Numeracy
Advancing Education in Quantitative Literacy

Volume 12 | Issue 2 Article 2

2019

Paired Measures of Competence and Confidence Illuminate Impacts of Privilege on College Students

Rachel M. Watson
University of Wyoming, rwatson@uwyo.edu
Edward Nuhfer
California State University (retired), enuhfer@earthlink.net
Kali Nicholas Moon
University of Wyoming, kalinicholas@gmail.com
Steven Fleisher
California State University Channel Islands, steven.fleisher@csuci.edu
Paul Walter
St. Edwards University, pauljw@stedwards.edu

See next page for additional authors

Recommended Citation

Watson, Rachel M., Edward Nuhfer, Kali Nicholas Moon, Steven Fleisher, Paul Walter, Karl Wirth, Christopher Cogan, Ami Wangeline, and Eric Gaze. "Paired Measures of Competence and Confidence Illuminate Impacts of Privilege on College Students." *Numeracy* 12, Iss. 2 (2019): Article 2. DOI: https://doi.org/10.5038/1936-4660.12.2.2

Paired Measures of Competence and Confidence Illuminate Impacts of Privilege on College Students

Abstract

We seek to understand how the experiences of groups that differ in gender, ethnicity, and sexual orientation produce college-level educational performances that differ from the experiences of the dominant majority group. We employ two datasets: a National Database of 24,701 participants and a Paired-Measures Database with 3,323 participants. Both datasets provide demographic information, socioeconomic conditions of status as first-generation student, English as a first language, and interest in majoring in science, and competency scores on understanding science as a way of knowing obtained from the Science Literacy Concept Inventory. The Paired-Measures Database includes additional self-assessed competence ratings that enabled quantifying affective confidence. We meld the ways of knowing of ethics, numeracy, and social justice, especially the social justice concept of Othering, to interpret our data. Two of three competing hypotheses about self-assessment encourage Othering. Our data strongly support the third—that all groups are good at self-assessment and merit equal respect. Women and men are equally competent in science literacy. Women, on average, are more accurate in their self-assessments whereas men, on average, are overconfident. Those with minority sexual orientations register higher competence than the binary-sexual majority but are less confident of their competency. Minority ethnicities, on average, produce significantly lower science literacy scores. With one exception (Middle Eastern), groups produce mean self-assessed competence ratings that are remarkably accurate predictors of their mean competence scores. The three socioeconomic conditions exert significant and unequal impacts across ethnic groups, with Hispanic, Middle Eastern and Pacific Islander data providing some unique results.

Keywords

social justice, numeracy, self-assessment, ways of knowing, ethnicity, gender, sexuality, assessment, science literacy, higher education

Cover Page Footnote

Rachel Watson is the director of the Science Initiative's Learning Actively Mentoring Program and the Director for the Queer Studies Minor at the University of Wyoming. Melding together science, humanism, feminism, and queer theory, Rachel's research interests focus on student learning assessment, active learning (action learning and problem-based learning) and social and environmental justice as they inform microbiology, biochemistry and environmental science curriculum design. For twenty years, Rachel has been the volunteer co-coach of the 13-time National Champion Nordic Ski Team and five-time head coach for Team USA at the World University Games.

Edward Nuhfer served as Director of Faculty Development and Educational Assessment and tenured Professor of Geology at four universities. His research interests are in metacognitive self-assessment, the role of the affective domain, and curricular design for reflective, higher-level thinking. He continues actively in writing, research, and assessment.

Kali Nicholas Moon graduated from the University of Wyoming, where she researched relationships between science literacy and self-assessment abilities of diverse groups. Throughout her studies, she

taught Life Science courses and played an active role in STEM faculty development. Kali now works in secondary education as a science teacher, applying and continuing her research on motivation, metacognitive self-assessment, diversity and inclusion, and science literacy.

Steven Fleisher is Instructional Faculty in Psychology at California State University Channel Islands. His expertise is in teacher-student relationships and instructional methodologies that support student autonomy and learning. His research focus is on metacognition, self-regulated learning, positive affective environments, self-assessment and reflective thinking, and the neurobiology of learning.

Paul Walter is an Associate Professor of Physics at St. Edward's University. His research interests in physics and science education include item response theory and developing tools for instructors to observe the transitions of their students' understanding.

Karl Wirth is an Associate Professor of Geology at Macalester College. His research focuses on metacognition, motivation, and undergraduate research experiences in support of best practices in teaching and learning in undergraduate STEM. As assessment coordinator for the Keck Geology Consortium, he seeks to improve undergraduate research experiences through the development of intentional curricular structures and mentoring practices.

Christopher B. Cogan is Instructional Faculty in Geography, Memorial University of Newfoundland, Canada. His research focuses on Geography, Environmental Science, and Geographic Information Science applied to biodiversity and teaching for exceptional learning. He was a researcher at the Alfred Wegener Institute in Germany, a member of the California State University design team for the Science Literacy Concept Inventory, and a winner of the best teaching award at CSU Channel Islands.

Ami Wangeline is faculty in the Biology Department at Laramie County Community College in Cheyenne, WY since 2008. She is the PI of an undergraduate research program where students conduct research on various aspects of fungal selenium hyperaccumulation engaging in authentic scientific practice from their first college semester. She also recently redesigned the General Biology for majors course to move from a content-based to a skills-based course in greater alignment with Vision and Change from AAAS.

Eric Gaze directs the Quantitative Reasoning (QR) program at Bowdoin College, is Chair of the Center for Learning and Teaching, and is a Lecturer in the Mathematics Department. He is the current President of the National Numeracy Network (2013 – 2015) and an associate editor of Numeracy. Eric has given talks and led workshops on the topics of Quantitative Reasoning course development and assessment.

Authors

Rachel M. Watson, Edward Nuhfer, Kali Nicholas Moon, Steven Fleisher, Paul Walter, Karl Wirth, Christopher Cogan, Ami Wangeline, and Eric Gaze

Introduction

In this paper, we try to understand how the experiences of a demographic group may produce a college-level educational performance that differs from the performances of the dominant majority group. The initial problem involves recognizing that substantial privileges afforded by membership in the dominant majority are real. Numeracy can uncover that reality, even when assumptions have hidden or obscured such privileges. The information we collected to help our understanding comes from measures of cognitive competence (knowledge) and associated self-assessed competence (people's feelings of confidence about knowledge) that are aligned tightly with our cognitive measures. We share our database in Appendix A included in the supplemental material linked on this article's metadata page.[1]

Our goals for this study are three-fold and aligned with our data collection strategies. First, we aim to confirm our earlier work (Favazzo et al. 2014; Nuhfer et al. 2016a, 2016b, and 2017) that the relationship between people's self-assessment and their actual competence is meaningful and that most people are adequate self-assessors. We use a competency measure as expressed by participants' scores on a special kind of multiple-choice test called a concept inventory. The Science Literacy Concept Inventory (SLCI) measures the ability to comprehend science as a way of knowing through conceptual thinking. It addresses the level of general citizen literacy with the reading ability of a high school graduate and tests thinking without advantaging those who possess significant rote knowledge of facts from any science discipline.

We derive measures of metacognitive self-assessed competence from participants' self-ratings expressed in a range from 0% to 100%. These self-ratings express affective self-perceived competence and are informed to varying degrees by awareness of actual competence and prior experiences from scores obtained through competency testing. In our recent work (Nuhfer 2015; Nuhfer et al. 2016a and 2017), we emphasized that a requisite for the validity of any such studies is that participants must clearly understand the challenges on which they are being asked to self-assess their competence. Both the measures of demonstrated competence and self-assessed competence must come from instruments that provide data of known reliability before one should attempt to interpret this paired data numerically. Our measures of self-assessed competence come from a Knowledge Survey of the SLCI (KSSLCI) and postdicted self-ratings of performance that participants express after having engaged with the

[1] https://scholarcommons.usf.edu/numeracy/vol12/iss2/art3/

entire Inventory. This paper is neither a science nor a STEM study, but instead, we use measurements of science literacy to furnish us direct measures of competence and metacognitive self-assessments of competence.

Our second goal is to employ numeracy's framework of reasoning to assess the existence of patterns of similarity or significant difference between *majority* and *minority* groups. In this paper, we use 'majority' in the numerical sense to refer to a group that makes up greater than half of the participants of any population category and 'minority' to refer to groups within the category represented by lower numbers. In our database, Caucasian whites are the majority race/ethnicity. Concerning gender, those who identify as binary male or female constitute the majority. Regarding sexuality, heterosexual participants formed the majority. Minority groups, by their very nature of being the minority, are often marginalized from the privileges that the majority group may not perceive and often takes for granted. Participants self-identified their affiliations with groups defined by gender, race/ethnicity, sexual orientation, and socioeconomic factors.

Finally, we aim to inform our interpretation of numeric data with both ethical and social justice ways of knowing and to describe intersectionalities and complexities in our data by considering the influences of three *socioeconomic* factors, which can extend or limit privilege. These measures are (a) status as a first-generation student, (b) status of English as a first language, and (c) expressed interest in majoring in a science. We use the term 'socioeconomic' to describe these three measures following the meaning established by the American Psychological Association (2018) in describing socioeconomic status (SES) inclusively: it "…encompasses not just income but also educational attainment, financial security and subjective perceptions of social status and social class." Socioeconomic is appropriate as an inclusive term that considers layers of privilege and disadvantage. Race/ethnicity is tightly tied to SES; in 2017, 8.7% of Whites lived below the poverty line whereas 18.3% of Hispanics and 21.2% of Blacks did so (U.S. Census Bureau 2017). Further, race/ethnicity is a determinant of the neighborhood of residence (Frazier et al. 2003). The place in which one lives affects access to education, employment, and services (Dreier et al. 2001). High SES is associated with higher achievement in language function tasks (Calvo and Bialystok 2014). Finally, Saw et al. (2018) show that low SES students are less likely to cultivate and retain an interest in science, technology, engineering, and math careers.

Self-Assessment, Self-Efficacy, and Self-Regulation

Although college graduation rates have increased for all students since the turn of the century (National Center for Education Statistics 2016), completion rates are

still very low, with just over half of students completing their degree within six years. This completion rate is even lower for marginalized minority students (Shapiro et al. 2017).

While there are many factors affecting college attrition and retention, one effort we can make is to teach student self-assessment. Integrating self-assessment exercises into classroom practices enhances student learning (Nicol 2009) while also improving student self-efficacy (Ross 2006; Keane and Griffin 2018). Self-efficacy, the strength of belief to develop one's capacity to achieve through effort and instruction, is a necessary component in obtaining an education. Unless students have confidence in their abilities to master their chosen subjects, they are unlikely to be able to either make wise choices when goal-setting or exercise the resilience and persistence needed to meet the inevitable challenges faced while becoming educated (Bandura 1997). For those minority students who have been historically marginalized, these challenges are more difficult, so self-efficacy is especially crucial. Ballen et al. (2017) find self-efficacy enhances the positive effect of active learning practices for underrepresented minority students in STEM courses, and helps close their gap in course performance as measured by grades and a Knowledge Assessment Inventory (a knowledge survey with an expanded Likert scale).

Improving self-assessment accuracy is a necessary component for developing self-efficacy, which involves increasing trust in one's own ability to learn. In turn, self-efficacy is a quality required for developing self-regulation. Self-regulation is a structured strategy employed for meeting one's learning goals. According to Pintrich (2004), self-regulation involves planning and activation, monitoring, control, reaction, and reflection. In planning and activation, learners set goals, identify their motivations, and plan for time and effort. In monitoring, learners become more metacognitively aware of their cognition (i.e., levels of understanding), motivation (i.e., extrinsic vs. intrinsic), and behavior (i.e., effort and use of time). In the control stage, learners select and adjust cognitive and behavioral strategies for learning and for managing their motivation. In reaction and reflection, learners make judgments and attributions about how they did and why. Engaging in these practices leads to improved learning and higher-order thinking.

Regarding self-assessment and self-regulation, McMillan and Hearn (2008, 40) state that "student self-assessment stands alone in its promise of improved student motivation and engagement, and learning. Correctly implemented, student self-assessment can promote intrinsic motivation, internally controlled effort, a mastery goal orientation, and more meaningful learning." Nicol and Macfarlane-Dick (2006) note that students already engage in assessing their academic work but do so less skillfully than they might if they had received instruction in strategies that inform their self-assessment efforts.

Blanch-Hartigan (2011) reports that understanding self-assessment and its improvement involves knowing the causes of student overestimation and underestimation. Teachers need to examine what is affecting self-assessment accuracy, i.e., "*how* students are being inaccurate and *who* is inaccurate" (p. 8, emphasis in the original). Accordingly, the alignment of self-assessment questions with the queries used for measures of performance is essential.

When instructors provide instruction to develop metacognitive monitoring (self-assessment of progress and achievement), metacognitive knowledge (understanding how learning works and how to improve it), and metacognitive control (changing efforts or strategies when needed), they help students to meet their aspirations of becoming educated (Dunlosky and Metcalfe 2009). Helping students who have been marginalized to gain awareness of motivation through cultivating self-assessment skills may help them identify their areas of greatest interest.

Teaching self-assessment is necessary to facilitate student self-efficacy and thus student retention. However, researchers have questioned the value of student self-assessment measures; the hypothesis that students who are least capable are those who are worst at self-assessment has dominated the literature (Mabe and West 1982; Falchikov and Boud 1989; Kruger and Dunning 1999; Dunning et al. 2003; Dunning 2011; Ehrlinger et al. 2008; Bell and Volckmann 2011; Dunning 2013, Ehrlinger and Shain 2014; Webb and Karatjas 2018; Anson 2018). This belief has slowed efforts to teach student self-assessment and promote student metacognition.

Ways of Knowing

In order to achieve our goals, we have blended three ways of knowing: numerical, ethical, and social justice. We have further used measures of understanding science as a way of knowing and measures of metacognitive self-assessed competence of such knowing as keys to discovering differential effects on the education of marginalized minority groups. We regard understanding and addressing these effects as a *wicked problem* (Kolko 2012), and thus we assemble a multidisciplinary cohort of authors with diverse ways of knowing to engage with it.

A *way of knowing* is an epistemology that develops over historical time through the collective contributions of successive generations of experts. Aspiring learners must realize that ways of knowing begin in response to a need and persist through historical time because they provide value. Learners can develop an understanding of diverse ways of knowing the world, themselves, and others by directly engaging with these ways cognitively, metacognitively, and affectively. Ways of knowing are deeply rooted and inextricably tied to identity. Cultural and

indigenous ways of knowing are the platform on which individuals build their worldviews in relational ways (Bang et al. 2007; Romm 2017). Women's ways of knowing develop in response to their experience within gendered learning environments where their positionality may prevent or affect their knowledge construction (Belenky et al. 1986; Hayes et al. 2000). We acknowledge these particular ways of knowing here because they exemplify the interplay between epistemology and self-assessment. Ways of knowing incorporate both competence and self-assessed competence. Flannery (2000, 78) puts this well when considering women's ways of knowing: "…identity and self-esteem are intertwined with learning, unlearning and relearning who we are and how we value ourselves."

Science is a metadisciplinary way of knowing that explains the physical world through testable knowledge. Other metadisciplinary ways of knowing valued in Western scholarship are the ways of the social sciences, arts, mathematics, humanities, and technology. We authors have all experienced students who majored in one area and then commonly ceased to seek the value of other ways of knowing, often wrongly presuming that the others have lesser value. When these students must make the greater leap to understand the value in a cultural way of knowing that is neither part of their education nor their life experience, many struggle. A teacher's presuming lesser or no value for an unfamiliar way of knowing curtails first the teacher's effort, engagement, and persistence that are needed for the teacher's understanding. Thereafter, this presumption communicates the teacher's misunderstandings back onto affected individuals or groups of students and curtails their learning. If the misunderstanding borne from presumption becomes culturally widespread through a majority group, the result inflicts epistemicide on ways of knowing unique to minority groups and removes them from the culture.

Ethics and Numeracy as Ways of Knowing. The incentive for this paper begins with two ethical concepts of autonomy and beneficence. Every participant in our study has already exercised the concept of autonomy by actively seeking the benefits of gaining an education from a college or university. Our study seeks to enact beneficence by learning how to aid participants to succeed in meeting their aspirations.

We employ *numeracy* as a way of knowing produced by reasoning through the language of mathematics. Numeracy's way of knowing is particularly suited for detecting symptoms that may be detrimental to meeting educational aspirations. Like the ways of knowing of science (Nuhfer et al. 2016b) and ethics (Anderson and Handelsman 2010), numeracy's way of knowing employs recognized concepts. Gaze et al. (2014) list the foundational concepts of numeracy as number sense, reading and interpreting graphs, basic probability and statistics, and reasoning. Numeracy serves us for identifying the quantitative

variations of measured cognitive competence and self-assessed competence between groups and within groups. Our study compiled a large dataset consisting of direct measures of these paired variables across categorical groups with known socioeconomic conditions. Such data offer unique opportunities to discover where socioeconomic conditions influence competence and self-assessed competence. When a socioeconomic condition significantly lowers either, we will refer to the measured quantity of lowering as a penalty that arises merely from the consequence of being in a group category exposed to a detrimental socioeconomic condition.

The way of knowing of ethics employs the ethical principles listed in Anderson and Handelsman (2010): *justice* (strive to treat every person fairly), *beneficence* (do good), *nonmaleficence* (do no harm), and *autonomy* (respecting the value of each individual and preserving their power to make life choices). An adequate ethical assessment is only achievable through simultaneous consideration of all four principles. We employ ethics to help us to describe how particular groups are affected. The ethical principles of justice and autonomy bear obvious relationships to numeracy. Justice is invariably described through variants of "equals should be treated equally and unequals unequally." However, the terms *equal* and *average* are fraught with danger when a language of description becomes taken as a language to justify particular actions to reduce inequality by enforcing compliance toward presumably desirable norms of an "average person" (Warner 1999; Rose 2016; Oughton 2018). Autonomy offers a balancing principle by articulating that an individual or a small minority group of individuals retains rights over self and rights of choice for self-determination. Mere status of existence as the majority group confers no rights to act on a person or a minority group by suppressing diversity, violating autonomy, or exercising maleficence.

When numeracy detects a difference that designates specific individuals or groups as less successful cognitively or less confident in meeting their aspirations, ethics directs us to understand whether forces or circumstances that we can control may be aiding their success (enacting beneficence) or disadvantaging individuals or collective groups (violating nonmaleficence) in meeting their aspirations.

The numerical difference that we employ in this paper to define two groups as different (or unequal) is the difference between their two means. A *statistically significant difference* in the language of numeracy expresses how large the difference between the means of two groups must be before we can reject the hypothesis that the relationship between two groups is primarily one of equality (sameness) rather than primarily one of inequality (difference).

Both ethics and numeracy recognize differences between groups and ranges of differences (variances) within groups. Ethics recognizes that each group consists of unique and autonomous individuals. The quality that designates an individual as a member of a defined group expresses only a minor part of the complexity that characterizes individuals. While numeracy employs a measure to detect individuals' groups as significantly different from one another, ethics employs considerations of beneficence and nonmaleficence to describe whether the difference reveals something harmful or beneficial to a group or individuals within a group. Destructive actions and discourse are not the products of using numerical or ethical reasoning frameworks but rather are the products of their inappropriate use.

Social Justice as a Way of Knowing. Social justice offers a third way of knowing that is informed by many different traditions and can be defined, practiced, and theorized in many different ways but that has value for taking decisive action. Bell (2016) describes social justice as the process that promotes the goal of inclusive affirmation of human agency through the democratic, respectful, and collaborative work of diverse others. They describe a socially just world as "a world in which the distribution of resources is equitable and ecologically sustainable, and all members are physically and psychologically safe and secure, recognized and treated with respect." Fook and Goodwin (2018) trace the historical pluralizing of the meaning of social justice, noting the addition of a second dimension in the late twentieth century. That is, to the original ethical focus on the unfair distribution of material goods, social justice added a focus on the recognition of how institutions privilege some groups and oppress others. From these latter traditions was born a recognition of power systems that sustain the oppression of the marginalized and affirm privilege.

Social justice is sometimes misunderstood as merely reactive or political rather than as a valid framework of reasoning and way of knowing (Guest et al. 2009). Guest et al. provide a list of traits that embody unifying concepts that are easily understood and particularly applicable to academic environments. These concepts include equality of rights and opportunities, provision for basic needs, and treatment with respect and safety—including safety from discrimination. The last two concepts are pertinent to a particular concept of social justice that we find especially relevant to this paper: 'Othering.' Othering describes a process of marking others as inferior so as to establish dominance (Chow 1989; Weis 1995; Donovan 2000; Fisher 2015; Merriam-Webster 2018), the process of marginalizing (James 1998), or the absence of functional empathy (Canales 2000). Othering is key to the discursive development of the subservient (Weis, 1995).

For centuries, writers such as Jefferson and Banneker (1791) and Nieto (1998) have called attention to the social tendency to stereotype racial/ethnic

minorities as having lesser intelligence. Likewise, Kosciw et al. (2004) explain how LGBTQ+ (lesbian, gay, bisexual, transgender, queer, with + denoting other genders, sexualities, and sexes and with + indicating our awareness that no acronym can be adequately inclusive of all identities [Scheller-Boltz n.d.]) also have been marginalized and associated with terms that denote lesser intelligence.

Donovan (2000) calls attention to the way in which society associates women with the affective domain, the emotional, and the irrational. Society associates marginalized groups with particular traits, and affective traits themselves, like the people with whom these traits are associated, are Othered. For example, "being emotional" is seen as less valuable than "being rational." Damasio (1999) relates how the emotional-affective domain itself was devalued in educational discourse and relegated to lesser animalistic qualities. Thus, student self-assessment, which provides a path to the development of metacognitive capacity by understanding the affective self, has been historically Othered.

Othering produces effects that numeracy detects as measures of difference expressed in lowered performance and lowered confidence. Ethics could be employed to describe these differences as mostly maleficent through the evidence that the effects are detrimental to particular individuals or groups. We can use social justice to illuminate the cause of detriment by noting how the privilege of membership in a majority group can, even without malicious intent, trigger Othering that harms the success of minority demographic groups.

Intersectionality

All individuals and groups construct an identity that includes many demographic descriptors in response to a network of inextricably linked systems of power (Mitchell 2014). *Intersectionality* describes these crossroads and acknowledges that many categorizations impact each person and group: race/ethnicity, sexuality, gender identity, ability, and others such as socioeconomic status. We cannot reduce individual identity or group identity construction to a single descriptor, and we must understand privilege and oppression as deriving from plural causes.

Othering is particularly attractive as a model for working within the complexity of intersectionality because it describes actions, situations, and behavior without singling out specific groups or using accusative descriptors that trigger detrimental reactions, such as defensiveness and disrespect. All of us internalize socially constructed Othering. Our disrespecting others or experiencing disrespect from others short-circuits our ability to remain aware that alternative, unfamiliar ways of knowing offer something of value. Othering is perhaps the arch-nemesis of drawing on the collaborative brain-power required to address the challenges of a wicked problem.

Methods

Measuring Demonstrated and Self-Assessed Competence

Demonstrated Competence Measures. We measure demonstrated competence using the Science Literacy Concept Inventory (SLCI—Nuhfer et al. 2016b). The SLCI consists of twenty-five items that map to twelve concepts assessing participants' comprehension of science as a way of knowing at the level of college undergraduates. The SLCI collects the following demographic data: binary gender identification, race/ethnicity, numbers of college science courses completed, academic rank (first-year college student through professor), and three measures of socioeconomic conditions: (1) status as a first-generation student, (2) an interest or lack of interest in majoring in a science, and (3) having as a first language the language of the measuring instrument (English). The SLCI National Database that we employed contains data from 24,701 participants.

Self-Assessed Competence Measures. We use a separate Paired-Measures Database consisting of 3,323 participants. It contains the SLCI data, but also adds, from each participant, one or more ratings of self-assessed competence. A knowledge survey of the SLCI (the KSSLCI) provides a self-assessment rating calculated from each participant's rating of her or his ability to respond to each of the SLCI's 25 items according to:

> A. I can fully address this item now for graded test purposes.
> B. I have partial knowledge that permits me to address at least 50% of this item.
> C. I am not yet able to address this item adequately for graded test purposes.

We refer to the collective KSSLCI rating as a *granular* self-assessment because it derives from an item-by-item, detailed self-assessment. The Paired-Measures database contains 1,825 granular self-assessments made from the KSSLCI ratings.

We also employ single questions that ask for *predicted* score estimates made from just a description of the SLCI before seeing the instrument (N = 2,465) and *postdicted* global estimates (N = 3,323) recorded after taking either the KSSLCI (N = 1,824), the SLCI (N = 2,810), or both. We refer to self-assessments made from single questions as *global* self-assessments. The global *postdicted* self-assessments of performance completed after taking either the KSSLCI or the SLCI itself correlated with the SLCI scores to degrees comparable to the scores computed from the granular ratings derived from the KSSLCI. Correlations between self-assessed competence and actual competence were positive and significant, between r = 0.5 and 0.6 for participant-by-participant paired measures

and greater than $r = 0.7$ for the twenty-five item-by-item paired measures. (See Appendix B included in the supplemental material linked on this article's metadata page,[2] Fig. B-1.)

The Master's thesis of Nicholas-Moon (2018) made a unique contribution to this Paired-Measures Database by adding paired measures from over 850 participants and by extending collection of demographic data into groups defined by self-identified sexuality and non-binary gender. In two separate questions, respondents were asked to 1) choose: heterosexual, gay/lesbian (homosexual), queer, bisexual, or other and 2) choose: female, gender queer, intersex, male, transgender, female to male transgender, male to female transgender, or other. However, so few respondents identified as gender queer, intersex, or transgender that we are unable to present any analyses.

The construction of the paired instruments, the conditions under which they collect data, and the psychometrics of both are already detailed in Nuhfer et al. (2016a; 2016b; 2017). We do not repeat that detail here. Our current Version 7 of the SLCI contains the 25 multiple-choice items plus two global self-assessment queries. Thus, it now collects paired measures of demonstrated and self-assessed competence in one instrument.

To describe self-assessment skill of individuals, we use the terminology following the taxonomy of Nuhfer et al. (2017). *Good* self-assessment accuracy is a self-assessed competency that lies within ±10 percentage points (ppts) of demonstrated competency. *Adequate* self-assessment accuracy is over ±10 ppts but less than or equal to ±15 ppts. *Marginal* self-assessment accuracy is over ±15 but less than or equal to ±20 ppts. *Inadequate* self-assessment is inaccuracy greater than ±20 ppts inaccuracy but less than or equal to ±30 ppts, and *extreme* self-assessment inaccuracy defines misestimates as greater than ± 30 ppts.

Nuhfer et al. (2016a) noted the difficulty of using language to describe the precise nature of *what* is being reported through the numerical expression of self-assessed competence. There is no doubt that an individual's self-assessed rating of competence is a felt affective response, but this affective feeling is one largely *informed* by the cognitive knowledge and experience related to the specific content relative to the challenge being assessed. However, when we aggregate groups of people with backgrounds of different privilege, the collective rating expressed in the group's mean self-assessment rating may offer a stronger reflection of a general state of *confidence* to address the challenge that has been formed by that group's experiences of past success as influenced by its history of socioeconomic privilege. In everyday informal use, it is tempting to describe self-assessed competence with the term 'confidence.' However, we try to restrict the

[2] https://scholarcommons.usf.edu/numeracy/vol12/iss2/art3/

use of ‘confidence’ to when we are describing the mean self-assessment ratings of groups. The distinction between ‘confidence’ and self-assessed competence is one of focus. We see ‘confidence’ as the intensity of belief of general feelings of competence, but self-assessed competence as the strength of belief to address a specific challenge with present capacities.

Data Analysis. We archived our data in Microsoft Excel spreadsheets and analyzed the data using Statistical Analysis System Institute’s JMP software for all data processing (reliability, correlations, ANOVA, Item Analyses according to Item Response Theory, and modeling). We used Adobe Illustrator to improve the quality of graphs rendered by JMP.

Our analyses in this study involve a balance of lumping and splitting of groups of interest. We begin with an initial lumping of participants into two parts comprised of an ethnic *majority* white Caucasian group and a composite ethnic *minority* group consisting of all who self-identified as other than white Caucasians and those who self-identified as other. This creates just two groups for an initial examination. Although ‘other’ creates a nebulous category, selecting it indicates a clear preference not to identify as a member of the majority.

In examining self-identifications of sexual orientation, we considered both LGBQ and 'other' as minority groups, but we decided against lumping the small populations of respondents with 'other' when we examined this specific data. Although ‘other’ is a clear expression of self-identification as outside the majority binary sexuality group, it reflects with equal clarity a choice to not self-identify with the alternative categories offered. It is possible that some individuals responding as 'other' would have marked a ‘questioning’ category if it had been provided. Snyder et al. (2018) confirm that respondents in this category may show a different trend than gay and bisexual identifying students.

Splitting involves subdividing the two major classifications and studying the groups of interest that lie within these. The limits to which we can carry splitting occur when a particular group of interest does not have enough data to reach interpretations that are likely going to prove reproducible in future studies. We deduce that a group may be “too small” by two tests.

(1) The data is insufficient to achieve reliability comparable to the reliability that we know is achievable from the large datasets collected by these particular instruments (from Nuhfer et al. 2016a; 2016b).

(2) The SLCI item-by-item order of difficulty collected from any group in our study of 3,323 paired measures seems inconsistent with the item-by-item order of difficulty established from Item Response Theory analysis of the SLCI (see Appendix B) of the National Database (24,701 participants).

Item Response Theory Analysis. Item response theory (IRT) allows for measuring latent traits, such as ability, based on participants’ performance on a

test. In Appendix B, we use IRT to display the registered differences in response to each of the 25 items by the demographic populations of interest.

Instead of a raw test score, IRT reports its equivalent rating as an *ability formula score* for each participant that commonly ranges from -3 to 3 instead of the traditional 0% to 100%. The lowest negative value expresses the lowest measured ability. The ability formula score depends not simply on an individual's getting an answer right or wrong but examines the entire populace that took the test and determines the individual's score both from counting the number of items answered correctly and also from weighting each item according to its proven difficulty and how well it contributes to discriminating high-scoring from low-scoring participants.

Nuhfer et al. (2016b) utilized IRT to validate the SLCI, but they published their paper by describing the results in terms of scores as percentages, percentage points, and classical reliability measures. The classical test analyses better meet the readership of educators who evaluate their classes in terms of percentage scores on tests rather than in terms of IRT criteria. However, Item Response Theory is a more robust way of evaluating a test and its items. Therefore, in this paper, we include in Appendix B a summary report of the SLCI. There we compare the demographic groups' relationships on an item-by-item basis, expressed as IRT parameters.

We ordered the tables in supplemental Appendix C in parallel with the presentations of figures in the main paper. These materials are supplemental and are not required for understanding the main paper. However, readers seeking more detail of our findings will find it there. In this paper, we have named the particular tables in Appendix C that correspond to sections of the text to make finding the corresponding table less onerous.

Results

In reporting these results, we have two purposes. One is to report our results on the study of demographic groups, the primary subject of this paper. The other is to confirm that our earlier work (Nuhfer et al. 2016a; 2016b; and 2017) remains replicable given newer, larger databases of (1) SLCI competency measures with 24,701 participants and (2) a subset of these scores (N = 3,323 participants) with paired data of SLCI competency scores and self-assessed competency ratings.

Our first step is to examine how representative our Paired-Measured Database is of the larger National Database. Our smaller database of paired measures draws from a somewhat different geographic region than the National Database, with the former having a significantly greater representation from

Wyoming institutions of higher education. A comparison of our two databases (Appendix C[3], Table C-1 in columns 8 and 9) reveals that the most striking differences is that our Paired-Measures Database has relatively fewer minorities (notably, fewer Hispanics), first-generation students, and students with English as a learned language, but a higher percentage of students interested in majoring in science.

The mean SLCI scores for the Paired-Measures Database are 2.8 ppts higher than the National Database. As we shall see, this difference is explainable by the different representations of demographic groups in the compositions of both database populaces. Although the two databases have some significant differences in demographic makeup, the two give similar overall results on the SLCI with a correlation of mean scores across equivalent categories (Table C-1, Appendix C, columns 4 and 7) of $r = 0.92$. The mean scores on each of the 25 SLCI items are very similar when rendered from the separate populaces of both databases, with a correlation of $r = 0.995$.

Demographic Results: Ethnic Majority versus Minority

White Caucasians constitute a majority ethnic group in our studied populace. *Majority* and *minority* are expressions with distinct quantitative meanings. Numeracy can clearly designate apparent privileges that accord with 'majority privilege,' but it seems unsuited to evaluating these designations as 'White privilege.' Therefore, we reference such privilege as 'majority privilege' wherever our evidence best supports that term.

Initially examining our study population for any statistically significant differences between these two broad majority and minority groups offers the statistical advantage of analyzing demographic groups with the largest number of participants. Thus, we begin by examining whether we can detect specific differences between the group consisting of majority white Caucasians and an aggregate minority group that consists of all those who self-identified their ethnicities by expressing non-affiliation with the Caucasian majority (African-American, Asian, Hispanic, Middle Eastern, Native American, Pacific Islander, and other). If we can do so, we can then attempt to drill down into both groups to try to discover what is occurring within the smaller groups that comprise these two.

Being members of the minority group incurs a highly significant penalty in competency (SLCI) scores that, on average, consists of 6.8 ppts for race/ethnicity minorities in our National Database. (See Fig. 1 and Table C-1, Appendix C).

[3] https://scholarcommons.usf.edu/numeracy/vol12/iss2/art3/

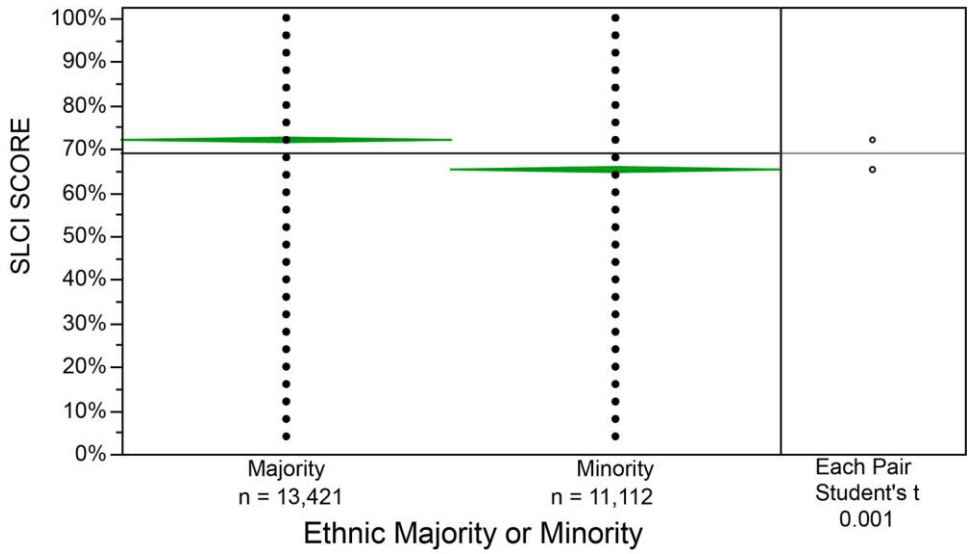

Figure 1. Categorical plot of racial/ethnic majority and minority scores on the Science Literacy Concept Inventory from the National Database. On this and all similar categorical graphs presented in this paper, the position of the mean for each group lies within the vertical center of the green diamonds. That the mean values lie within the diamonds is at greater than 99.9% confidence. The diameters of the small circles shown in the box on the right side of such graphs are the heights of the corresponding diamonds. Vertical distances separating the circles reveal the statistical degree of significant difference between the means of the groups graphed. The differences shown here are highly significant at $P < 0.001$.

Our smaller Paired-Measures Database also confirms a parallel competency penalty (Fig. 2) for those in the racial/ethnicity minority group, which although larger (8.5 ppts) is consistent with minority status reflecting significantly lower mean competence scores. However, Figure 2 allows something unique to this study; it allows us to examine whether these differences in cognitive competence are related to significant differences in affective feelings of confidence as expressed by ratings of self-assessed competence.

Panel 2-A in Figure 2 shows that the lower competency scores recorded by the minority group in Figure 1 mirror the lower confidence ratings produced by the group. Further, both groups' averages of their self-assessed competence are very close to their average demonstrated competence.

Panel 2-B taken from the paired measures of the SLCI scores and the granular self-assessments produced by the knowledge survey (KSSLCI) reveals just how strongly the collective self-assessments of groups correlate with their actual demonstrations of competence. The group self-assessments reveal a collective felt difficulty about each of the twenty-five items that closely parallels the difficulty confirmed by the relative scores both groups separately achieved on

each item. Both minority and majority groups achieve high correlations between self-assessed competence and demonstrated competence.

The minority group scores lower on every one of the 25 Inventory items and consistently expresses lower confidence than the majority group on every item. The aggregate felt differences in item-by-item difficulty are consistent between the two groups, with a correlation of $r = 0.92$ (not shown in a figure) for the KSSLCI self-assessments by the majority and minority groups.

The two panels in Figure 2 are consistent with the results of Nuhfer et al. (2016a; 2017) and are devastating to claims that people's self-assessments are meaningless random noise, that students cannot self-assess, and that knowledge surveys cannot be recommended as instruments to assess achieved competence. Instead, aligned paired measures of self-assessed competence and demonstrable competence could be the most important of all measures to take. They reveal that groups with reduced ability to demonstrate competence actually do feel less competent than groups that register higher competence, although individuals within any group are probably largely unaware of this.

As we proceed next to examine these groups in further detail, it is good to consider that we have already documented that lack of privilege could affect both performance and affective feelings of confidence to perform. Affect and cognition seem inseparably involved in learning.

Demographic Results: Effects from Socioeconomic Conditions

We have confirmed above that minority ethnicities, as a collective, score lower and are less confident as a whole about their understanding of science as a way of knowing than the Caucasian majority. Next, we examine whether our data can provide any factors that numeracy's way of knowing could designate as possible causes for the observed difference.

Nuhfer et al. (2016b) show that the three socioeconomic conditions of (1) status as a first-generation student, (2) having no expressed interest in majoring in science, and (3) having English as a learned language all produce significantly lowered mean competency scores as registered by the SLCI. We find that the effects of lowered competency scores associated with these three socioeconomic factors are replicated again here in this paper from our National SLCI Database and our Paired-Measures Database (Table C-1, Appendix C).

In this paper, we drill deeper to study whether these factors could inequitably affect the collective self-assessed competence (a measure of confidence) of demographic groups. The factors, on average, influence self-assessed ratings between the demographic groups at high levels of significance ($P < 0.001$, Table C-2, Appendix C) and to almost precisely the same degree as they influenced the measures of demonstrated competence. In fact, the mean self-assessed ratings for

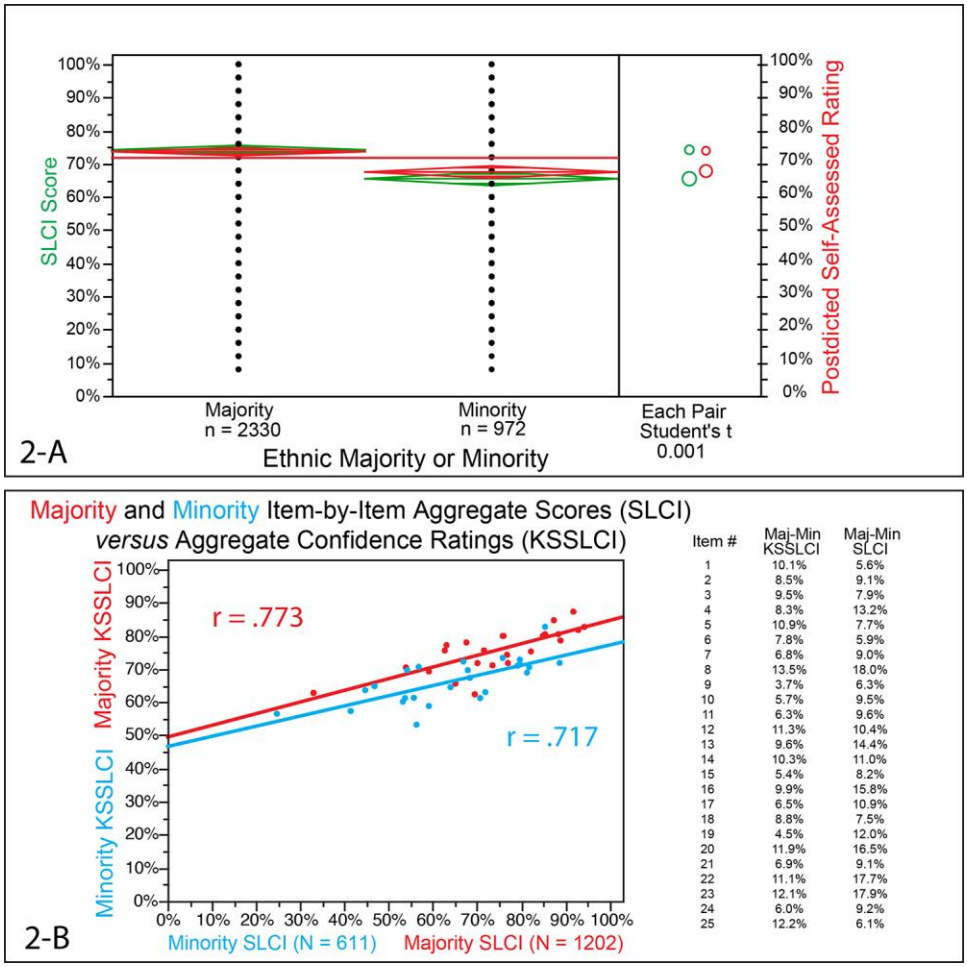

Figure 2. Comparisons of mean self-assessed competence as group confidence ratings and competence demonstrated by test scores by race/ethnicity majority vs. minority status. Data comes from the smaller Paired-Measures Database. Panel 2-A plots the data categorically and employs the self-assessed ratings from the postdicted global items and the demonstrated competence measure from the SLCI scores. The majority group, on average, was more accurate in its self-assessment, whereas the minority group was significantly less confident, and slightly overestimated its actual scores. Panel 2-B displays a scatter plot of item-by-item mean confidence vs. competence measures. The self-assessed confidence ratings come from the Knowledge Survey of the Science Literacy Concept Inventory (KSSLCI). Each dot, color-coded as the majority (red) and minority (blue) groups, represents the respective group's aggregate confidence ratings (KSSLCI) plotted against the group's aggregate scores (SLCI) on each of the 25 items addressed in common by both instruments. The confidence ratings and competence scores were higher for the majority group on every item (data columns right side of Panel 2-B). The trends in response to both instruments were similar for both groups with an item-by-item correlation of $r = 0.92$ for the KSSLCI ratings and $r = 0.97$ on the SLCI scores between the majority and minority groups. Subsequent figures in this paper reveal how applying confidence–competence measures with reliable instruments provides valuable information through which to assess and understand learning.

all six groups were only about 1 ppt different from their actual mean competency scores (Table C-2, Appendix C, Column 5). The results in that Table C-2 further confirm the observations by Nuhfer et al. (2016a, Fig. 10) that, given reliable data derived from well-aligned, paired instruments, the collective self-assessment by a group of people of their competence provides a remarkably accurate estimate of a measure of the group's average competence.

Demographic Results: Effects of Binary Gender

We employ tests of statistical significance of difference between means of different categories of populations. Unless otherwise noted within the text or tables (see Appendix C for details), we use $P < 0.001$ to define a "significant difference." In Nuhfer et al. (2016b), we explain why we employ this high bar of $P < 0.001$ or 99.9% confidence to try to ensure that the findings we report should continue to prove reproducible. We are fortunate to be able to employ a high standard because our studied populace yielded a very large database, and our instruments with which we measure documented competence and self-assessed competence are aligned and have good reliability (Nuhfer et al. 2016a). However, this does not mean that an investigator will not find important significant differences on a small population such as a sample taken from a few classes on a single campus. Demographics differ between institutions, and they do produce data that shows significant differences. However, comparisons of different campuses are not our focus in this paper, but are the focus of our paper in preparation.

The binary categories of male (men) and female (women) across the entire study populace offer as broad an overview as that offered by the categories of majority and minority race/ethnicity. The differences in mean SLCI scores between those who self-reported their gender identity as male as opposed to female on both the National and Paired databases is not statistically significant at $P < 0.001$. Likewise, we rarely found statistical difference in the responses between men and women in many subgroups of the population such as those who were first-generation students or those who were not; those who expressed an interest in majoring in science and those who did not, and those for whom English is a first language. However, among those for whom English is not a first language, women and men were significantly different, with women outscoring men within this socioeconomic category on average by about 2.8 ppts on the National Database (Table C-3, Appendix C). We attribute failure to detect this difference in the Paired Measures Database to the small number of participants whose first language is not English in the smaller database.

Nuhfer et al. (2017) paired SLCI scores and self-assessment ratings from the KSSLCI for 371 men and 664 women and observed that women were more accurate in self-assessment than men. In this paper, we paired the SLCI scores

and postdicted global self-assessment ratings of 2,101 women and 1,204 men. Our results replicate the reported observation. While the 1.4 ppts difference between men's and women's SLCI scores proved not to be a significant difference, the difference between women's and men's self-assessed scores of their performance on the SLCI was highly significant. The self-assessed average estimated by women was only 0.8 ppts below their actual mean score of 71.2%, whereas men overestimated their mean score of 72.6% by an average of 2.8 ppts (Fig. 3).

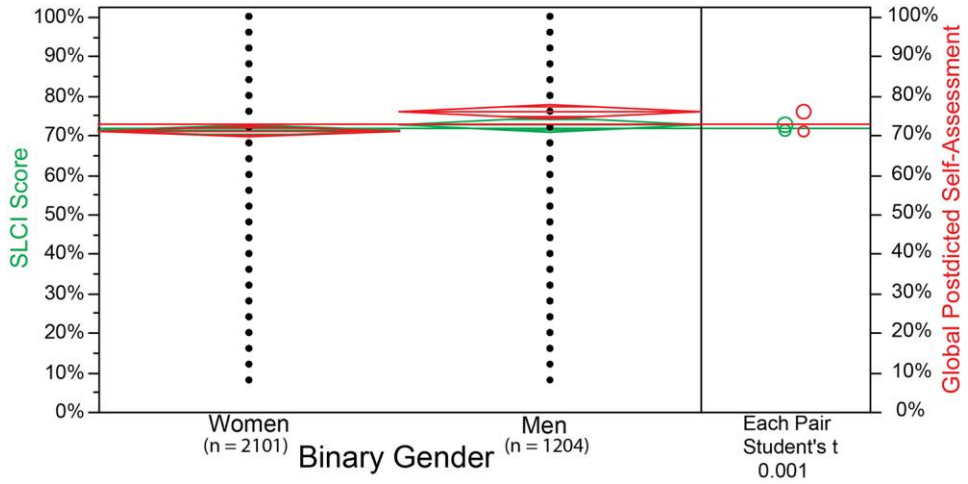

Figure 3. Categorical plot showing women's and men's demonstrated and self-assessed competence (confidence). Demonstrated competence measures come from SLCI scores (green) and self-assessed competence measures come from global postdicted self-assessed competency ratings (red). All data is from the Paired-Measures Database. Colored horizontal lines reveal the grand means for each measure. While there is no highly significant difference between women's and men's competency scores, the mean self-assessment score differences are highly significant at $P < 0.001$. Collectively, women assess their competence with minimal error, whereas men significantly overestimate their competence.

Demographic Results: Influence of Ethnic Affiliation

Next, we drill deeper into the data to learn whether there are variations by ethnicities that we cannot see when all are lumped together within the broad categories examined above. We have two databases: our smaller Paired Measures Database, and our National Database that contains only SLCI scores and no self-assessment measures. Some ethnicities are represented by very small numbers in our Paired Measures Database. We would like to use our Paired Measures Database as much as possible because it contains the additional information of participants' self-assessed competence. However, we need to learn how representative that smaller dataset seems to be of the National Database. We first do so by comparing the measured competencies by SLCI scores of the ethnicities across both databases (Fig. 4).

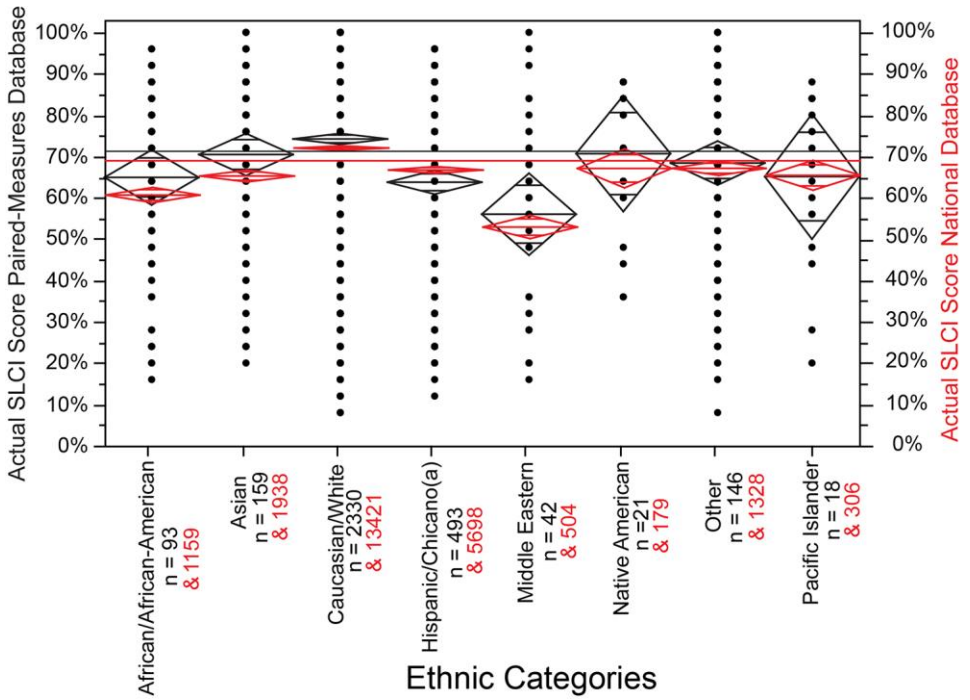

Figure 4. Categorical plot that compares mean competency scores for two databases by ethnicity. The competency means come from SLCI scores. The relative variations between ethnic categories are similar in both databases, but the grand mean scores (shown by colored horizontal lines) are lower in the National Database (red) and generally lower for most groups than in the Paired-Measures Database, except for the Hispanic ethnicity. Numbers of participants represented in each database appear with category labels along the ordinate. Confidence that the mean lies within the vertical heights of the colored diamonds is 99.9%.

The means are lower in the National Database than in the Paired-Measures Database across nearly all ethnicities, and we know the exact positions of the true mean values less precisely and with less confidence in the Paired-Measures Database for each ethnicity than we do in the National Database. Yet, the relative variations between ethnic categories prove remarkably consistent, even for ethnicities like Pacific Islander and Native Americans with tiny representations in the Paired-Measures Database. The correlation of the mean scores across respective ethnicities from both databases is $r = 0.89$. In short, the smaller Paired-Measures Database, which is one-seventh the size of the National Database, offers a good representation of the national data.

We can next learn how closely the means of self-assessment competency track with the measures of actual competency from our paired measures (Fig. 5). Except for Pacific Islanders, mean self-assessed competencies are lower than those of the Caucasian majority. All minority ethnic groups that comprise the collective minority group (depicted in Fig. 2) generate lower mean SLCI scores

than Caucasians, and their corresponding collective self-assessments indicate that they have some collective awareness of their disadvantage. Only one ethnic group (Middle Eastern) significantly overestimated its mean competency scores.

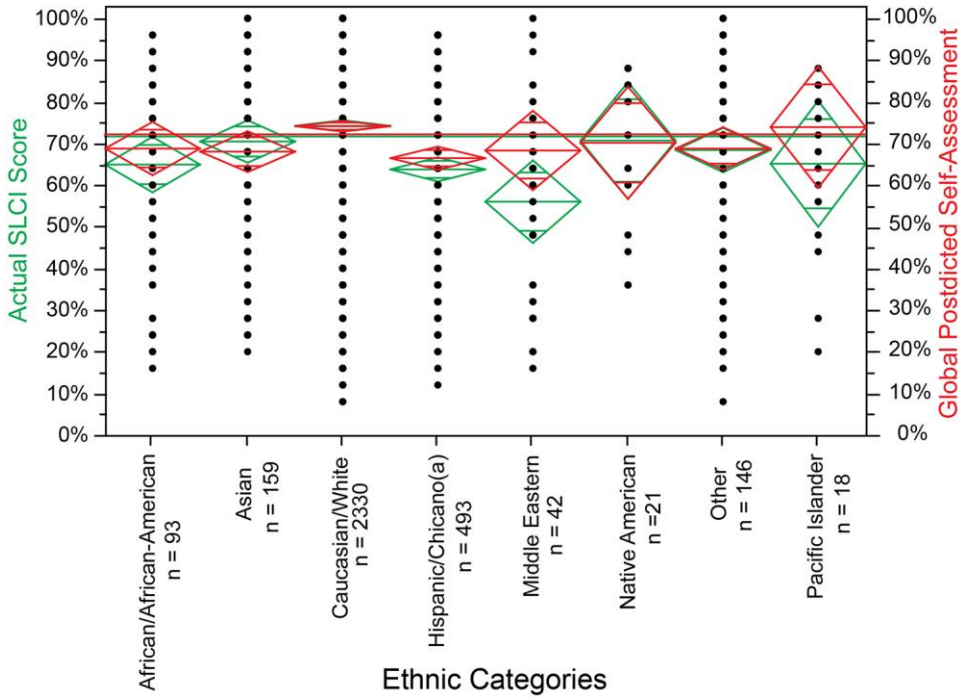

Figure 5. Categorical plot of competence and confidence by ethnicity. Data comes from the Paired-Measures Database with demonstrated competence measured from SLCI scores (green) and self-assessed competence (confidence) measures coming from postdicted self-assessment ratings (red). Numbers of participants within each category appear with labels along the ordinate. The means of each category are known to be within the red and green diamonds at a confidence of 99.9%.

Figures 4 and 5 confirm that ethnic minority groups score differently from the Caucasian majority and from one another. Different minority groups are not uniformly affected by influences that cause reduced test scores and confidence. Different factors may be at work in affecting different minority groups, or a few factors may be impacting all the minority groups but not by equal degrees.

Recall that, overall, the binary genders revealed no significant competence difference but significantly different self-assessment abilities. We now examine whether these gender trends behave the same within or across ethnicities. We can use the larger National Database to examine competence, as measured by the SLCI, to compare the genders across ethnicities (Fig. 6).

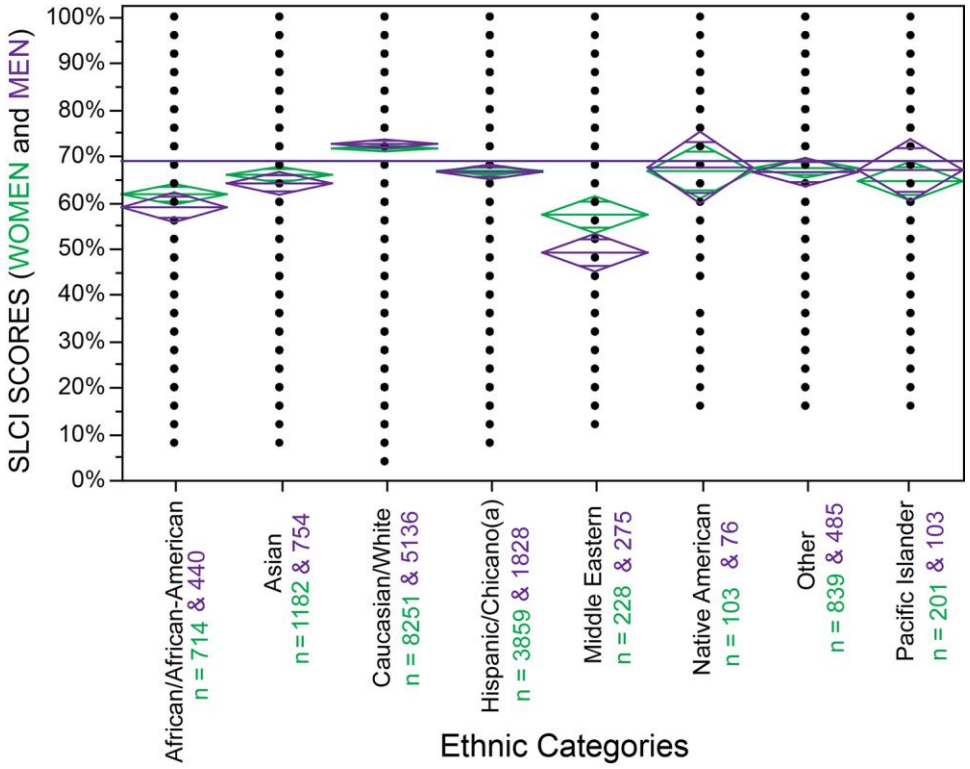

Figure 6. Categorical plot of binary genders' SLCI competency scores by ethnic category. Women (green) and men (violet) within the same ethnic group have similar mean competence scores, except for the Middle Eastern ethnic group, which shows significant binary gender differences in mean SLCI scores at the $P < 0.001$ level.

We see that women in the ethnicities African American and Asian outscore men, whereas in the Caucasian majority men very slightly outscore women. However, the only group in which these gender score differences are significant ($P < 0.001$) is the Middle Eastern ethnicity where women vastly outscored men.

Next, we can add self-assessment data as we examine whether the three socioeconomic conditions that we collected can help us to understand the differences that we see within our demographic groups. Self-assessment data consists of a valid self-assessment signal and random noise, and for this reason, paired measures of competence-confidence like these require critical masses of data to tease out an interpretable self-assessment signal from the noise (Nuhfer et al. 2016a). The amount of data these researchers recommend is 400 paired measures because that amount clearly depicts pure random noise on a Kruger-Dunning type graph when the worst case occurs—that the data consists of nothing but noise. It is essential to know what pure random noise looks like on all graphical conventions used in the peer-reviewed literature of self-assessment so as not to interpret the patterns of noise as the products of human behavior. Some

researchers have done so. At least three of the four case studies presented in the 1999 seminal paper on self-assessment (Kruger and Dunning, 1999) portray noise rather than human self-assessment data, and all four cases are done on small paired-measures databases not nearly approaching the 400 participants recommended (Nuhfer et al. 2016a; 2017).

When a self-assessment signal exists in the data, one may get some information worth interpreting from less data than 400 paired measures. Small data sets can be checked for any interpretive value by (a) determining whether the data in the set collected from both instruments have reliability of $R > 0.7$, and (b) determining whether the item difficulty for the 25 items calculated by Item Response Theory for the participants representing the members of the group of interest correlates well ($r > 0.7$) with the item difficulty known from the psychometrics of the instrument. We proceed next by compiling the SLCI competence score results from our National Database (Table C-4, Appendix C).

Readers engaging with the supplementary materials of the Appendices should use caution by recognizing that large studies that look at both affective feelings of self-assessed competence and demonstrated competence are rare and break new ground. We have much yet to learn. However, the average score from a test is a very different concept from perceiving these averages as descriptors of people. We are averaging focused measures obtained by specific instruments. We are not averaging human heterogeneity, which would be nonsensical. Any misperceptions that we are averaging people leads to beliefs in "the average person" as an ideal. Rose (2016) details why this misperception is particularly dangerous.

As detailed by supplementary material in Table C-4 (Appendix C), ours is a very complex problem. Each ethnic group consists of heterogeneous individuals, many of whom are succeeding as top performers and others who are failing in their efforts to meet their aspirations, irrespective of whether they self-identified as members of a group that, on average, is collectively privileged or disadvantaged. We have greatly added to the complexity by considering three socioeconomic measures.

Some members of each ethnic group are affected by only one of the socioeconomic conditions, others by all three. Gender differences exist in language ability, but we do not know how or if these vary across ethnic groups. Just varying the percentages of men and women within a group could produce changes in the test scores obtained from these different populaces. A complete analysis, one which we do not attempt in this paper, would seek to quantify the synergistic effects of all three conditions across all of the individual participants in all the different ethnicities. The complex heterogeneity of people challenges this effort.

We turn now to our Paired-Measures Database to learn more about ethnicities through seeing the relationships between the socioeconomic conditions and

metacognitive self-assessment measures (Table C-5, Appendix C). For some ethnic groups (Pacific Islander, Native American, and, perhaps, Middle Eastern), our small numbers of participants are split into six subcategories that consist of so little data that it is unsafe to interpret.

To make interpretations about competency measures and penalties by ethnicity, we will use our larger National Database summarized in Table C-4 Appendix C. To investigate confidence-competence relationships by ethnicities, we will employ our paired measures of demonstrated competence from Table C-5 and self-assessed competence (confidence) from Table C-6 (detailed in Appendix C), which draw their data from the same participants.

Here, we summarize our results. Being a member of the Caucasian majority confers privilege that numeracy detects as statistically higher competency scores and generally higher confidence as determined by self-assessment ratings. Men were not distinguished as a group to be significantly advantaged (privileged); women and men displayed equal competence as manifest in the SLCI scores that measure the competence in understanding science's way of knowing. The Middle Eastern ethnicity offers a striking exception through higher women's scores. Women, in general across all ethnicities, revealed measurably more accurate self-assessment skill than men.

In order of overall importance, three socioeconomic conditions, each considered separately, produce statistically significant penalties on overall SLCI scores: English as a non-native language (7.5-ppt penalty), lack of committed interest to science (5.5-ppt penalty) and status as a first-generation student (4.6-ppt penalty). However, while considered separately to calculate their effect, the three do not act separately within society. These conditions can occur together with their synergistic effects differing from their acting alone and differing in having unequal effects across separate ethnic groups.

Demographic Results – Sexual Orientation

Above, we noted the criteria that we use to evaluate whether a small dataset of fewer than 400 participants can be meaningfully interpreted. The SLCI data for the LGBQ group (N = 49) yields a Cronbach's alpha reliability of R = 0.87, and the item-by-item scores for this group correlate at r = 0.80 with the negative of the item difficulty values (–IRT Difficulty) computed from IRT analysis of the SLCI's 25 items (Appendix B) in the Paired-Measures Database. These results indicate that the small amount of data representing the competencies for this group contains information that merits interpretation (Fig. 7) (Nicholas-Moon 2018).

The mean SLCI scores for the LGBQ students are significantly higher (12.5 ppts) than heterosexual-identifying students. However, students who identified as LGBQ notably underestimated their average score by 8 ppts. Students who self-

identify as heterosexual represent the significant majority, and they have more accurate self-assessments with a mean of only 0.3 ppts above their mean SLCI score. The few students (14 of 859) who identify as having a sexuality of 'other' represent a group too small to allow interpretation. However, this group of 14 produced a lower mean SLCI score and lower mean confidence rating than did the LGBQ group (Fig. 7).

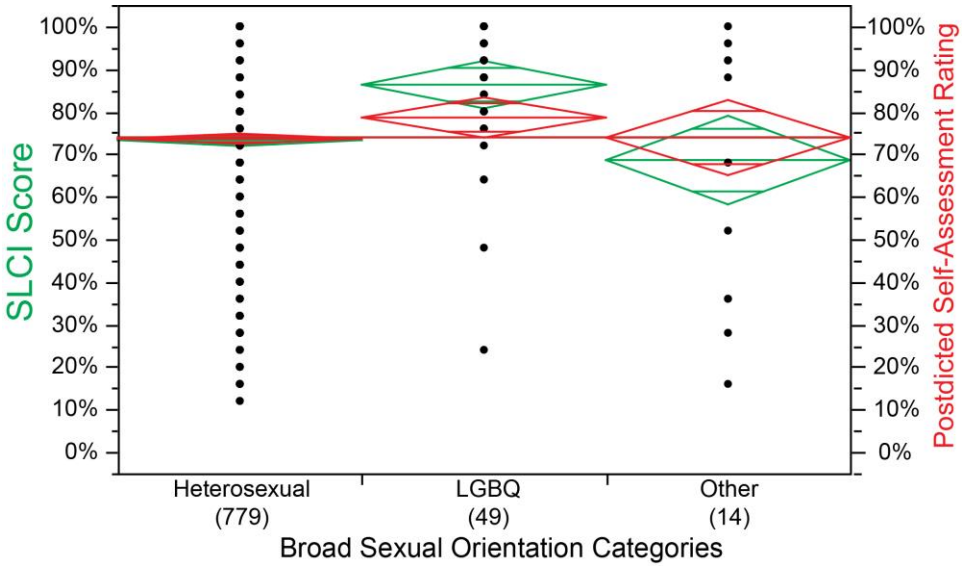

Figure 7. Categorical plot of broad sexual orientations. The figure displays the mean competence scores from the SLCI (green diamonds) and mean self-assessed competence ratings (red diamonds) from the postdicted self-assessment ratings. The diamonds contain the means at the 99.9% confidence level. The horizontal colored lines representing the grand means of the SLCI scores and the self-assessment ratings are congruent at 73.9%.

Of the 49 LGBQ respondents, a large majority identified as Caucasian/White (84%), grew up speaking English as their first language (97.9%), or had one or more parents graduate from college (81.6%). Additionally, most are science majors, indicating interest in science (85.7%). Few of the 49 LGBQ students incur penalties from socioeconomic conditions (9 first-generation, 7 non-science majors, and 1 English-language learner).

Linear Modeling Relationships

We verify that the three socioeconomic conditions exert statistically significant effects on the competency scores when examined across the entire study populace. Our present data reveal these conditions as inequitably distributed across demographic groups of interest. Past research (Nuhfer et al. 2016b) reveals

all of the demographic groups as inequitably distributed across different educational institutions.

Linear modeling offers a way to view the data from the three socioeconomic conditions simultaneously and see their synergistic effects. We use the JMP software's Fit Model Option to construct a simple model from groups rather than from individual participants. We draw on our National Database to tabulate the percentages of students affected by each socioeconomic condition in each ethnic category. To see whether the model might apply to a different, smaller subset of data, such as that which an institution might produce, we tested the model on our smaller Paired-Measures Database.

Our linear model:

$$
\begin{aligned}
Y = 0.00798 \\
&+ (0.13953)*\text{First-generation Student? Yes)} \\
&+ (0.27003*\text{First-generation Student? No)} \\
&+ (0.31752*\text{Interest in Science major? Yes)} \\
&+ (0.24772*\text{Interest in Science major? No)} \\
&- (-0.01873*\text{English as first language? Yes)} \\
&- (0.02666*\text{English as first language? No)}
\end{aligned}
$$

where Y represents SLCI score, provides a good approximation of the mean scores for each ethnic group and even for the scores of binary gender groups which were not included in creating the model (Table C-7, Appendix C). The closeness of the approximation of the model-derived mean scores to the actual mean scores confirms the importance of the influence of the three socioeconomic conditions on students' performance that can arise from variations in social privilege.

Having the kinds of data available for one's institution that we have for our research should enable an institution to estimate the degree to which the members of groups in the institution are incurring competency-measure penalties. For example, few institutions we studied seemed aware that those socioeconomic effects are not equitably distributed across the binary genders in their student populace.

Discussion

Self-assessment and Its Role in Social Justice

This paper connects the numeracy of self-assessment to social justice. We address three competing hypotheses used to explain the relationship between demonstrated and self-assessed competence (Nuhfer et al. 2016a; 2017):

(1) The relationship is just random noise, and the two measures have no meaningful relationship to one another (Porter 2013).

(2) The relationship is meaningful, and studies show that people have a strong propensity toward overconfidence in their actual abilities. Those who are least competent exhibit the greatest overconfidence. (Mabe and West 1982; Falchikov and Boud 1989; Dunning and Kruger 1999; Dunning et al. 2003; Dunning 2011; Ehrlinger et al. 2008; Bell and Volckmann 2011; Ehrlinger and Shain 2014; Webb and Karatjas 2018).
(3) The relationship is meaningful and, overall, people's self-assessed competence is in accord with the competence that they can demonstrate (Favazzo et al. 2014; Nuhfer et al. 2016a and 2017; Keane and Griffin 2018).

The three hypotheses promote very different respective philosophical ideas about self, others, and metacognition:

(1) Humans cannot self-assess. Therefore, students' metacognitive self-assessments of their capabilities have no value.
(2) Some with expertise can self-assess with accuracy because they are aware of their capabilities, but most people have overly inflated views of their knowledge and skills. Thus, when we label a group to be unskilled and unaware, we are likely to be correct.
(3) Those unlike us have different experiences and ways of knowing that we probably do not have from our own experiences and ways of thinking. When we judge a group as largely unskilled and unaware, we are very likely to be wrong.

Hypothesis 1 now seems untenable (Nuhfer et al. 2017). While the volume of peer-reviewed literature in the behavioral sciences gives the highest credibility to Hypothesis 2, that literature, especially the literature that employs the graphical conventions created by Kruger and Dunning (1999) and variants of (y-x) versus (x) type graphs, now appear built on faulty numeracy (Nuhfer et al. 2016a and 2017). Data sets that support Hypothesis 2 when using the numerical approaches employed in Kruger and Dunning (1999) will likely support Hypothesis 3 when researchers apply newer numerical approaches to their data.

Because the second hypothesis has persisted since 1999, it has received two decades of support in the peer-reviewed literature and now has a public recognition as "The Dunning-Kruger Effect." A web search of the term reveals how often people invoke the Effect to label and malign targeted groups. This was observed prominently when we viewed posts generated during the 2016 election. Othering often comes with citations of the Effect (see Dolan 2018 for just one example) as a way to use the credibility of peer-reviewed research to justify maligning a targeted group.

Other peer-reviewed studies have also failed to find evidence that students are unaware of what they do not know and suggest that the typical relationships observed using the Kruger and Dunning convention reflect a constant effect, perhaps an awareness of grading norms (Clayson 2005). Krajc and Ortmann (2008) argue that, based on the highly selective nature of the student body in the initial Kruger and Dunning study, the unskilled-and-unaware problem results from signal extraction differences by those skilled and unskilled. Our study contains participants from a wide spectrum of selectivity, and Nuhfer et al.

(2016a; 2017) traced the problems to numeracy and the collection and analyses of data rather than to any group of participants' inabilities to self-assess.

Our growing dataset continues to support the contention that people across all demographic groups are generally good at self-assessment. This evidence strongly supports Hypothesis 3. Additionally, this evidence begs us to change our cultural value for individuals' awareness of their ability from one that marginalizes those with lower competence as being oblivious of their naivety, to one of respect for individuals' abilities to assess their competence. From this should follow more socially just views and behaviors: value others as a first response, even those with whom we do not identify, listen respectfully to learn why others value their ways of knowing, and resist social pressures to Other.

Numeracy and Social Justice

Nowhere are the differences between the reasoning involved in performing computation and that involved in numeracy made more apparent than when dealing with the calculated differences between social groups. Our data illuminates such statistically significant differences. We often confirm a lowered understanding of science as a way of knowing and reduced confidence in understanding when we compare minority groups to the dominant majority group. This confirmation of competency differences may seem uncomfortably like the computations of test score differences between various racial groups in *The Bell Curve* (Herrnstein and Murray 1994), which created a firestorm of controversy. In that case, some abandoned both numeracy and social justice. Instead, they seized on the computed differences as "evidence" to justify labeling specific racial groups as inferior — as less skilled in thinking and, by association, less aware of their inferiority.

Any weaponizing of "The Dunning-Kruger Effect" can be as inexcusable as misuse of the computed differences reported in *The Bell Curve*. We contend that rather than labeling people as "unskilled and unaware of it," those who would employ such computed differences to justify Othering and maligning others risk exposing themselves as "innumerate and unaware of it."

Numeracy departs from computing by seeking an understanding of the nature of numbers and what they represent in a study. Such understanding requires gaining information from other ways of knowing before we can understand how to interpret differences and perhaps decide which among several competing hypotheses (Chamberlin 1897) is most reasonable. Here, we seek to understand the complexities contributing to the computed differences we find by employing the way of knowing developed in social justice studies. We begin by valuing the reality of extreme diversity *within* all single-group categories.

Responses to the demographic questions on the SLCI were not required. However, over 99% of participants voluntarily assigned themselves to the groups

that we examined under the categories of gender, ethnicity, and sexuality. Each of the groups contains diversity in both measured competence and privilege. Our numbers record measures of scores and ratings, not privilege. Every group contains participants who are wealthy and participants who are homeless. A champion for social justice who says to a person "you are white, and therefore you are a high achiever" reveals as much lack of numeracy as the racist who says "you are black, and therefore you are an at-risk student." After one perceives a group as having great variance within it and a group's privilege as better defined by probability rather than by a label, one is less likely to make such statements. Most certainly, the diversity existing *within* any ethnic group precludes the use of group membership for making judgments about individuals.

The likelihood that a member of a particular group enjoys privilege is highly contextual, and there are varying probabilities of enjoying privilege depending on the groups in which people find themselves. We learned about the distribution of privilege and penalties within and between groups by learning about our participants' socioeconomic conditions.

Socioeconomic Conditions and Majority and Minority Intersections

Regarding academic privilege, we find that being a member of a demographic group determines the probability that one will engage with college either already advantaged or penalized by socioeconomic conditions. These penalties prove measurable through their impacts on the understanding of science as a way of knowing and self-assessed competence of that understanding. Collectively, self-identifying groups have a remarkably accurate group awareness of their disadvantage. This observation furthers our belief in the validity of human self-assessment, and it indicates the depth to which Othering is rooted in the very fabric of our socio-demographic systems.

Educational systems have historically contributed to Othering, often unintentionally. For example, K-12 schools largely dismantled *tracking*, the practice of assigning students to college-prep, general, or vocational curricula groups based on early testing, after tracking was recognized as a system designed for sorting Others into curricula and schools that restricted their social mobility. In the remainder of the Discussion, we hope to interpret these intersectional findings in ways that attend to ethics and social justice.

Socioeconomic Conditions and Binary Gender Intersections

We find no significant differences in science literacy for male and female students except when women (N = 2,979) significantly outscored men (N = 1,716) by 2.8 ppts in the population for whom English was a non-native

language. We attribute this difference to women's greater ability to utilize a learned (second) language, an interpretation informed by research like that of van der Slik et al. (2015) who find that women outperform men in the mastery of a second language. Their observation further carries a neurological explanation (Burman et al. 2008).

Interestingly, this is an example where internalization of Othering may lead to an advantage. That is, women are associated with "soft skills." While soft skills have been Othered, we now understand that acquiring soft skills advantages functioning in society (Crawford et al. 2011).

The overconfidence of male respondents and the more accurate but slightly under-confident self-assessment by female respondents is, perhaps, attributable to stereotypes about traditionally gendered abilities (Tindall and Hamil 2004; Wood and Eagly 2012; Perez-Felkner et al. 2017). Bian et al. (2017) show that at the age of 5, boys and girls associate being "really, really smart" with their own gender. However, girls at ages 6 and 7 are significantly less likely to associate being "really, really smart" with their own gender. Bian et al. (2017) dub this tendency of learners to associate high-level cognitive abilities with men the "brilliance = males" stereotype. For men who are internalizing a brilliance stereotype, language mastery, which requires iterative failure, may be less accessible.

Undoubtedly, "Internalization of Otherness," the propensity "to identify oneself through the eyes of the dominant group in society" (Donovan 2000, 149), is playing a role. Both women and men may internalize the "brilliance = males" stereotype, and this could account for the overconfidence of males versus the slight under-confidence of females seen in our research. However, this study begs us to look even more deeply at how the findings can be applied to the interpretation of our research. Experiencing the three socioeconomic conditions almost always precedes students' experiencing a college education. Thus, the differences we see in this study may have their origins in childhood.

Six- and 7-year-old girls were likely to select girls as having the best grades. Boys selected boys as having the best grades. This shows a disconnection between girls' perceptions of boys' cognitive capability and their perceptions of their actual performance. Girls know that girls are out-scoring boys, but girls still think that boys are smarter. Thus, girls as young as 6 perceive the gap that our research has experimentally shown between boys' self-assessment (confidence) and their competence. Girls are calling attention to the phenomenon of overconfidence in men.

Bian et al. (2017) also show that girls and boys are equally interested in a game for children who try hard, but girls are less interested in the game for smart children. These authors interpreted this finding as indicating that women would be less likely to select careers associated with brilliance (e.g., physics). This may instead be an indicator of growth versus fixed mindset. Children interested in a

game for children who try hard are indicating a value for challenging themselves and building qualities; children interested in a game for smart children are valuing fixed, existing qualities (Dweck 2002). It is interesting to consider that the overconfidence that we see in men in our study might reveal overconfidence as an effect of fixed mindset.

One might reframe current research and instructional attention from students with moderate under-confidence to concentrate instead on students who demonstrate great overconfidence. Educators recognizing the importance of teaching self-assessment rather than fostering overconfidence use different intervention methods for students with varying levels of confidence to support metacognitive development. McDonald (2009) demonstrates that self-assessment training for males in high school produces improvements in academic achievement compared to untrained students.

Socioeconomic Conditions and Ethnicity Intersections

When we examine each ethnicity separately, we recognize that, on average, SLCI scores are lower for all minorities, and their lower group confidence ratings showed that most minority groups are aware of their disadvantage. We also know that socioeconomic conditions that inflict a measurably predictable penalty on test scores and confidence ratings are more prevalent in minority ethnic groups. Yet, we find some unanticipated surprises.

Hispanic Ethnicity. The Hispanic ethnicity is the fastest growing minority group in the U.S. Despite significant differences existing in the socioeconomic conditions affecting the Hispanic populaces in the two datasets, no simple explanation accounts for why Hispanics in our National Database significantly outscore those represented in our Paired-Measures Database (Table C-8, Appendix C). Hispanics in every socioeconomic and binary gender category in the National Database outscore those in the corresponding category of the Paired-Measures Database, despite some socioeconomic conditions being stacked unfavorably against those in the larger National Database.

We do know that mean SLCI scores from students grouped by institutions correlate highly with the selectivity of the institutions (Nuhfer et al. 2016b, 150 Fig.4). Thus, it is tempting to explain the differences by claiming that the National Database probably represents more selective schools than the Paired-Measures Database.

We test this claim by seeing if other ethnicities reflect the same pattern of difference (Table C-9, Appendix C), and we refute the claim. The scores of all ethnicities, save one, indicate that the Paired-Measures Database probably draws from more selective schools than does the National Database. That one exception is the population of Hispanic/Chicano(a). Whatever produces the significant difference between the two populations of Hispanics represented in the different

datasets is not explainable by institutional selectivity or the socioeconomic penalties we assessed. We require focused, detailed studies of specific ethnic categories that we cannot provide here before we offer explanations in which we can have confidence.

The numbers we see between the two datasets in the Results section reveal a general, glaring difference: the National Database contains a richer diversity, the Caucasian-white majority is less dominant in the National Database, and the ratio of Hispanics to Caucasians in the National Database is double that of the Paired-Measures Database. The trend revealed is that, as a group's percentages of a populace grow significantly toward the percentages of the dominant majority, at some point the minority group's competency test scores also begin to approach those of the dominant majority.

Pacific Islander Ethnicity. The self-assessed competency of Pacific Islanders was exceptional among minorities, in that it was as high as the majority Caucasian group. The small number of Pacific Islanders in our study does not offer a sufficient representation to interpret with confidence.

Middle Eastern Ethnicity. While Muslims in the United States are diverse, Middle Eastern Arabs and non-Arabs make up the largest percentage of this group. Since the terrorist attacks of September 11, 2001, Muslims have endured increased stigmatization; students on college campuses who practice Islam have to cope with hostile campuses increasingly fraught with Islamophobia (Ali and Bagheri 2009). Thus, it is not surprising that participants identifying as Middle Eastern might produce unique data in our study. Middle Eastern Ethnicity constitutes: (1) the group that produces the lowest mean score on the SLCI competency measure, (2) the only ethnic group that significantly overestimates its mean competency score, and (3) the only ethnic group in which its women significantly outscore its men. This last finding creates a complex intersection of socioeconomic conditions, ethnicity, and gender.

The socioeconomic query on the SLCI, "Is English your first language?," nets a wide spectrum of disadvantage. The ease of transition between one language and another is variable depending upon the languages. Languages such as French and Spanish are very similar to English making the timeline to transition between English and these languages or vice versa shorter. However, the transition between a language like Arabic and English is a transition that is considered the most difficult (Foreign Service Institute n.d.).

The Middle Eastern students for whom science literacy is lowest are most disadvantaged by disproportionately being first-generation and non-native English speakers. Nuhfer et al. (2016b), working from a smaller dataset, estimate that the markedly lower SLCI scores for this group seem primarily explained by the Middle Eastern group's markedly higher proportion of participants working to

become educated within a non-native language. Here, we suggest that the gap in measured competence scores between men and women is a gap that should be expected to increase in magnitude, in accord with the increased difficulty of the language transition. Given the data that we have and what we know from the literature, these explanations may be as far as we can go with an evidence-based interpretation. Beyond this, we offer possible alternate explanations that we cannot test adequately with our present data.

Gender equity varies according to the sources of foreign national students. In some cases, particularly Middle Eastern, where culture privileges men for selection for higher education, it may be that women must undergo a more rigorous process of selection before they are allowed to come to the U.S. for a college education. In such cases, it is also possible that the women who do manage to come to the U.S. bring a greater representation from the upper economic classes, and thus a better preparation for education than do men. In such cases, women's higher scores on a concept inventory may reflect a marker of privilege.

Sexual Orientation

LGBTQ+ students experience a different climate on college campuses than do their heterosexual peers. Harassment, discrimination, and interpersonal violence are disproportionately high (Kosciw et al. 2004; Renn 2017; Snyder et al. 2018) and this negatively impacts persistence in college and health and wellness (Hurtado et al. 2008). Thus, it is paramount that we expand demographic data to encompass this often voiceless Other and remember that the opposite of liberation is absence.

The performance-versus-confidence gap observed for LGBQ-identifying students likely has some explanations that root in the impacts of "Internalization of Otherness." In this case, LGBQ students may have come to accept themselves as embodying the narrative of the dominant group. The word 'gay' has gained a contemporary meaning beyond sexuality, and that meaning is negative. "That's so gay" now paraphrases "That is so dumb or stupid." In the 2003 National School Climate Survey, 89.5% of students reported hearing this phrase frequently in school (Kosciw et al. 2004). Undoubtedly this phrase becomes internalized and likely causes LGBQ students to have less general confidence, and this transfers into underestimating their abilities in science literacy. Certain homophobic narratives may compound issues of under-confidence as they represent heterosexual men as the rational or learned, "…the further one can distance oneself from what are seen as feminine characteristics, the more of a *real* (read heterosexual) man one is." (Weis 1995, 26).

The significantly higher science literacy of LGBQ students also may root in the "Internalization of Otherness." Tobias (1976) proposes the 'Best Little Boy in

the World' hypothesis stating that those with marginalized sexuality might divert attention from their sexuality through over-achievement in other areas: academics, athletics, prestigious jobs, relationships and so forth. Pachankis and Hatzenbeuhler (2013) substantiates this hypothesis. Moreover, that study shows that objectively measured stigma predicted the extent to which men with non-normative sexualities sought affirmation through competition.

Some argue that Othered groups might be driven to high academic achievement when fueled by the idea of resistance to domination (Pechenkina 2017). The idea of proving the dominant group wrong may be an active driver of achievement. Additionally, we speculate that living with daily stigma may bring a heightened ability to cope with the challenge. This enhanced ability may transcend the academic sphere. It is, however, essential to make the point that this is only possible for Othered groups that have the resources needed to allow resistance through achievement. As none of the LGBQ-identifying students in this study reported English as a learned language, and the majority had parents who went to college and were interested in science, they had advantages that may have provided support and allowed the options for resistance through achievement.

Conclusion

Consistent with our previous work, this study shows that, collectively, groups' ratings of confidence and competence correlated at highly significant levels and reconfirms that people are surprisingly good judges of their abilities. Numeracy is the basis on which we contradict the bulk of self-assessment literature that claims that people cannot self-assess or that people characteristically maintain wildly overinflated views of their capabilities. Views that devalue people for their believing in themselves encourage Othering.

If we instead adopt the view that learners are adequate self-assessors who, with instruction, could become good self-assessors, we are on track to liberate learners from being perceived as unknowing and unaware of their ignorance. That view, if embraced by educators, takes students from the 'oppressed' (Freire 1970) to autonomous individuals who deserve respect. If teachers also work to increase respect for many ways of knowing, then the ways of knowing contributed by minorities can be accepted and valued for the unique perceptions that underlie them and the merit they may add.

Moreover, by showing that this skill of self-assessment is largely maintained over diverse demographics, we hope to compel readers to recognize the capability of all those who have long been Othered. We further this plea by calling attention to several socioeconomic conditions that partially account for the lowered understanding of science as a way of knowing. This turns us away from views that would essentialize differences in performance and provides additional

evidence that access to resources that allow students to become familiar with the language of learning and science as a way of knowing will close gaps in competency scores.

We also acknowledge the limitations of our data. We only consider three socioeconomic conditions. We did not gather data on the socioeconomic status (SES) of the students' high schools. Hattie (2009) notes that the SES of schools that feed colleges and universities are, in fact, more important than the SES of the individual student. One reason that students may not have interest in majoring in science is that the schools in which many students without privilege find themselves do not offer opportunities for students to develop their awareness that science majors and careers in science are within their capability. Where this occurs, schools act as power structures that, without intent, limit the choices of future college students. Gathering information about the advantage or disadvantage that feeder schools provide to students prior to their attending college should allow even more understanding of socioeconomic disadvantages and their complex intersections.

Our work supports some earlier studies on the positive academic performance impacts of certain pre-college home environments. Educated parents speak the language of schooling in their homes, and they are better prepared to expand their children's choices and support their positive ambitions (Hattie 2009). Such home environments advantage students in their later academic achievement. Rosenzweig (2001) also finds that participatory, supportive parenting positively impacts student performance, whereas harsh discipline has a negative impact. These impacts are demographically stratified: those of higher socioeconomic status are most positively impacted; Asian and Latinos/as are more positively impacted than white and African American students. This stratification of socioeconomic status seems to parallel our findings of first-generation, non-native English speakers, and not having an interest in science all being associated with lowered mean test scores. Additionally, our data shows the differential impact of these conditions on demographic groups. We show that a few conditions impact all of the ethnic groups but not by equal degrees, and the intersectionalities within and between groups are highly complex. While we have employed the three socioeconomic conditions and several ways of knowing to gain some explanatory power about privilege, we do not make claims about their relative importance to other conditions or factors that we did not measure.

We look to continue this work by examining institutions of varied selectivity and demographic distributions. We hope to gather more responses from our neglected LGBTQ+ communities. Perhaps melding qualitative approaches with this data collection could allow greater understanding of the gap in competence and self-assessed competence seen in this group. Understanding those who self-identify in the 'other' category on our forms may require a qualitative study. The

extremely low number of respondents identifying as gender queer, intersex, and transgender may reflect stigma for these identifications. Thus, we hope to continue our data collection and to learn better ways to ask these questions.

This project offers a unique opportunity to employ the conceptual ways of knowing from ethics, numeracy, and social justice. In higher education, the disciplines can be the groups that frequently commit Othering. We authors come from different disciplines and very different backgrounds. What we learned from this study helped us to elevate respect in ways that helped us truly hear one another and to work to give voice to each other. We found it useful to start by articulating the role through which each way of knowing could best contribute and inform the whole. Creating this paper for the theme collection of Numeracy was an exceptional learning experience for us.

Acknowledgements

We thank the thousands of participants and faculty who created the data by completing our surveys and the IRB boards from various universities that have overseen this data collection and monitored its use. What we especially want to acknowledge is a rare event of collegial collaboration. The authors of our two prior *Numeracy* papers had continued to amass an ever-growing dataset with demographic data. When we obtained Kira Hamman's invitations to propose a submission for a social justice theme issue of *Numeracy*, we recognized we had valuable data, but none of us had expertise in social justice. However, we knew of three potential contributors in Wyoming: Rachel Watson who had also published on knowledge surveys and who used the SLCI in her classes; Kali Nicholas Moon who just completed her MS thesis by using the SLCI and self-assessment in a social justice context; and Ami Wangeline, who was employing the paired measures to gain understanding about the students and their needs at her community college. Kali's research stipend was funded by her University's Active Learning Initiative and the Center for Teaching and Learning. The authors of our *Numeracy* paper agreed unanimously to invite Paul Walter, who has supervised collection of thousands of SLCI assessments, to bring his experience in Item Response Theory (IRT) to this project and to invite the three Wyoming researchers to take the lead role in the authorship of a social-justice theme paper, with other authors' support role being to continue to develop the competence-confidence parameters that we had established. Fortunately, all four enthusiastically accepted their invitations and Paul's wife Alexa Plunkett Walter contributed her good reviewing and editing abilities to our draft. It is rare that researchers from disparate backgrounds and institutions will risk sharing their databases in this way, and maybe rarer that they succeed in producing a collaborative product. Here, we wish to acknowledge the trust and good spirits

that permitted giving shared voice needed to contribute to the understanding of what constitutes a "wicked problem" (Kolko 2012) that no single discipline or way of knowing can resolve.

References

Ali, Saba Rasheed, and Elham Bagheri. 2009. "Practical Suggestions to Accommodate the Needs of Muslim Students on Campus." *New Directions for Student Services* 125: 47–54. https://doi.org/10.1002/ss.307.

American Psychological Association. 2018. "Ethnic and Racial Minorities & Socioeconomic Status." Retrieved from https://www.apa.org/pi/ses/resources/publications/minorities.aspx.

Anderson, Sharon K., and Mitchell M. Handelsman. 2010. *Ethics for Psychotherapists and Counselors: A Proactive Approach*. Chichester, UK: Wiley-Blackwell. https://doi.org/10.1002/9781444324303.

Anson, Ian G. 2018. "Partisanship, Political Knowledge, and the Dunning-Kruger Effect." *Political Psychology* 39(5): 1173–1192. https://doi.org/10.1111/pops.12490.

Bang, Megan, Douglas L. Medin, and Scott Atran. 2007. "Cultural Mosaics and Mental Models of Nature." *Proceedings of the National Academy of Sciences* 104(35): 13868–13873. https://doi.org/10.1073/pnas.0706627104.

Ballen, Cissy J., Carl Wieman, Shima Salehi, Jeremy B. Searle, and Kelly R. Zamudio. 2017. "Enhancing Diversity in Undergraduate Science: Self-efficacy Drives Performance Gains with Active Learning." *CBE Life Sciences Education* 16:ar56(2): 1–6. https://doi.org/10.1187/cbe.16-12-0344.

Bandura, Albert. 1997. *Self-Efficacy: The Exercise of Control.* New York, NY: W.H. Freeman and Company.

Belenky, Mary F., Blythe M. Clinchy, Nancy R. Goldberger, and Jill M. Tarule. 1986. *Women's Ways of Knowing: The Development of Self, Voice and Mind.* New York: Basic Books.

Bell, Lee Anne. 2016. "Theoretical Foundations for Social Justice Education". In *Teaching for Diversity and Social Justice,* edited by Maurianne Adams and Lee Anne Bell, 3–26. New York: Routledge.

Bell, Priscilla, and David Volckmann. 2011. "Knowledge Surveys in General Chemistry: Confidence, Overconfidence, and Performance." *Journal of Chemical Education* 88(11): 1469–1476. https://doi.org/10.1021/ed100328c.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. 2017. "Gender Stereotypes About Intellectual Ability Emerge Early and Influence Children's Interests." *Science* 355(6323): 389–391. https://doi.org/10.1126/science.aah6524.

Blanch-Hartigan, Danielle. 2011. "Medical Students' Self-Assessment of Performance: Results from Three Meta-Analyses." *Patient Education and Counseling* 84(1): 3–9. https://doi.org/10.1016/j.pec.2010.06.037.

Burman, Douglas D., Tali Bitan, and James R. Booth. 2008. "Sex Differences in Neural Processing of Language among Children." *Neuropsychologia* 46(5): 1349–1362. https://doi.org/10.1016/j.neuropsychologia.2007.12.021.

Calvo, Alejandra, and Ellen Bialystok. 2014. "Independent Effects of Bilingualism and Socioeconomic Status on Language Ability and Executive Functioning." *Cognition* 130(3): 278–288. https://doi.org/10.1016/j.cognition.2013.11.015.

Canales, Mary. K. 2000. "Othering: Toward an Understanding of Difference." *Advances in Nursing Science* 22(4): 16–31. https://doi.org/10.1097/00012272-200006000-00003.

Chamberlin, Thomas C. 1897. "The Method of Multiple Working Hypotheses." *Journal of Geology* 5: 837–848. (Published version of Chamberlin's original 1890 paper). https://doi.org/10.1086/607980.

Chow, Rey. 1989. '"It's You, and Not Me': Domination and 'Othering' in Theorizing the 'Third World."' In *Coming to Terms: Feminism, Theory, Politics*, edited by Elizabeth Weed: 152–161. New York: Routledge.

Clayson, Dennis E. 2005. "Performance Overconfidence: Metacognitive Effects or Misplaced Student Expectations?" *Journal of Marketing Education* 27(2): 122–129. https://doi.org/10.1177/0273475304273525.

Crawford, Pat, Suzanne Lang, Wendy Fink, Robert Dalton, and Laura Fielitz. 2011. *Comparative Analysis of Soft Skills: What Is Important for New Graduates? Perceptions of Employers, Alum, Faculty and Students*. East Lansing, Michigan: Association of Public and Land-grant Universities (APLU) and University Industry Consortium (UIC). http://www.aplu.org/members/commissions/food-environment-and-renewable-resources/CFERR_Library/comparative-analysis-of-soft-skills-what-is-important-for-new-graduates/file.

Damasio, Antonio. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Orlando: Harcourt Inc.

Dolan, Eric. W. 2018. "Study: People with Less Political Knowledge Think They Know a Lot About Politics." *PsyPost*, April 16. https://www.psypost.org/2018/04/study-people-less-political-knowledge-think-know-lot-politics-51062.

Donovan, Josephine. 2000. *Feminist Theory: The Intellectual Traditions*. New York: Continuum.

Dreier, Peter, John Mollenkopf, and Todd Swanstrom. 2001. *Place Matters: Metropolitics for the Twenty-First Century*. Lawrence: University Press of Kansas.

Dunlosky, John, and Janet Metcalfe. 2009. *Metacognition.* Los Angeles, CA: Sage Publications.

Dunning, David. 2011. "The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance." *Advances in Experimental Social Psychology* 44: 247–296. https://doi.org/10.1016/B978-0-12-385522-0.00005-6.

Dunning, David. 2013. "The Problem of Recognizing One's Own Incompetence: Implications for Self-Assessment and Development in the Workplace." In *Judgment and Decision Making at Work*, edited by Scott Highhouse, R. S. Dalal, & E. Salas. 37–56. New York: Taylor & Francis.

Dunning, David, Kerri Johnson, Joyce Ehrlinger, and Justin Kruger. 2003. "Why People Fail to Recognize Their Own Incompetence." *Current Directions in Psychological Science* 12(3): 83–87. https://doi.org/10.1111/1467-8721.01235.

Dweck, Carol S. 2002. "Beliefs That Make Smart People Dumb." In *Why Smart People Can Be So Stupid*, edited by Robert J. Sternberg, 24–41. New Haven: Yale University Press.

Ehrlinger, Joyce, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. 2008. "Why the Unskilled Are Unaware: Further Explorations of (Absent) Self-Insight Among the Incompetent." *Organizational Behavior and Human Decision Processes* 105(1): 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.

Ehrlinger, Joyce, and E. Ashley Shain. 2014. "How Accuracy in Students' Self Perceptions Relates to Success in Learning." In *Applying Science of Learning in Education: Infusing Psychological Science into the Curriculum*, edited by Victor A. Benassi, C. E. Overson and C. M. Hakala, 1–10 http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=B93F6C94B68473D9E1FE541EF035A27A?doi=10.1.1.722.5574&rep=rep1&type=pdf.

Falchikov, Nancy, and David Boud. 1989. "Student Self-Assessment in Higher Education: A Meta-Analysis." *Review of Educational Research* 59(4): 395–430. https://doi.org/10.3102/00346543059004395.

Favazzo, Lacey, John D. Willford, and Rachel M. Watson. 2014. "Correlating Student Knowledge and Confidence Using a Graded Knowledge Survey to Assess Student Learning in a General Microbiology Classroom." *Journal of Microbiology & Biology Education* 15(2): 251–258. https://doi.org/10.1128/jmbe.v15i2.693.

Flannery, Daniele D, "Identity and Self-Esteem." In *Women as Learners: The Significance of Gender in Adult Learning*, ed. by Elizabeth J. Tisdell (San Francisco: Jossey-Bass, 2000) 53–78.

Fisher, Greg. 2015. "Othering." Review of *The Faces of the Other: Religious Rivalry and Ethnic Encounters in the Later Roman World"* (2011) Edited by

Maijastina Kahlos. *The Classical Review* 65(1): 226–228. https://doi.org/10.1017/S0009840X14002339.

Fook, Jan, and Susan Goodwin. 2018. "Introducing Social Justice." In *Everyday Social Justice and Citizenship: Perspectives for the 21st Century*, edited by Anne Marie Mealey, Pam Jarvis, Jan Fook, and Jonathan Doherty, 3–13. New York: Routledge.

Foreign Service Institute. n.d. "*FSI's Experience with Language Learning*." School of Language Studies. Accessed July 16, 2018. https://www.state.gov/m/fsi/sls/c78549.htm.

Frazier, John W., Florence M. Magai, and Eugene Tettey-Fio. 2003. *Race and Place*. Boulder: Westview Press.

Freire, Paolo. 1970. *Pedagogy of the Oppressed*. Translated by Myra Bergman Ramos. New York: Herder and Herder.

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.

Guest, Andrew M., James M. Lies, Jeff Kerssen-Griep, and Thomas J. Frieberg. 2009. "Concepts of Social Justice as a Cultural Consensus: Starting Points for College Students of Different Political Persuasions." *Journal of College and Character*, 10(6): 1–13. https://doi.org/10.2202/1940-1639.1446.

Hattie, John A. C. 2009. "The Contributions from the Home." In *Visible Learning: A Synthesis of Over 800 Meta-analyses Relating to Achievement*. 61–71. New York: Routledge. https://doi.org/10.4324/9780203887332.

Hayes, Elisabeth and Daniele D. Flannery with Ann K. Brooks, Elizabeth J. Tisdell and Jane M. Hugo. 2000. *Women as Learners: The Significance of Gender in Adult Learning*. San Francisco: Jossey-Bass.

Herrnstein, Richard J., and Charles Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Hurtado, Sylvia, Kimberly Griffin, Lucy Arellano, and Marcella Cuellar. 2008. "Assessing the Value of Climate Assessments: Progress and Future Directions." *Journal of Diversity in Higher Education* 1(4): 204–221. https://doi.org/10.1037/a0014009.

James, Stanlie M. 1998. "Shades of Othering: Reflections on Female Circumcision/Genital Mutilation." *Signs: Journal of Women in Culture and Society*, 23(4): 1031–1048. https://doi.org/10.1086/495300.

Jefferson, Thomas, and Benjamin Banneker. 1791. History Gallery C3, *Benjamin Banneker Scientific Thinker*. Smithsonian National Museum of African American History & Culture. Washington, D.C.

Keane, Lainey, and Claire P. Griffin. 2018. "Assessing Self-Assessment: Can Age and Prior Literacy Attainment Predict the Accuracy of Children's Self-

Assessments in Literacy?" *Irish Educational Studies*. 37(1): 127–147. https://doi.org/10.1080/03323315.2018.1449001.

Kolko, John. 2012. "Wicked Problems: Problems Worth Solving." *Stanford Social Innovation Review*. March 6, 2012. https://ssir.org/articles/entry/wicked_problems_problems_worth_solving.

Kosciw, Joseph. G., Emily A. Greytak, Neal A. Palmer, and Madelyn. J. Boesen. 2004. *The 2003 National School Climate Survey: The School-Related Experiences of Our Nation's Lesbian, Gay, Bisexual and Transgender Youth*. New York: A Report from the Gay, Lesbian & Straight Education Network (GLSEN). https://www.glsen.org/sites/default/files/2013%20National%20School%20Climate%20Survey%20Full%20Report_0.pdf.

Krajc, Marian, and Andreas Ortmann. 2008. "Are The Unskilled Really That Unaware? An Alternative Explanation." *Journal of Economic Psychology* 29(5): 724–738. https://doi.org/10.1016/j.joep.2007.12.006.

Kruger, Justin, and David Dunning. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments." *Journal of Personality and Social Psychology* 77(6): 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.

Mabe, Paul A., and Stephen G. West. 1982. "Validity of Self-evaluation of Ability: A Review and Meta-analysis." *Journal of Applied Psychology* 67 (3): 280–296. https://doi.org/10.1037/0021-9010.67.3.280.

McDonald, Betty. 2009. "Exploring Academic Achievement in Males Trained in Self-Assessment Skills." *Education* 37(2): 145–157. https://doi.org/10.1080/03004270802069244.

McMillan, James H., and Jessica Hearn. 2008. "Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement." *Educational Horizons* 87(1): 40–49. http://files.eric.ed.gov/fulltext/EJ815370.pdf.

Merriam-Webster. 2018. "Can 'Other' Be Used as a Verb? 'Other': A Verb That Sets Itself Apart." *Merriam Webster Words We're Watching*. https://www.merriam-webster.com/words-at-play/other-as-a-verb.

Mitchell, Donald, Jr. (Ed.). 2014. *Intersectionality & Higher Education: Theory, Research, & Praxis*. New York: Peter Lang Publishing Inc.

National Center for Education Statistics. 2016. Digest of Education Statistics. Table 326.10. *Graduation Rate from First Institution Attended for First Time, Full-time Bachelors Degree-seeking Students at 4-year Postsecondary Institutions by Race/Ethnicity, Time to Completion, Control of Institution and Acceptance Rate: Selected Cohort Entry Years, 1996-2009*. U.S. Department of Education. https://nces.ed.gov/programs/digest/d16/tables/dt16_326.10.asp.

Nicholas-Moon, Kali. 2018. "Examining Science Literacy Levels and Self-Assessment Ability of University of Wyoming Students in Surveyed Science Courses Using the Science Literacy Concept Inventory with Expanded Inclusive Demographics." Master's thesis, University of Wyoming.

Nicol, David. 2009. "Assessment for Learner Self-Regulation: Enhancing Achievement in the First Year Using Learning Technologies." *Assessment & Evaluation in Higher Education* 34(3): 335–352. https://doi.org/10.1080/02602930802255139.

Nicol, David, and Deborah Macfarlane-Dick. 2006. "Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31(2): 199–218. https://doi.org/10.1080/03075070600572090.

Nieto, Sonia. 1998. "Fact and Fiction: Stories of Puerto Ricans in US Schools." *Harvard Educational Review* 68(2): 133–163. https://doi.org/10.17763/haer.68.2.d5466822h645t087.

Nuhfer, Edward B. 2015. "Clarification to Points in 'Correlating Student Knowledge and Confidence Using a Graded Knowledge Survey to Assess Student Learning in a General Microbiology Classroom.'" *Journal of Microbiology & Biology Education* 16(2): 125–126. https://doi.org/10.1128/jmbe.v16i2.986.

Nuhfer, Edward, Christopher Cogan, Steven Fleisher, Eric Gaze, and Karl Wirth. 2016a. "Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency." *Numeracy* 9(1): Article 4: 1–24. https://doi.org/10.5038/1936-4660.9.1.4.

Nuhfer, Edward B., Christopher B. Cogan, Anya Goodman, Carl Kloock, Christopher W. Wheeler, Gregory G. Wood and Natalie Zayas Delgado. 2016a. "Using a Concept Inventory to Assess the Reasoning Component of Citizen-Level Science Literacy: Results from a 17,000-Student Study." *Journal of Microbiology & Biology Education* 17(1): 143 – 155. https://doi.org/10.1128/jmbe.v17i1.1036.

Nuhfer, Edward, Steven Fleisher, Christopher Cogan, Karl Wirth, and Eric Gaze. 2017. "How Random Noise and a Graphical Convention Subverted Behavioral Scientists' Explanations of Self-Assessment Data: Numeracy Underlies Better Alternatives." *Numeracy*10 (1): Article 4: 1–31. https://doi.org/10.5038/1936-4660.10.1.4.

Oughton, Helen. M. 2018. "Disrupting Dominant Discourses: A (Re)Introduction to Social Practice Theories of Adult Numeracy." *Numeracy* 11(1) Article 2: 1–18. https://doi.org/10.5038/1936-4660.11.1.2.

Pachankis, John E., and Hatzenbeuhler, Mark L. 2013. "The Social Development of Contingent Self-Worth in Sexual Minority Young Men: An Empirical

Investigation of the 'Best Little Boy in the World' Hypothesis." *Basic and Applied Social Psychology* 35(2): 176-190. https://doi.org/10.1080/01973533.2013.764304.

Pechenkina, Ekaterina. 2017. "'It Becomes Almost an Act of Defiance': Indigenous Australian Transformational Resistance as a Driver of Academic Achievement." *Race Ethnicity and Education* 20(4): 463–477. https://doi.org/10.1080/13613324.2015.1121220.

Perez-Felkner, Lara, Samantha Nix and Kirby Thomas. 2017. "Gendered Pathways: How Mathematics Ability Beliefs Shape Secondary and Postsecondary Course and Degree Field Choices." *Frontiers in Psychology* 8(386): 1–11. https://doi.org/10.3389/fpsyg.2017.00386.

Pintrich, Paul R. 2004. "A Conceptual Framework for Assessing Motivation and Self-Regulated Learning in College Students." *Educational Psychology Review* 16(4): 385–407. https://doi.org/10.1007/s10648-004-0006-x.

Porter, Stephen R. 2013. "Self-Reported Learning Gains: A Theory and Test of College Student Survey Response." *Research in Higher Education* 54(1): 201–226. https://doi.org/10.1007/s11162-012-9277-0.

Renn, Kristen. 2017. "LGBTQ Students on Campus: Issues and Opportunities for Higher Education Leaders." *Higher Education Today*, April 10, 2017: Accessed July 4, 2018. https://www.higheredtoday.org/2017/04/10/lgbtq-students-higher-education/.

Romm, Norma R.A. 2017. "Conducting Focus Groups in Terms of an Appreciation of Indigenous Ways of Knowing." In *Handbook of Research Methods in Health Social Sciences*, edited by Pranee Liamputtong, 1–15. Singapore: Springer. https://doi.org/10.1007/978-981-10-2779-6_46-1.

Rose, Todd. 2016. *The End of Average: Unlocking Our Potential by Embracing What Makes Us Different*. New York: Harper-Collins.

Rosenzweig, Charlotte. 2001. "A Meta-Analysis of Parenting and School Success: The Role of Parents in Promoting Students' Academic Performance." In *American Educational Research Association 2001 Conference* 1–44. Seattle: Educational Resources Information Center. https://files.eric.ed.gov/fulltext/ED452232.pdf.

Ross, John A. 2006. "The Reliability, Validity, and Utility of Self-Assessment." *Practical Assessment, Research & Evaluation* 11(10): 1–13. http://pareonline.net/getvn.asp?v=11&n=10.

Saw, Guan, Chi-Ning Chang and Hsun-Yu Chan. 2018. "Cross-sectional and Longitudinal Disparities in STEM Career Aspirations at the Intersection of Gender, Race/Ethnicity, and Socioeconomic Status." *Educational Researcher* 47(8): 525–531. https://doi.org/10.3102/0013189X18787818.

Scheller-Boltz, Dennis. n.d. "LGBT? LGBTQ+? LGBTTQQFAGPBDSM? Or just: QUEER! Critical Remarks on an Acronym in Slavonic and Non-Slavonic Languages." *Academia.edu*. San Francisco: Academia, Inc. https://www.academia.edu/35992619/LGBT_LGBTQ_LGBTTQQFAGPBDSM_Or_just_QUEER_Critical_Remarks_on_an_Acronym_in_Slavonic_and_Non-Slavonic_Languages. Accessed November 5, 2018

Shapiro, Doug, Afet Dunbar, Faye Huie, Phoebe K. Wakhungu, Xin Yuan, Angel Nathan, and Youngsik Hwang. 2017. *A National View of Student Attainment Rates by Race and Ethnicity – Fall 2010 Cohort (Signature Report No. 12b)*. Herndon, VA: National Student Clearinghouse Research Center. https://nscresearchcenter.org/wp-content/uploads/SignatureReport12.pdf.

Snyder, Jamie A., Heidi L. Scherer, and Bonnie S. Fisher. 2018. "Interpersonal Violence among College Students: Does Sexual Orientation Impact Risk of Victimization?" *Journal of School Violence* 17(1): 1–15. https://doi.org/10.1080/15388220.2016.1190934.

Tindall, Tiffany, and Burnette Hamil. 2004. "Gender Disparity in Science Education: The Causes, Consequences and Solutions." *Education* 125(2): 282–295.

Tobias, Andrew. 1976. *The Best Little Boy in the World*. New York: Ballantine.

U.S. Census Bureau. 2017. *U.S. Poverty Report*. Retrieved November 4, 2018 from: https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pov/pov-01.html.

van der Slik, Frans W. P., Roeland W. N. M. van Hout, and Job J. Schepens. 2015. "The Gender Gap in Second Language Acquisition: Gender Differences in the Acquisition of Dutch among Immigrants from 88 Countries with 49 Mother Tongues." *Public Library of Science PLoS ONE* 10(11): e0142056. https://doi.org/10.1371/journal.pone.0142056.t004.

Warner, Michael. 1999. *The Trouble with Normal: Sex, Politics and the Ethics of Queer Life.* Cambridge, MA: Harvard University Press.

Webb, Jeffrey. A., and Andrew G. Karatjas. 2018. "Grade Perceptions of Students in Chemistry Coursework at All Levels." *Chemistry Education Research and Practice* 19(2): 491–499. https://doi.org/10.1039/C7RP00168A.

Weis, Lois. 1995. "Identity Formation and the Processes of 'Othering': Unraveling Sexual Threads." *The Journal of Educational Foundations* 9(1): 17-33. https://eric.ed.gov/?id=EJ510975.

Wood, Wendy, and Alice H. Eagly. 2012. "Biosocial Construction of Sex Differences and Similarities in Behavior." *Advances in Experimental Social Psychology* 46: 55–123. https://doi.org/10.1016/B978-0-12-394281-4.00002-7.