosyncratic to specific module topics and not yet validated.

A Meta-analytic Approach to Assessment

Although we had the benefit of using the same outcome measures in each module, it is often the case that QR initiatives differ both in the interventions themselves and the outcome measures used. For example, independent researchers interested in the impact of software usage on math anxiety might use one type of homework and software in a calculus course but a different type of homework and software—and perhaps a more focused measure of statistics anxiety—in a statistics course. Such projects might be carried out at the same institution or at different ones but nevertheless share an underlying research question that could be addressed by combining separate study effects rather than by pooling individual level data.

Meta-analysis is an analytic technique designed for such circumstances, using the results or "effects" found in some set of studies as the raw data for additional analysis, rather than pooling the original data from those studies and reanalyzing it collectively. The basic logic behind meta-analysis is that tests of the same or similar hypotheses can yield effect sizes of different magnitudes for a host of reasons (Borenstein et al. 2009; Cumming 2012). Individual studies test distinct participant populations with samples of varying size and under conditions that differ in numerous ways. Thus, although some of the variation in effect sizes from study to study might reflect normal sampling variability, some variation might also reflect differences due to specifiable conditions or study features—attributes that can be coded and incorporated into the meta-analysis as moderator variables. In estimating overall average effect sizes, confidence intervals around those estimates, and differences between average effect sizes for different categories of studies, a meta-analysis shifts the focus toward the magnitude of effects and not

[2] Our goal was to encourage wide participation by individual instructors as part of an institutional intervention rather than to conduct experimental investigations of each module. Because instructors were typically teaching only one section of the relevant course, we did not pursue a non-equivalent control design for each class. The usual limitations of simple pre-post designs do constrain our conclusions, and we take up this issue in the discussion section.

just their "significance," while also drawing attention to the replicability and generality of findings.

Meta-analysis has become a widely embraced technique for integrating a diverse, existing research literature on a given topic or hypothesis, including fields as varied as biomedical research, ecology, education, psychology, criminology, and business (Borenstein et al. 2009). In most cases, meta-analysts carefully search the published and unpublished literature based upon their inclusion criteria, capitalizing on whatever work happens to have been completed by interested parties and made available for analysis. The analysts do not determine what studies are done but instead gather those appearing relevant and then code them for a range of attributes that might account for differences in reported effects.

What is perhaps less commonly appreciated, however, is the fact that meta-analysis can be considered prospectively as a framework for integrating a planned—as opposed to merely "available"—set of studies (e.g., Klein et al. 2014; Open Science Collaboration 2015). For example, Cumming (2014) and Maner (2014) have noted the potential benefits of conducting a meta-analysis on the different studies that might appear within a single, multi-study/replication paper. Such approaches, sometimes referred to as mini meta-analyses or internal meta-analyses (Goh et al. 2016; Ueno et al. 2016), provide a means of obtaining estimates of effects less subject to single-study sampling variability. Indeed, a set of studies which find similar but statistically nonsignificant effects can yield an unambiguous and statistically significant finding when integrated meta-analytically.

It is this application of meta-analysis to a planned set of studies—within a single school or across multiple schools and sites—that this paper seeks to illustrate as a tool for QR researchers. In a sense, all the relevant studies for inclusion are known and need not be retrieved from a broader and often unpublished, difficult-to-access literature. This practice is possible even when the studies are based on relatively small-*N*, single-class interventions and include differences (e.g., in course subject matter or syllabi) that would preclude an analysis that simply collapsed raw data into a single, larger data set.

Meta-analysis is not a panacea, and we discuss certain limitations in more detail in the context of our reported module intervention findings. In general, however, a meta-analysis reflects the limitations of the constituent studies. For example, aggregating a large number of pre-post studies does not alter the limitation inherent in the lack of randomly assigned control groups. Moreover, to the extent that the moderator variables explored are non-randomized features of the studies themselves (e.g., the disciplines of participating classes), then any evidence of differing effect sizes for different groupings of studies would have to be interpreted in correlational terms.

Despite such limitations, meta-analysis is a technique well-suited to evaluating certain outcomes for QR interventions. To illustrate its potential value, we report here the meta-analysis we conducted for our 20-course QR module intervention. We begin with a brief description of the design and the psychometric characteristics of the mathematics self-doubt and perceived utility measures used to assess outcomes. We then report results by first addressing the possibility of participant attrition effects in our pre-post design. Next we report on the reliability evidence for outcome measures in the present samples. Finally we report on the meta-analytic findings themselves, including a moderator variable analysis based on a broad disciplinary classification of the participating courses.

Method

Participants and Course Implementation

Undergraduate liberal arts faculty members at our small, private university developed new QR modules in their courses. From three to five new one-term offerings were developed and implemented each semester over a period of five semesters for a total of 20 interventions. Final samples for each of the course/QR module interventions reflected students who completed both the pre- and posttest measures of QR-related beliefs. These *n*'s ranged from 5 to 35, yielding a combined total of $N = 349$. Additional cases in each class were excluded because matchable posttest scores were not available for some students ("non-completers"). In the Results section, we report data comparing completers and non-completers to assess the possibility of selective attrition effects.

As noted above, participating instructors worked with the campus QUAD Center to develop a QR intervention that complemented their existing courses. The specific content and timeline of the projects/modules varied based on course topic and instructor interest, but all reflected the primary goals of embedding QR into the courses in new and interesting ways that might reduce mathematics self-doubt and enhance the perceived usefulness of quantitative analysis.

Concerns specific to conducting internal or "mini" meta-analyses include the possibility that studies might be selectively omitted or that new studies might be added only until a particular statistical result is achieved (Goh et al. 2016; Ueno et al. 2016). In the present analysis, all 20 interventions funded by the grant were included, without regard to individual or cumulative effect sizes. Only one instructor participated more than once, but in topically distinct offerings.

Participating courses included classes at all levels and were distributed across arts & humanities (9), quantitative behavioral sciences (5), and natural sciences & mathematics (6). This classification scheme was developed through independent, program-specific criteria based on the course content and the home department's curriculum. For example, all courses in the latter two categories were offered and

taught by faculty within departments that have formal mathematics/statistics and quantitative design requirements for majors; the other 9 courses came from major programs structured without any such requirement. Classification, department, course level, and sample size for each included course is reported in Table 2.

Table 2
Participating Courses and Sample Sizes by Category and Department

Arts & Humanities	*Quantitative Behavioral Sciences*	*Natural Sciences & Mathematics*
Anthropology 2XX (*n* = 14)	Economics 3XX (*n* = 15)	Biology 1XXA (*n* = 22)
Art History 2XX (*n* = 15)	Psychology 2XX (*n* = 28)	Biology 1XXB (*n* = 35)
English 3XX (*n* = 14)	Psychology 3XX (*n* = 24)	Biology 2XX (*n* = 17)
History 3XXA (*n* = 11)	Psychology 4XX (*n* = 5–6)	Exercise Science 1XX (*n* = 12)
History 3XXB (*n* = 14)	Sociology 2XX (*n* = 15)	Exercise Science 2XX (*n* = 7–8)
History 3XXC (*n* = 14)		Mathematics 1XX (*n* = 35)
Politics 2XX (*n* = 28)		
Politics 3XXA (*n* = 12–13)		
Politics 3XXB (*n* = 10)		

Note: Sample sizes reflect only the cases for which students completed measures at both pretest and posttest. Ranges are given in the few cases where students did not have complete data for one of the two outcome measures. Course numbers are masked to protect instructor anonymity, but leading digit levels reflect the imprecise local numbering system: 100–200 designations for introductory/early exposure courses, 200–300 designations for intermediate level (often with mixes of major and non-major students), and 400 designations for classes taken primarily by junior/senior majors. Appended letters are used to indicate different courses at the same numerical level.

Affective Outcome Measures

There were no specific, objective quantitative skills common to or measured across all modules. Instead, consistent with our goals, our assessment relied on two locally developed and pretested self-report measures of general comfort/discomfort with thinking numerically and belief in the utility of quantitative information. We refer to these here as the Brief Mathematics Self-Doubt (BMSD) scale and the Preference for Numerical Information—Utility (PNIU) scale, respectively. In each participating class, students completed the BMSD and the PNIU during regular class periods early in the semester and again near the end of the semester (see footnote 2). To minimize experimenter demand effects, assessments made no mention of the module-related course activities prior to completion of these measures and relied only on student-generated anonymous codes for identification.

Given the uniqueness of these measures, we briefly review preliminary reliability and validity data below. The Brief Mathematics Self-Doubt scale (BMSD; Friedrich 2010) is an eight item self-report measure constructed by adapting the wording of a more general trait self-doubt scale (Oleson et al. 2000). The wording of the original scale's self-doubt items avoided specific references to academic course performance or test anxiety and instead focused on broader

positive/negative reactions or "comfort" in dealing with situations. The BMSD's adapted items (see Table 3) use minimal rephrasing to focus them on quantitative issues. For example, the original wording in the following item was supplemented with quantitative wording [in brackets]: "When engaged in an important task [*involving quantitative reasoning*], most of my thoughts turn to bad things that might happen (e.g., failing) rather than to good things." Item agreement/endorsement on the BMSD is indicated on a 1 (strongly disagree) to 6 (strongly agree) scale and averaged across items after reversing oppositely-keyed items for scoring. High scores indicate higher self-doubt or discomfort with respect to quantitative thinking.

Table 3
Brief Mathematics Self-Doubt (BMSD) Scale

1) When engaged in an important task involving quantitative reasoning, most of my thoughts turn to bad things that might happen (e.g., failing) rather than to good things.
2) For me, with math problems the emotional impact of avoiding failure (e.g., sense of relief) is greater than the emotional impact of achieving success (e.g., joy, pride).
3) More often than not I feel unsure of my mathematical abilities.
4) I sometimes find myself wondering if I have the ability to succeed at important activities if they involve numerical analysis.
5) I wish that I felt more certain of my strengths in understanding statistical information.
6) As I begin an important activity involving quantitative information, I usually feel confident in my ability.
7) Sometimes I feel that I don't know why I have succeeded at a quantitative task.
8) As I begin an important activity involving numerical information, I usually feel confident in a successful outcome.

Note: Items are adapted from Oleson et al. (2000). Item responses are on a 1 (strongly disagree) to 6 (strongly agree) scale. Total score is computed as the average across items after reverse scoring items 6 and 8.

The BMSD has shown promising reliability and validity in previous work. In an introductory psychology population (Friedrich et al. 2013), the BMSD demonstrated high internal consistency (Cronbach alpha = 0.90) and expected moderate negative correlations with measures of preference for analytical thinking (r = -0.48; scale from Bartels [2006]), basic numeracy (r = -0.37; scale from Lipkus et al. [2001]), and self-reported SAT-Quantitative Reasoning scores (r = -0.48), while showing appropriate discriminant validity with respect to SAT-Critical Reading scores (r = 0.02). Friedrich et al. (2013) also found that lower BMSD scores were significantly associated with better recall of experimentally presented drug risk information. In a separate study (Friedrich 2010), among women who saw a breast cancer drug portrayed *favorably* in an advertisement but accompanied by *unfavorable* clinical trial statistics, lower self-doubt scores were associated with lower effectiveness ratings (i.e., ratings more consistent with the clinical trial evidence).

What we have labeled the PNIU scale—our separate measure of the perceived usefulness of numerical information—consists of eight content-appropriate items drawn from the broader Preference for Numerical Information (PNI) scale (Viswanathan 1993; items 4, 6, 8, 11, 13, 15, 16, and 18). Agreement with statements such as "Numerical information is very useful in everyday life" and "Numbers are not necessary for most situations" (reverse keyed) is indicated

on a 1 (strongly disagree) to 6 (strongly agree) scale and averaged across items after reversing some for scoring. High scores on the PNIU indicate greater perceived usefulness of numerical information.

The PNIU showed high internal consistency in local pilot work with introductory psychology students (Cronbach alpha = 0.81) and an expected significant but moderate negative correlation with BMSD scores (r = -0.36)—a correlation subsequently replicating at r = -0.52 (Friedrich et al. 2013). It is worth noting that Gaze et al. (2014) employed a subset of five of these same items (accessed through independent sources) as a similar measure in validating their Quantitative Literacy and Reasoning Assessment (QLRA). In a much larger and heterogeneous sample, they observed a Cronbach alpha = 0.75 for their five-item scale and a significant correlation of r = 0.37 with their objective measure of QR performance.

It is important to note that the BMSD and PNIU were not included as simple proxies for quantitative skill. Although it is unsurprising that people reporting low self-doubt and people reporting high perceived utility tend to score higher on objective performance measures, the correlations are modest and the constructs being measured are not interchangeable. Individuals with objectively strong quantitative skills, for example, may nevertheless avoid engaging them because they find quantitative thinking distressing or see it as "not being worth the effort." In addition, individuals who experience discomfort when thinking quantitatively may nevertheless recognize the utility of their own (or others') rigorous quantitative analysis. Our assessment interest was in these elements of self-doubt and perceived utility—beliefs that might bear upon motivation to engage in QR.

Results

Individual study statistics, including confidence intervals on individual effect sizes, are included in the meta-analysis forest plots shown in Figures 1 and 2. Studies are grouped by category (see Table 2) and ordered by effect size. Given the two-mean, pre-post nature of each course's design, the reported meta-analyses adopt "d_Z" as the effect size measure (Lakens 2013), where the difference in means is standardized by the standard deviation of individual change scores.[3]

[3] The difference between pre- and posttest means in a paired sample *t*-test can be standardized by dividing by either the pooled standard deviation or by the standard deviation of the differences ("change scores") for individuals. Although there is some debate about which approach to standardizing is most informative, Lakens (2013) notes that when all included studies share the same within-subjects design, and when changes for each individual are the outcomes of interest, standardizing by the standard deviation of change scores is appropriate (labeled d_Z by Lakens).

DOI: https://doi.org/10.5038/1936-4660.12.2.8

Conventional guidelines suggest values of 0.2, 0.5, and 0.8 as benchmarks for small, medium, and large effects, respectively.

In the following sections, we report results separately for the mathematics self-doubt measure (BMSD, for which negative change and d_Z values indicate a reduction in self-doubt) and the utility of numerical information measure (PNIU, for which positive change and d_Z values indicate an increase in perceived utility). We first explore possible participant attrition effects, then assess reliabilities for the outcome measures in the present samples, and finally report on the overall meta-analyses with course category as a moderator variable.

Exploring Potential Non-completion Effects

Due to normal add/drop changes, withdrawals, and absences at the time of posttesting, it is possible that the samples for the meta-analysis reflect selective attrition based on QR-related factors. To explore this possibility, we compared the BMSD and PNIU pretest scores of those who, for whatever reason, did not complete posttesting to those who did, and thus are included in the final matched-sample effect sizes. An initial comparison between completers' and noncompleters' BMSD and PNIU scores done separately for each course (in cases having the requisite two or more non-completers for analysis) revealed no significant differences, but such small-*N* comparisons are severely underpowered.

Aggregating across all 20 courses, the difference between BMSD scores for completers (M = 3.33, n = 349) and non-completers (M = 3.29, n = 134) did not approach significance: $t(481) = 0.41$, $p = 0.68$. The difference in PNIU scores between completers (M = 4.28, n = 349) and non-completers (M = 4.29, n = 137) also failed to approach significance: $t(484) = -0.15$, $p = 0.88$. Separate comparisons of completers and non-completers grouped within each of the three broad course classifications also failed to show significant differences on either measure (all ps > 0.4). Thus, even with relatively high-powered comparisons, there was no evidence to indicate selective attrition occurred as a function of pretest scores on the outcome variables of interest.

Outcome Measure Reliability

Effect sizes can be substantially underestimated in the presence of measurement error. Cronbach alpha measures of internal consistency were calculated separately for each course and—given that small sample estimates are unstable—also for the overall sample of completers at both pretest and posttest. For BMSD pretest scores, the median course-based alpha value was 0.88, and the combined-samples alpha was 0.87 (N =349). For BMSD posttest scores, alphas were 0.89 and 0.89, respectively. For PNIU pretest scores, the median course-based alpha value was 0.81, and the combined-samples alpha was also 0.81 (N = 348). For PNIU posttest

scores, alphas were 0.83 and 0.84 respectively. Thus, these measures showed high internal consistency reliability both early and late in the semester.

Because interventions target underlying constructs, and individuals might respond differently to any given intervention, one would expect test-retest correlations to be somewhat lower. Nevertheless, they capture consistency in people's relative standing in scores over time. For the BMSD, the median course-based pre-post correlation was 0.81, with a combined sample pre-post correlation of 0.78 (N = 349). For the PNIU, the respective values were 0.67 and 0.72 (N = 349). Thus, independent of any shift in pre-post averages, both measures reflected substantial temporal consistency in relative standing.[4]

Meta-analyses and Tests of Moderation

Separate meta-analyses were conducted for the set of BMSD effects (see Fig. 1) and the set of PNIU effects (see Fig. 2). The figures include test statistics and effect size confidence intervals both for individual courses and for the meta-analytic aggregates. A critical decision involves whether to pursue a "fixed effects" or a "random effects" analysis (Borenstein et al. 2009). Because each module intervention was quite different and because the courses were diverse in terms of discipline, we had reason to believe that the true population effect size underlying each sample effect would not be a single, fixed value as assumed in a fixed effect analysis but would instead vary—for example, by disciplinary category—as assumed in a random effects analysis. Thus, on theoretical grounds, we believed that a random effects model would be most appropriate analysis here. For completeness, however, we report the outcome for an initial fixed effect analysis along with the full random effects analysis.

In each case, the d_z measures of standardized differences in pre and post means were subjected to an initial fixed effect meta-analysis, which also yields a test for heterogeneity using the Q statistic. A significant Q value indicates greater variability in individual study effects than one would expect under the single or "fixed" population effect null hypothesis, suggesting the potential influence of a moderator variable. The I^2 statistic indicates the percentage of variation in observed study effect sizes that is likely attributable to differences in true effect sizes, with rough benchmark values of <25% for low, 25–50% for moderate, and >50% for large values (Borenstein et al. 2009). We then report the preferred random effects meta-analysis that allows for variability in true effect sizes. A random effects analysis also allows one to explore the influence of moderator

[4] As expected, BMSD and PNIU scores for the full sample of completers were moderately negatively correlated with each other at both pretesting, r = -0.35, and at posttesting, r = -0.41 (ps < 0.001). Given the high reliabilities of the measures, these moderate values suggest that meaningfully related but non-redundant constructs are being assessed by the self-doubt and perceived utility scales.

Group by Course Category	Study name	Std Paired Difference	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Arts & Humanities	History_3XXA	-0.54	0.32	0.10	-1.17	0.09	-1.67	0.10
Arts & Humanities	Politics_2XX	-0.52	0.20	0.04	-0.91	-0.12	-2.58	0.01
Arts & Humanities	Anthropology_2XX	-0.38	0.28	0.08	-0.93	0.16	-1.39	0.16
Arts & Humanities	ArtHistory_2XX	-0.35	0.27	0.07	-0.87	0.17	-1.32	0.19
Arts & Humanities	English_3XX	-0.28	0.27	0.07	-0.81	0.25	-1.03	0.30
Arts & Humanities	Politics_3XXA	-0.14	0.29	0.08	-0.71	0.43	-0.49	0.63
Arts & Humanities	History_3XXB	-0.09	0.27	0.07	-0.62	0.43	-0.35	0.73
Arts & Humanities	History_3XXC	-0.03	0.27	0.07	-0.55	0.50	-0.11	0.92
Arts & Humanities	Politics_3XXB	0.14	0.32	0.10	-0.49	0.76	0.43	0.67
Arts & Humanities		-0.27	0.09	0.01	-0.45	-0.08	-2.85	0.00
Behavioral Sciences	Psychology_4XX	-0.49	0.43	0.19	-1.33	0.36	-1.13	0.26
Behavioral Sciences	Economics_3XX	-0.39	0.27	0.07	-0.92	0.13	-1.46	0.14
Behavioral Sciences	Psychology_2XX	-0.30	0.19	0.04	-0.68	0.08	-1.57	0.12
Behavioral Sciences	Psychology_3XX	-0.11	0.20	0.04	-0.51	0.29	-0.55	0.58
Behavioral Sciences	Sociology_2XX	-0.09	0.26	0.07	-0.60	0.42	-0.35	0.73
Behavioral Sciences		-0.24	0.12	0.01	-0.47	-0.01	-2.07	0.04
Natural Sciences & Math	ExerciseScience_3XX	-0.39	0.37	0.13	-1.10	0.33	-1.05	0.29
Natural Sciences & Math	Biology_1XXB	-0.18	0.17	0.03	-0.52	0.15	-1.08	0.28
Natural Sciences & Math	Biology_1XXA	0.09	0.21	0.05	-0.32	0.51	0.44	0.66
Natural Sciences & Math	Math_1XX	0.24	0.17	0.03	-0.09	0.58	1.42	0.16
Natural Sciences & Math	ExerciseScience_1XX	0.35	0.30	0.09	-0.23	0.94	1.19	0.23
Natural Sciences & Math	Biology_2XX	0.77	0.28	0.08	0.23	1.31	2.78	0.01
Natural Sciences & Math		0.13	0.10	0.01	-0.06	0.32	1.33	0.18
Overall		-0.12	0.06	0.00	-0.23	-0.00	-2.02	0.04

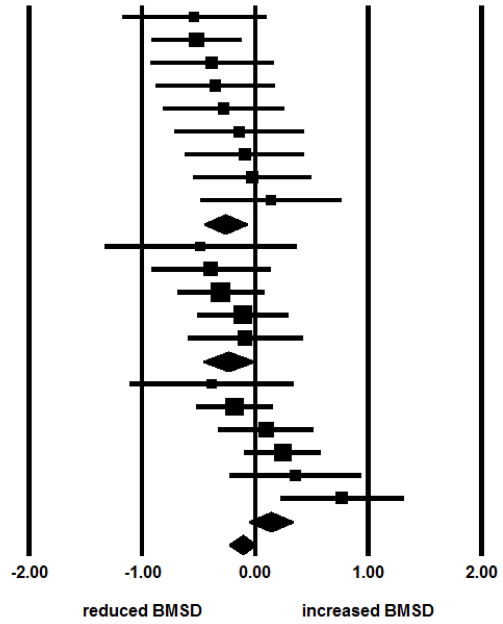

Figure 1. Forest plot showing BMSD effects for individual studies and meta-analytic course groupings. Each course's standardized paired difference (d_Z value) for Brief Mathematics Self Doubt (BMSD) scores is represented by a square within a 95% confidence interval. Negative d_Z values indicate a reduction in self-doubt from pre- to posttesting, with the size of each square proportional to that study's relative weight in the meta-analysis. Confidence intervals for course group and overall effects are depicted as diamonds centered on the relevant average effect.

variables—specifically the disciplinary course group categories in the present data set—in accounting for effect size heterogeneity. All analyses were performed using Comprehensive Meta-Analysis (CMA) Version 2 (Borenstein et al. 2005).[5]

Analysis of Brief Mathematics Self-Doubt (BMSD) Effects. An initial fixed effects analysis yielded significant heterogeneity, with $Q(19) = 30.11$, $p = 0.05$, and $I^2 = 36.9\%$. The random effects analysis yielded a non-significant overall mean effect size of $M_d = -0.12$, $Z = -1.71$, $p = 0.09$, $CI_{95\%}[-0.26, +0.02]$. A subsequent meta-analysis incorporating course group as a moderator tested for the presence of subgroup differences and estimated mean effect sizes within each subgroup according to a random effects model using a pooled variance term (see Borenstein et al. 2009). The effect of the course group moderator was significant—$Q(2) = 10.09$, $p = 0.006$—with arts & humanities and behavioral sciences average effects revealed as different from the effect size for the natural sciences & mathematics group.

Recalling that a drop in mathematics self-doubt over time would appear as a negative d_Z value, the weighted mean effect for the five behavioral science courses was $M_d = -0.24$, $Z = -2.07$, $p = 0.04$, $CI_{95\%}[-0.47, -0.01]$. Similarly, the weighted mean effect for the nine arts & humanities courses was $M_d = -0.27$, $Z = -2.85$, $p = 0.004$, $CI_{95\%}[-0.45, -0.08]$. In contrast to this self-doubt reduction, the weighted mean effect for the six natural sciences & mathematics courses was a positive but non-significant $M_d = 0.13$, $Z = 1.33$, $p = 0.18$, $CI_{95\%}[-0.06, +0.32]$. Thus, the modest and non-significant mean d_Z value of -0.12 observed across the 20 studies reflected the negative (self-doubt reduction) effects associated with the arts & humanities and behavioral sciences groups along with the non-significant positive effect (increase) observed for the natural sciences & mathematics group.

Analysis of Preference for Numerical Information—Utility (PNIU) Effects. A parallel series of analyses performed on the set of PNIU effects once again yielded significant heterogeneity, with $Q(19) = 43.51$, $p = 0.001$, and $I^2 = 56.33\%$. An increase over time in the perceived usefulness of numerical information is reflected in a positive d_Z value, and the random effects meta-analysis for the 20 studies yielded a significant overall mean effect size of $M_d = 0.27$, $Z = 3.10$, $p = 0.002$, $CI_{95\%}[0.10, 0.44]$. Following the same approach as with the BMSD effects, a meta-analysis incorporating course group as a moderator did not reveal significant differences between the subgroup M_d values, with $Q(2) = 2.14$, $p = 0.34$. Examination of mean effect sizes for subgroups is nevertheless informative.

[5] Although we used a commercial software package for these analyses, user-friendly open-source options such as those included in the Jamovi Project (2018, v0.9; www.jamovi.org) are available.

Group by Course Category	Study name	Std Paired Difference	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Arts & Humanities	English_3XX	-0.34	0.27	0.08	-0.88	0.20	-1.24	0.21
Arts & Humanities	Politics_3XXA	-0.03	0.28	0.08	-0.58	0.51	-0.11	0.91
Arts & Humanities	History_3XXB	0.06	0.27	0.07	-0.46	0.59	0.24	0.81
Arts & Humanities	History_3XXC	0.30	0.27	0.07	-0.24	0.83	1.10	0.27
Arts & Humanities	Politics_2XX	0.39	0.20	0.04	0.01	0.78	2.00	0.05
Arts & Humanities	Anthropology_2XX	0.45	0.28	0.08	-0.10	1.00	1.59	0.11
Arts & Humanities	ArtHistory_2XX	0.45	0.27	0.07	-0.08	0.99	1.67	0.09
Arts & Humanities	History_3XXA	0.63	0.33	0.11	-0.02	1.27	1.90	0.06
Arts & Humanities	Politics_3XXB	0.87	0.37	0.14	0.14	1.59	2.34	0.02
Arts & Humanities		0.29	0.14	0.02	0.02	0.55	2.12	0.03
Behavioral Sciences	Economics_3XX	-0.27	0.26	0.07	-0.79	0.24	-1.03	0.30
Behavioral Sciences	Psychology_2XX	-0.03	0.19	0.04	-0.40	0.34	-0.16	0.87
Behavioral Sciences	Psychology_4XX	0.67	0.50	0.25	-0.30	1.64	1.36	0.17
Behavioral Sciences	Sociology_2XX	1.17	0.33	0.11	0.51	1.82	3.49	0.00
Behavioral Sciences	Psychology_3XX	1.23	0.27	0.07	0.70	1.76	4.55	0.00
Behavioral Sciences		0.47	0.19	0.03	0.11	0.84	2.53	0.01
Natural Sciences & Math	Biology_1XXB	-0.07	0.17	0.03	-0.40	0.27	-0.39	0.70
Natural Sciences & Math	Biology_2XX	0.11	0.24	0.06	-0.36	0.59	0.47	0.64
Natural Sciences & Math	Biology_1XXA	0.15	0.21	0.05	-0.27	0.57	0.72	0.47
Natural Sciences & Math	Math_1XX	0.16	0.17	0.03	-0.17	0.50	0.96	0.34
Natural Sciences & Math	ExerciseScience_1XX	0.17	0.29	0.08	-0.40	0.74	0.58	0.56
Natural Sciences & Math	ExerciseScience_3XX	0.27	0.36	0.13	-0.44	0.97	0.74	0.46
Natural Sciences & Math		0.12	0.15	0.02	-0.18	0.42	0.77	0.44
Overall		0.27	0.09	0.01	0.10	0.45	3.05	0.00

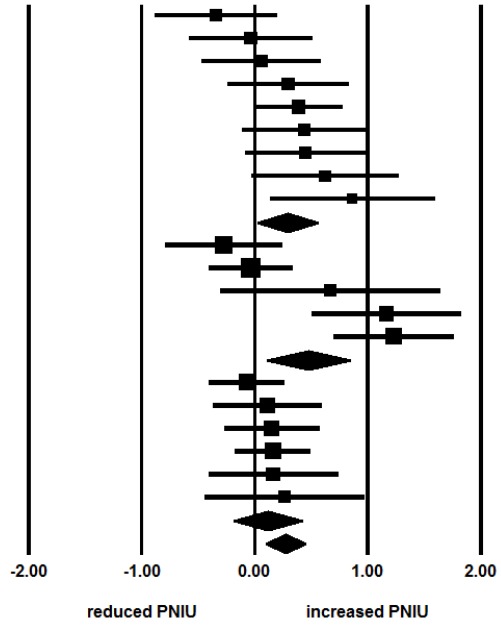

Figure 2. Forest plot showing PNIU effects for individual studies and meta-analytic course groupings. Each course's standardized paired difference (d_Z value) for Preference for Numerical Information—Utility (PNIU) scores is represented by a square within a 95% confidence interval. Positive d_Z values indicate an increase in perceived utility from pre- to posttesting, with the size of each square proportional to that study's relative weight in the meta-analysis. Confidence intervals for course group and overall effects are depicted as diamonds centered on the relevant average effect.

The weighted mean effect for the five behavioral science courses was M_d = 0.47, Z = 2.53, p = 0.01, $CI_{95\%}$[0.11, 0.84]. The weighted mean effect for the nine arts & humanities courses was also positive and significant, with M_d = 0.29, Z = 2.12, p = 0.03, $CI_{95\%}$[0.02, 0.55]. Finally, the weighted mean effect for the six natural sciences & mathematics courses was positive but did not approach significance, with M_d = 0.12, Z = 0.77, p = 0.44, $CI_{95\%}$[-0.18, +0.42]. The absence of a significant Q value for the moderator indicates the current data cannot rule out chance-level differences between the mean effects for the subgroups, and thus the between-study heterogeneity does not appear to be readily explained by course grouping. Nevertheless, it is consistent with the BMSD analysis that significant (now positive) shifts in perceived usefulness of quantitative reasoning were seen for both the behavioral sciences and the art & humanities groups, but not for the natural sciences & mathematics group.

Discussion

In this assessment study, we focused on a diverse array of QR module interventions at a single institution, investigating their implementation in 20 distinct courses. In designing the modules, we worked within a framework of quantitative literacy that, among other elements, emphasized beliefs "embracing" quantitative reasoning (see Table 1). This led us to focus specifically on students' beginning-to-end-of-term changes in both subjective comfort with quantitative thinking (BMSD scores) and belief in its utility (PNIU scores). Actual quantitative skills are of course essential to effective QR, but such skills may not be used effectively when people find such thinking to be unpleasant or fail to view quantitative information as potentially useful and worth the processing effort (Rheinlander and Wallace 2011).

Both the BMSD and PNIU scales demonstrated strong psychometric qualities, consistent with past work. We integrated their effect sizes across the 20 course interventions and explored the potential moderating role of subject area via random effects meta-analysis. A standardized mean difference (d_Z) value of 0.2 is generally considered a small effect, with 0.5 considered medium sized. By this rule of thumb, our average effects were modest but in expected directions. The significant, average effect sizes for BMSD and PNIU changes for arts & humanities courses indicated a reduction in self-doubt (M_d = -0.27) and an increase in perceived usefulness (M_d = 0.29). Significant mean effect sizes were also observed for the quantitative behavioral sciences courses, with M_d = -0.24 for the self-doubt measure and M_d = 0.47 for the perceived utility measure. It was only in the natural sciences & mathematics group of courses that mean effect sizes failed to reach significance, with M_d = 0.13 for the BMSD and M_d = 0.12 for the PNIU.

Limitations to Meta-analytic Integration

There are, of course, a number of important limitations to consider. Some are specific to meta-analysis as an approach, and others are common challenges facing educators implementing and assessing programs such as the one we describe. For example, all of the studies incorporated in our meta-analysis were simple pre-/posttest designs, and such studies are ill-suited to making strong causal claims about the effects of each individual module. However, when QR and similar curricular interventions are adapted from other models and implemented by practitioners, it is not always the case (and, one might argue, actually seldom the case) that each attempt is evaluated as a randomized clinical trial.

In the module intervention presented here, we were concerned primarily with collecting data in ways that were minimally intrusive for instructors and that held at least the potential for identifying positive changes. Our assessment goal was to determine whether things appeared to be moving in the desired direction—a sign that the interventions *might* be a plausible source of change. The fact that we could not attribute positive changes to the selective attrition of students with initially higher self-doubt or lower perceived utility scores rules out certain confounds, but it is worth noting that we did not have fine-grained information on the reasons for non-completion. Individuals might have dropped after pre-testing but before the intervention, dropped after the intervention and possibly in response to the QR focus, or simply missed class on the date of the posttesting. These are concerns for future researchers to consider, but we did not find specific evidence to suggest that incomplete data for students varied as a function of scores on the outcome measures in ways that might unfairly favor our hypotheses.

The additional fact that beliefs changed in theoretically meaningful ways as a function of course category strengthens our conclusions in other ways. Nevertheless, these findings do not rule out other alternative accounts. For example, it is obviously the case that additional, non-module content was included in every course and that students were taking a host of other courses concurrently—both factors that could have influenced the observed changes. We also note here that we subsequently discovered 23 students (out of N= 349 completers) who had at some point participated in two module courses in different semesters over the five-year span of the intervention. Such a small and dispersed percentage of repeat participants would do very little to impact the presumed statistical independence of the effect sizes being integrated in the meta-analysis, but non-independence of samples is certainly a factor to keep in mind when there are multiple interventions at a single institution. For these few students, it seems likely that QR experiences before the second module course—whether from the prior module experience or through other coursework—would have pushed them

closer to a ceiling in terms of scores and thus worked to limit the change expected by the intervention. While causality is uncertain in our design, a consistent *absence* of change for the sample as a whole would have raised legitimate concerns about the potential of our module program.

The broader point is that meta-analysis is not a panacea for addressing all limitations in the constituent studies. Non-causal designs yield meta-analytic findings that are themselves non-causal in their interpretation. This caution is especially important when considering moderator variables. In theory, there is no limit to the number of moderating variables one could explore. Given significant heterogeneity of effects, one could group studies by type of institution, level of course, type of module intervention, gender of instructor, and so on. As such, it is tempting to go on a "fishing expedition" with moderator variable analyses that capitalize on chance differences (an issue similar to uncorrected post hoc comparisons in traditional inferential analysis). In the present study, we limited ourselves to the course-type moderator because our grant had specifically proposed this kind of diversity of courses and because it could provide a meaningful illustration of how moderator analysis can inform assessment efforts.

Additional concerns of meta-analysis include access to the relevant studies and specification of inclusion criteria. In our present, prospective design, we were fortunate to have access to all the relevant interventions. Determining in advance what studies to include avoids the obvious problem associated with adding studies only to the point that a significant average effect is achieved (Ueno et al. 2016). When analyses are instead based on retrieval of prior work, considerable care is required to identify qualifying (and often unpublished) reports of interventions in order to address what is commonly referred to as the *file drawer* problem (Cumming 2014). Specifying the criteria for inclusion is also critical for generating analyses that combine effects of meaningfully related phenomena as well as for avoiding accusations of post hoc "cherry picking" of studies that favor a desired outcome. Fortunately, a broad procedural literature and established guidelines (e.g., Kepes et al. 2013) already exist to guide QR researchers interested in exploiting the strengths of meta-analytic approaches.

Benefits of Meta-Analytic Integration

Despite the modest size of our effects and limitations of the analysis that we have noted, our meta-analytic findings provided additional valuable feedback regarding our institutional intervention. Perhaps most importantly, an inspection of the individual significance tests and confidence intervals on effect sizes (Figs. 1 & 2) shows that at the individual course/study level, nearly all of the individual effects (34 of 40) failed to achieve statistical significance. Given the institutional environment in which we implemented our module program, all the studies had sample sizes that left "by-course" analyses severely underpowered. As a result, it

would have been easy to miss the impact of the overall intervention had we simply focused on individual courses and conventional significance tests.

Faculty participants in our intervention were recruited based on their interest in institutional QR goals and the diversity of their courses' topical areas and levels, but no interested parties were discouraged or excluded based on size of enrollment. To illustrate the importance of power concerns, the probability of obtaining a statistically significant effect with a two-tailed 0.05-level paired *t* test when the true population effect is a small one ($d_Z = 0.2$) is only 0.065 with $n = 5$ (our smallest sample) and 0.21 with $n = 35$ (our largest sample). Thus, even before beginning, each individual course would have been quite likely to miss the effect (non-significant results) if evaluated alone when the *true population effect* was on the smaller side as suggested in our meta-analysis. With small sample sizes, it is in fact possible to find sample effect sizes of the opposite sign. Only by integrating across interventions were reliable underlying effects revealed.

Of course, such concerns with sampling variability and low power hold equally true for evaluating other types of QR interventions, as do the benefits of meta-analytic integration. Of particular concern, the limitations of single-study null hypothesis tests are likely to be poorly understood by participating faculty who lack an inferential statistics background. Not only might they be discouraged when their single course implementations appear to "fail," but institutions may decline to support broader programmatic changes if assessments yield highly variable and frequently non-significant individual effects. Even if individual instructors might feel somewhat discouraged by a statistical analysis of their own course's data, meta-analytic findings can show a different picture for an intervention at the institutional level.

The analysis of the course-group moderator variable also offers potentially useful insights for institutional intervention. For example, our grant proposal was particularly concerned with showing effects in courses that are often more heavily enrolled by our less quantitatively oriented students. Students in arts & humanities disciplines likely receive much less QR exposure than those in STEM and quantitative behavioral science disciplines, and they may self-select into courses and programs based in part on their QR beliefs. Of course, classes in our arts & humanities group certainly served students beyond just majors, and embedded QR modules in these areas hold potential for enhancing the perceived relevance of quantitative thinking for all who enroll.

It was also noteworthy that we did not see the anticipated changes over time in our natural sciences & mathematics group, although there was no evidence of significant negative effects on our outcome measures. These courses often place much greater demands on QR skills. For example, in a mathematics course or a majors' biology course, students are often evaluated on tasks that aggressively push their current QR skill sets. As a result, QR activities could have been more

highly associated with anxiety and/or potential failure in such courses, in contrast to the perhaps more novel and less technically demanding experiences in our arts & humanities offerings. The module interventions might nevertheless have generated other benefits for the natural sciences & mathematics group—especially in QR skill areas—and thus it would be inappropriate to assert that modules were of no benefit. Regardless, such a meta-analysis of potential moderators can generate important local discussion about courses and parts of the curriculum to target given finite resources and support.

One such example of a potential moderator would be to explore breakdowns based on gender and representation in historically underrepresented groups—especially within STEM areas. Given the small sample sizes for most of our module courses, we had not anticipated that within-course breakdowns would be statistically reliable, but such information could certainly be of interest. In meta-analyses based on retrieval of past studies, such breakdowns might not be available in those sources for input as separate effect sizes in a meta-analysis. However, incorporating systematic collection of demographic information into the design of planned, prospective meta-analyses is certainly appropriate and—given sufficiently large samples for subgroups within each study—holds promise for uncovering important relationships.

One somewhat unique aspect of our analysis is that we were able to make use of locally developed and validated measures and employed the same ones in all of the constituent studies. It is important to note, however, that such customization and consistency are not essential for meta-analysis so long as the measures across studies validly assess the same or similar constructs. The data points of analysis are the studies' standardized effect sizes, and these are measure-independent in terms of scaling. Differences in measures or types of measures can simply be incorporated as potential moderator variables if need be. Nevertheless, for the kind of institutional meta-analysis we have promoted in our example, consistent measurement tools are often an option, and evaluating reliability and construct validity as we did with the BMSD and PNIU strengthens any conclusions.

Concluding Remarks

The meta-analysis described here synthesized research within a single institution, but the technique is equally applicable to studies conducted across multiple institutions. For example, collaborative partnerships with replications at different schools can be planned in advance with a meta-analytic synthesis in mind. Whether the constituent studies reflect single or multiple institutions, identical or diverse measures of a common construct, or structural differences that can be explored as potential moderator variables, meta-analysis as a general technique helps to shift assessment focus away from mere statistical significance to the magnitude of effects and their consistency across implementations.

Given meta-analysis' emphasis on effect size, it is worth considering how large an effect needs to be in order to be meaningful. A computed average effect size, by definition, will not be as impressive as the largest effect sizes often reported in individual studies. Single, successful studies are more likely to be accepted for publication, even though many such studies reflect Type 1 errors or give unrealistically large estimates of the true effect size (Cumming 2014). Underpowered/small-*N* studies can require sample effects that are well above the true population effect size in order to achieve conventional levels of statistical significance, biasing the magnitude of effects reported in published sources. This bias can lead to unrealistic expectations among adopters of empirically supported QR interventions.

As applied to the current assessment, our own findings yielded reliable evidence for effects that were favorable but modest by conventional rules of thumb. The true meaning of any effect size, however, depends on a variety of factors, with even small effects often proving to be important when considered over time or across large populations. Although one might hope to generate dramatic changes with well-designed interventions, such hopes are often unrealistic. For example, stable QR beliefs like the ones we assessed form through a lifetime of experiences both in and out of the educational system.

It is also the case that we limited our analysis to the immediate effects of the intervention. Our data do not address whether the changes we observed are lasting, or whether they might be cumulative across multiple module experiences if such interventions were implemented more broadly in the curriculum. Students experiencing not single modules, but multiple modules over time could experience a cumulative effect not evident in our short-term and individual course intervention assessment.

Thus, our analysis gives only a snapshot regarding the potential of QR module interventions like ours in the broader curriculum. We suspect that our initiative was not unique in this regard and that many evidence-based QR interventions are launched with the hope that broader and longer exposures might yield a stronger and more enduring impact. A particular area of potential interest in assessing outcomes involves exploring effectiveness with students who have historically been underrepresented in STEM and quantitative behavioral science areas. Given sufficient sample sizes and appropriate demographic information, a meta-analytic approach to assessment might effectively highlight promise that could otherwise escape the attention of individual instructors and policy makers.

Many if not most QR researchers are already familiar with the fundamentals of meta-analysis, as such reports are frequently referenced in the theory development and literature review sections of intervention papers. Nevertheless, in preparing this manuscript, we uncovered no examples in this journal for which meta-analysis was an integral part of the actual QR assessment effort. Through

the illustration with our own module program, we hope that others might consider its potential benefits for coordinating research efforts both within and between institutions and for strengthening the evidence base supporting effective and generalizable QR interventions.

Acknowledgments

We wish to thank the instructors who participated in the module project. We also thank the Teagle Foundation for its support in this research through a grant from its Engaging Evidence initiative.

References

Bartels, Daniel M. 2006. "Proportion Dominance: The Generality and Variability of Favoring Relative Savings Over Absolute Savings." *Organizational Behavior and Human Decision Processes* 100: 76–95. https://doi.org/10.1016/j.obhdp.2005.10.004

Borenstein, Michael, Larry Hedges, Julian Higgins, and Hannah Rothstein. 2005. Comprehensive Meta-Analysis (Version 2). Windows. Engelwood, NJ: Biostat.

Borenstein, Michael, Larry V. Hedges, Julian P.T. Higgins, and Hannah R. Rothstein. 2009. *Introduction to meta-analysis*. Chichester: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386

Bressoud, David. 2009. "Establishing the Quantitative Thinking Program at Macalester." *Numeracy* 2(1): Article 3. https://doi.org/10.5038/1936-4660.2.1.3

Cumming, Geoff. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. New York, NY: Routledge. https://doi.org/10.4324/9780203807002

Cumming, Geoff. 2014. "The New Statistics. Why and How." *Psychological Science* 25: 7–29. https://doi.org/10.1177/0956797613504966

Diefenderfer, Caren. 2012. "The Joy of Quantitative Reasoning." *Numeracy* 5(1): Article 1. https://doi.org/10.5038/1936-4660.5.1.1

Follette, Katherine, Sanlyn Buxner, Erin Dokter, Donald McCarthy, Beau Vezino, Lori Brock, and Edward Prather. 2017. "The Quantitative Reasoning for College Science (QuaRCS) Assessment 2: Demographic, Academic and Attitudinal Variables as Predictors of Quantitative Ability." *Numeracy* 10(1): Article 5. https://doi.org/10.5038/1936-4660.10.1.5

Friedrich, James. 2010. "Numeracy and Mathematics Self Doubt: Exploring Potential Confounding in Judgment Contexts." (poster presentation at the Society for Judgment and Decision Making Annual Conference, St. Louis,

MO, November 21).
Friedrich, James, Jonathan Wenger, Kirstin Demezas. 2013. "Subjective Numeracy and Mathematics Self Doubt as Predictors of Numeracy-Related Constructs and Risk Information Processing." (poster presentation at the Society for Judgment and Decision Making Annual Conference, Toronto, ON, November 17).
Gaze, Eric C., Aaron Montgomery, Semra Kilic-Bahi, Deann Leoni, Linda Misener, and Corrine Taylor. 2014. "Towards Developing a Quantitative Literacy/Reasoning Assessment Instrument." *Numeracy* 7(2): Article 4. https://doi.org/10.5038/1936-4660.7.2.4
Goh, Jin X., Judith A. Hall, and Robert Rosenthal. 2016. "Mini Meta-analysis of Your Own Studies: Some Arguments on Why and a Primer on How." *Social and Personality Psychology Compass* 10(10): 535–549. https://doi.org/10.1111/spc3.12267
Hillyard, Cinnamon. 2012. "Comparative Study of the Numeracy Education and Writing Across the Curriculum Movements: Ideas for Future Growth." *Numeracy* 5(2): Article 2. https://doi.org/10.5038/1936-4660.5.2.2
Karaali, Gizem, Edwin H. Villafane Hernandez, and Jeremy A. Taylor. 2016. "What's in a Name? A Critical Review of Definitions of Quantitative Literacy, Numeracy, and Quantitative Reasoning." *Numeracy* 9(1): Article 2. https://doi.org/10.5038/1936-4660.9.1.2
Kepes, Sven, Michael A. McDaniel, Michael T. Brannick, and George C. Banks. 2013. "Meta-analytic Reviews in the Organizational Sciences: Two Meta-analytic Schools on the Way to MARS (the Meta-analytic Reporting Standards)." *Journal of Business and Psychology* 28(2): 123–143. https://doi.org/10.1007/s10869-013-9300-2
Klein, Richard A., Kate A Ratcliff, Michelangelo Vianello, Reginald B. Adams Jr., Štepán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, Brian A. Nosek. 2014. "Investigating Variation in Replicability: A 'Many Labs' Replication Project." *Social Psychology* 45(3): 142–152. https://doi.org/10.1027/1864-

9335/a000178
Lakens, Daniel. 2013. "Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-tests and ANOVAs." *Frontiers in Psychology* 4: Article ID 863. https://doi.org/10.3389/fpsyg.2013.00863
Lipkus, Isaac M., Greg Samsa, and B. K. Rimer. 2001. "General Performance on a Numeracy Scale Among Highly Educated Samples." *Medical Decision Making* 21(1): 37–44. https://doi.org/10.1177/0272989X0102100105
Maner, Jon K. 2014. "Let's Put Our Money Where Our Mouth Is: If Authors Are to Change Their Ways, Reviewers (and Editors) Must Change With Them." *Perspectives on Psychological Science* 9(3): 343–351. https://doi.org/10.1177/1745691614528215
Maxwell, S. E., M. Y. Lau, and G. S. Howard. 2015. "Is Psychology Suffering from a Replication Crisis?: What Does 'Failure to Replicate' Really Mean?" *American Psychologist* 70(6): 487–498. https://doi.org/10.1037/a0039400
Oleson, Kathryn C., Kirsten M. Poehlmann, John H. Yost, Molly Lynch, and Robert M. Arkin. 2000. Subjective Overachievement: Individual Differences in Self-doubt and Concern with Performance." *Journal of Personality* 68(3): 491–524. https://doi.org/10.1111/1467-6494.00104
Open Science Collaboration. 2015. *Science* Aug 28: 349(6251):aac4716. https://doi.org/10.1126/science.aac4716
Rheinlander, Kim, and Dorothy Wallace. 2011. "Calculus, Biology and Medicine: A Case Study in Quantitative Literacy for Science Students." *Numeracy* 4(1): Article 3. https://doi.org/10.5038/1936-4660.4.1.3
Steele, Benjamin, and Semra Kilic-Bahi. 2010. "Quantitative Literacy: Does it Work? Evaluation of Student Outcomes at Colby-Sawyer College." *Numeracy* 3(2): Article 3. https://doi.org/10.5038/1936-4660.3.2.3
Sundre, Donna L., and Amy D. Thelk. 2010. "Advancing Assessment of Quantitative and Scientific Reasoning." *Numeracy* 3(2): Article 2. https://doi.org/10.5038/1936-4660.3.2.2
Ueno, Tajii, Greta M. Fastrich, and Kou Murayama. 2016. "Meta-analysis to Integrate Effect Sizes within an Article: Possible Misuse and Type I Error Inflation." *Journal of Experimental Psychology: General* 145(5): 643–654. https://doi.org/10.1037/xge0000159
Vachar, H.L. and Emily Lardner. 2010. "Spreadsheets Across the Curriculum, 1: The Ideas and the Resource." *Numeracy* 3(2): Article 6. https://doi.org/10.5038/1936-4660.3.2.6
Viswanathan, Madhubalan. 1993. "Measurement of Individual Differences in Preference for Numerical Information." *Journal of Applied Psychology* 78(5): 741–752. https://doi.org/10.1037/0021-9010.78.5.741
Wenner, Jennifer M., Eric M. Baer, Cathryn A. Manduca, R. Heather Macdonald, Samuel Patterson, and Mary Savina. 2009. "The Case for Infusing

Quantitative Literacy into Introductory Geoscience Courses." *Numeracy* 2(1): Article 4. https://doi.org/10.5038/1936-4660.2.1.4

Wilkins, Jesse L. 2000. "Preparing for the 21st Century: The Status of Quantitative Literacy in the United States." *School Science and Mathematics* 100(8): 405–18. https://doi.org/10.1111/j.1949-8594.2000.tb17329.x