Cite as: Feldon, D. F. (in press). Implications of measurement issues for advancing the socialization framework. In J. C. Weidman & L. DeAngelo (Eds.), *Socialization in higher education and the early career: Theory, research, and application.* New York: Springer International Publishing.

Implications of Measurement Issues for Advancing the Socialization Framework

David F. Feldon

Utah State University

Abstract

This chapter examines issues of measurement in socialization research and their implications for socialization as a theoretical framework. Although the majority of socialization research relies on qualitative methods, quantitative studies to assess the generalizability of core tenets of socialization require psychometrically sound and valid measures. Better understanding ways in which relevant constructs interact within the socialization process, including the possibility that some have greater influence than others in driving outcomes, has the potential to inform interventions and the strategic investment of resources to optimize their effectiveness for enhancing student success. Calling upon relevant data from published studies,

the chapter examines potential strengths and weaknesses of existing measures on an empirical basis and discusses the reciprocal implications that measurement and theory development can hold for one another.

In broad terms, socialization theory is the dominant theoretical framework for understanding graduate education in the United States (Gardner, 2010). Although differing versions exist (e.g., Austin, 2002; Bragg, 1976; Gardner & Mendoza, 2010; Tierney & Rhoads, 1994; Van Maanen & Schein, 1979; Weidman, Twale & Stein, 2001), there is a broad agreement that graduate students progress through stages to actualize their identities as legitimate members of an academic discipline with "individual and social roles, personalities and social structures becom[ing] fused" (Thornton & Nardi, 1975, p. 880). Further, the process of socializing into one's discipline as a professional entails core elements developed through engagement with the social and programmatic structures at each stage: knowledge acquisition, investment, and involvement (Weidman, Twale, & Stein, 2001). Thus, graduate students develop through these mechanisms to a point where, if socialization is successful, they are both "motivated and able to perform [a professional role] in a[n]… acceptable fashion" (Merton, Reader, & Kendall, 1957, p. 41), having learned "the relevant skills, knowledge, habits,

attitudes and values of the group" (Austin & McDaniels, 2006, p. 400). In current literature, indicators of such motivation often include persistence in degree programs or intent to pursue a faculty or research position (Austin, 2002; Lovitts, 2001). Indicators of performance are typically scholarly productivity (e.g., publications) and awards (Paglis, Green, & Bauer, 2006), but they may also be conceptualized as the reproduction of normative day-to-day scholarly practices in professional contexts (Reybold, 2003). In short, these are the outcomes of the socialization process. The purpose of this chapter is to examine existing strategies for the quantitative measurement of socialization's outcomes and processes, existing challenges for appropriate measurement, and their implications for the further development of socialization as an explanatory theory of graduate education.

Although readers of this book are likely to be familiar with socialization's major constructs, it merits reiterating that socialization has the foundational properties of a theory. It specifies a set of functions and constructs (i.e., mechanisms; Rojas, 2017) that interact to shape the evolving identity and subsequent identity-linked choices

of graduate students. This inherent linking of process and outcome suggests that knowing something about the socialization experiences of a student or group of students should enable us to predict something about whether and how students choose to participate in the broader community of scholars during and subsequent to their graduate training.

Despite this widely held supposition, there have been very few longitudinal studies linking socialization experiences to outcomes. Those that have been published typically use quantitative methods to assess the predictability of outcomes based on data regarding students' graduate school experiences (but see Holley [2018] and Wulff, Austin, Nyquist, & Sprague [2004]). For example, Paglis et al. (2006) administered surveys to 130 participants three times over 5.5 years, with a focus on mentorship experiences as predictors of subsequent scholarly productivity, self-efficacy, and commitment to a research career. While the strength of mentoring positively predicted productivity and self-efficacy, it did not predict research career commitment. This is noteworthy, because the development of values and identity consistent with the culture of the aca-

demic discipline is a central emphasis of the mechanisms of socialization, among which mentoring figures prominently (Austin & McDaniels, 2006; Weidman et al., 2001).

Similarly, an ongoing longitudinal study of 336 Ph.D. students in the biological sciences in the United States (Feldon, Jeong, et al., 2017) has identified unexpected trends in the relationships between socialization factors and outcomes. For instance, when comparing the experiences and outcomes of domestic Asian students, domestic White students, and international Asian students, Roksa, Jeong, Feldon, and Maher (in press) found that, after the first two years of doctoral study, rates of scholarly productivity did not parallel reported levels of socialization experiences. Specifically, domestic Asian students reported access to and participation in scholarly activities and interactions with faculty and peers at the same levels as their domestic White peers, with international Asian students reporting significantly less. However, the scholarly productivity of international Asian students was on par with that of domestic White students, with both groups publishing significantly more than domestic Asian students. Thus, contrary to the predictions of sociali-

zation theory and the findings of Paglis and colleagues (2006), favorable socialization did not predict stronger outcomes. Likewise, Roksa, Feldon, and Maher (in press) found that, across all participants, socialization experiences were not predictive of students' commitment to completing their doctoral degree programs. In another study examining the socialization experiences of graduate students in STEM disciplines, Feldon, Maher, Roksa, and Peugh (2016) report that the widening research skills gap between two groups of students over time could not be explained by access to mentorship, opportunities to collaborate on publications, or other socialization factors typically associated with faculty or academic departments.

Collectively, these findings raise questions about the ability of the mechanisms specified by socialization theory to predict outcomes. However, as quantitative studies, their data depend on the ways in which core constructs are operationalized and measured. This consideration is critical in light of the strong traditions of qualitative inquiry and thick description that have fueled much of the scholarship on socialization (Austin, 2002; Gardner, 2008a). If the instruments used to measure the socializing features of a doctoral

program are insufficiently nuanced, they would be unable to detect substantive differences that might account for differences in outcomes. Similarly, if the strategies used to validate the instruments or analyze the data they provide do not represent best practices, the resulting findings and inferences may be spurious. To examine these issues, the following sections of this chapter discuss the properties of existing measures used within the socialization framework and discuss their implications relevant to future research.

Underlying Methodological Assumptions

Weidman et al. (2001) characterize socialization theory in higher education as "a structural-functional approach to describe the relationships among the stages of socialization, core socialization elements, and fundamental outcomes of professional socialization" (p. 21). As such, its ontological and epistemological assumptions are aligned with a postpositivist or realist framework (Burrell & Morgan, 2016; Schwandt, 1997), which asserts an underlying reality external to the perceptions of participants, while recognizing the fun-

damentally subjective or constructed nature of individuals' experiences of that material reality. Most current studies of socialization in graduate education typically engage qualitative methods to more deeply explore the constructed perspectives of participants as a window into the lived experience of socialization, using inductive and descriptive strategies of constant comparison to identify emergent themes through the analysis of interviews and observations (e.g., Austin, 2002; Gardner, 2010; Holley, 2009). Thus, it is typically the case that such research does not position itself to test hypotheses. However, this trend does not mean that the corpus of qualitative inquiry around socialization exists at odds with postpositivist or realist approaches to testing theories (e.g., Glaser & Strauss, 1967; Maxwell, 2004; Shadish, Cook, & Campbell, 2002).

The role of a theory within a postpositivist or realist worldview is to articulate a set of mechanisms that give rise to predictable outcomes (Shadish et al., 2002). In order to test the tenets of a theory, we can assess the extent to which postulated mechanisms work as anticipated by providing opportunities for them to fail. Should outcomes repeatedly arise that diverge from those pre-

dicted by a theory under appropriate circumstances, a likely explanation must be crafted that identifies the role of the research method, the context within which the study was executed, a revision to the causal assertions of the theory, or some combination of these in producing the unanticipated results. Such theory testing is ideally suited for quantitative approaches that lend themselves to deductive modes of reasoning (Kelly, 2017). Appropriate research strategies in the study of graduate socialization may include hypothesis testing through quasi-experimental studies or predicting the relative strengths of relationships between independent variables (i.e., sociological structures or functions) and dependent variables (i.e., participant outcomes, such as those identified in the previous section). As Rojas (2017, p. xxiii) explains, "the translation of theoretical ideas into research agendas requires a link between the concepts that motivate theory…and the specific things that can be measured."

Contrary to the stereotype of quantitative research paradigms, these modes of inquiry do not necessitate the disregard of local meaning and context. Indeed, Miles and Huberman's (1984) notion of "local causality" makes clear that the meaning given to social

phenomena by participants is essential to understanding the ways in which mechanisms lead to outcomes. However, relevant local meanings and sociocultural features can be represented effectively through either qualitative or quantitative symbols, as long as the nuances of meaning and experience are adequately considered in constructing categories (Feldon & Tofel-Grehl, 2018; Maxwell & Mittapalli, 2010). When the construction of quantitative measures does not reflect these nuances sufficiently, the products cannot serve as valid instruments. In such cases, inferences derived from them through statistical analyses are flawed, introducing systematic measurement error from the perspective of traditional quantitative research and a misrepresentation of causal mechanism from the perspective of traditional qualitative research (Maxwell, 2004). Thus, even if an instrument is statistically reliable in terms of the internal consistency in its pattern of responses, the resulting data may not accurately or adequately reflect the underlying structures or functions (Pedhazur & Schmelkin, 1991). However, if an instrument is valid, it will inherently be reliable, because it will reflect the phenomena of interest in a manner compatible with the underlying mechanisms.

The concepts of validity and reliability are traditionally associated with quantitative research. However, analogous concepts extend to qualitative inquiry as well (Patton, 2002). Lincoln and Guba (1985, p. 300) discuss "dependability" as an important aspect of qualitative research, in which it is possible to document the analytic process for the purpose of allowing others to trace the path of analysis from raw data through to the researchers' conclusions to verify that they are well-grounded in the data. Similarly, two forms of validity discussed in reference to qualitative research are essential to understanding data collected for any study of socialization: Interpretive validity engages the question of whether or not the inferences drawn from the collected data adequately reflect the perspectives of the participants, and theoretical validity represents the extent to which a theoretical construct is applied appropriately in the interpretation of qualitative data (Maxwell, 1992).

Measurement Characteristics

In research focused specifically on socialization within graduate education, there are several well-known survey instruments developed to elicit information from enrolled students[1]. I briefly describe three of these in the following section. Thereafter, aspects of their design and validation are discussed in relation to socialization theory and common measurement practices.

Examples

Graduate and Professional Education Socialization Scales (Weidman & Stein, 2003)

Weidman and Stein (2003) developed a survey instrument to elicit socialization information from Ph.D. students in six areas: participation in scholarly activities, student-faculty interactions, student-peer interactions, supportive faculty environment, department

[1] Lovitts's (2001) seminal work on doctoral attrition used survey measures in addition to interviews, but the retrospective nature of her instrument differentiates it substantially from those discussed in this chapter.

collegiality, and student scholarly encouragement. Participation in scholarly activities was assessed through a list of 11 items to which participants provided a binary response (yes/no) indicating if they had engaged in each of the listed activities (e.g., peer critique of writing, grant writing, manuscript writing, journal article submission, etc.). Likewise, student-faculty and student-peer interactions were assessed using 4 binary items each to gauge these types of engagement (i.e., sometimes engage in social conversation, often discuss topics in the field, often discuss other topics of intellectual interest, ever talk about personal matters). Items assessing supportive faculty environment and department collegiality were based on a 5-point Likert scale, ranging from 1 (lowest level of agreement with the presented statement) to 5 (highest level of agreement) for items such as “I feel free to call on the faculty for academic help” and “the faculty sees me as a serious scholar.” The student scholarly encouragement subscale prompted participants to respond using a 3-point Likert (not at all true, somewhat true, completely true) to items stating that their department “promotes scholarly interchange between

students and faculty," "fosters and develops self-confidence in students," and "encourages the scholarly aspirations of all students."

Survey of Doctoral Student Finances, Experiences, and Achievements (SDSFEA; Nettles & Millett, 2006)

The SDSFEA (Nettles & Millett, 2006) consists of 88 items across seven sections: application and enrollment process, current doctoral program experience, attendance patterns, financing your doctoral education, future plans, undergraduate experiences, and background. Of these, the section focusing on current doctoral program experiences specifically elicits data on socialization while in the doctoral program. Other related aspects, such as whether students were provided with teaching or research assistantships were assessed under the financing section. This instrument also collected information about outcomes, such as scholarly productivity and degree completion.

Response formats across items varied. However, those focusing on socialization experiences utilized 5-point Likert responses ranging from "strongly disagree (1)" to "strongly agree (5)" in response to nine items like "At least one faculty member in my pro-

gram has had a strong impact on my intellectual development" and "I am satisfied with the level and types of student organizations and committees in my program." Participants' perceptions of their experiences were captured using a different 5-point Likert scale (i.e., "very dissatisfied (1)" to "very satisfied (5)") in response to twelve prompts targeting quality of instruction, collegial atmosphere, quality of academic advising, and faculty interest in participants' research. The survey also elicited information about the frequency of participants' discrete scholarly activities, asking that participants indicate the number of times they did things like participate in an independent study, publish a research article in a refereed journal, and apply for an external research grant with a faculty member. Response options for each consisted of 0, 1, 2, 3, 4, or "5 or more" for 22 activities. Ongoing activities like participating in informal study groups or receiving feedback about academic progress (10 items) provided a Likert response format ranging from "never (1)" to "very often (5)." Factor analysis identified aggregated scores for peer interaction, student/faculty social interaction, academic interactions with faculty, and interactions with advisors. Additionally, partici-

pants were asked to select the professional position they anticipated they would hold immediately after completing their doctoral degrees. Fifteen options were provided, including "faculty at a college or university," "researcher in the private sector," "homemaker," and "other (specify)".

Survey of Mentoring and Doctoral Student Outcomes (Paglis et al., 2006)

The survey developed by Paglis et al. (2006) consists of four subscales that each use Likert responses. The first subscale, Career Commitment, presents five items (e.g., "I am committed to a research career") to which participants respond on a 5-point scale from "strongly disagree (1)" to "strongly agree (5)." The second subscale, Self-Efficacy, includes ten items to which participants rate each stated academic skill (e.g., "be an effective co-author on a paper," "design and conduct effective research") on an 11-point scale, ranging from "not at all confident (0)" to "very confident (10)." The third subscale examines psychosocial mentoring experiences, presenting 14 items (e.g., "My adviser shares history of his/her career with me," "I try to imitate the work of my advisor.") to which participants re-

spond on a 5-point scale ("to a very slight extent (1)" to "to a very great extent (5)"). The final subscale, Career-Related Mentoring, consists of six items (e.g., "My adviser helps me to meet new colleagues," "My adviser gives me assignments or tasks that prepare me for a research position after I graduate.") to which participants respond using the same response options as the psychosocial mentoring subscale.

Implications of Response Format

Broadly speaking, response format is an important consideration for issues of validity, because the structure of possible responses that an instrument affords participants makes certain assumptions about the nature of their experiences and constrains their ability to communicate it. For example, face validity is usually assessed by recruiting experts in the area to review survey items and verify that they represent the target constructs. However, one of the potential shortcomings of this approach is that the experts typically recruited for such activities hold firm theoretical stances on the subject matter. As such,

the items are considered in light of a priori conceptions of the relevant constructs and may be inappropriately worded or conceptualized for the student perspectives they are designed to capture. Therefore the following sections discuss item format with respect to ecological validity, defined by Bronfenbrenner (1977) as "the extent to which the environment experienced (i.e., survey response options in this case) by the subjects in a scientific investigation has the properties it is supposed or assumed to have by the experimenter" (p. 516).

Likert items

It is noteworthy that all three surveys discussed make extensive use of Likert scale items and frequency counts as a strategy for capturing participants' graduate education experiences. This is an intuitive strategy, as it is both common in the social sciences and does not impose as much burden on respondents as would an interview or focus group. However, Likert items must be carefully assessed in terms of their ability to effectively represent the full range of meaningful responses that participants might wish to provide for a given prompt (Cummins & Gullone, 2000). If the response range is too restricted, it will force the responses of people with experiences

that differ in important ways into the same response value. By default, this can enhance the internal consistency of the scale (i.e., Cronbach's alpha) by homogenizing responses, but do so at the expense of the instrument's underlying validity. Further, compressing the response scale in such a way limits the ability to detect meaningful variation, and in turn, the ability of statistical analyses to detect trends that may exist (Cohen, 1983). Conversely, if the response range is too broad, it risks diffusing fundamentally similar meanings from respondents into differing response values that are arbitrarily selected, obscuring otherwise informative trends.

A related issue is one of the semantic distances between response labels (i.e., anchors). When responses to Likert items are analyzed using parametric statistics, the underlying assumption is that the conceptual distance between each response option and the next is the same, such that the data will function like an interval scale (Fraenkel & Wallen, 1996). While there is extensive empirical evidence in the social sciences that Likert items frequently behave like interval data, such trends are arguably the result of the extensive efforts typically invested in the development of valid survey response

items (Carifio & Perla, 2008). If the assumption of equal conceptual distance does not hold, however, then interpretations of the results will be skewed (Jamieson, 2004). For example, Weidman and Stein (2003) use 3-point Likert response scales to elicit participants' experiences of scholarly encouragement within their home academic departments, with response options consisting of "not at all true," "somewhat true," and "completely true." The first and third response options provide clear meanings. However, the middle option arguably presents a much larger range of possible impressions than the other two options. "Somewhat true" may likely confound multiple differentiated meanings that are more representative of participant experiences and reflect more even semantic spacing, such as "mostly true," "moderately true," and "slightly true." Several empirical studies have attempted to determine the optimal number of Likert response options. While there is some variation in findings, in general, Likert scales offering between 5 and 10 response items appear to produce equivalent distributions (Dawes, 2008). However, for more nuanced judgments, there may be advantages to using the

10-point response scale (e.g., quality of life; Cummins & Gullone, 2000).

Another challenge in constructing Likert items is to ensure that the stem encompasses only a finite range of a phenomenon. For example, Nettles and Millett (2006) ask respondents to score the statement "at least one faculty member in my program has had a strong impact on my intellectual development" on a scale from "strongly disagree" to "strongly agree." However, this item necessarily equates the same level of agreement with the statement if a student's intellectual development was (a) strongly impacted by only one faculty member during the entire scope of her degree program or (b) strongly impacted by many on a regular basis. These two scenarios would likely have substantially different impacts on a student's socialization and subsequent outcomes, yet they yield the same score in the context of the survey. It therefore seems worthwhile to investigate the frequency of these types of events, along with the typical number of different people involved in them prior to constructing fixed response items if the specifics are of theoretical interest. If they are not, then constructing items to elicit the extent of impact of

faculty interactions generally may effectively avoid unnecessary threats to item validity.

A third factor to consider is whether or not the respondent is in a good position to knowledgeably answer the question asked. For example, Paglis and colleagues (2006) assess mentoring in part by directing students to rate the extent to which their advisers "give assignments or tasks that prepare [them] for a research position after [they] graduate." However, any response is inherently speculative, as the respondents will not have had firsthand experience to know the extent to which a given task may or may not prepare them for a position as an independent scholar. Indeed, previous research clearly indicates that doctoral students often lack familiarity with what preparation is necessary to take on that role—both in general and in terms of specific post-degree employment options (Austin, 2002; Holley, 2018; Lovitts, 2008; Pole, 2000). Even when respondents are asked questions that they do not feel knowledgeable enough to answer or do not hold an opinion, research in survey methods indicates that they will respond to the item. Further, providing a ‘‘don't know/no opinion’’ option fails to mitigate this tendency even when

respondents are highly educated, as is the case with any study of graduate education (Bishop, Tuchfarber, & Oldendick, 1986; Krosnick et al., 2002; Schwarz, 1999).

Binary and count items

Both Weidman and Stein (2003) and Nettles and Millett (2006) ask respondents to indicate if they have participated in various socializing activities, such as writing a grant. In the first case, participants indicated either "yes" or "no," which captures neither the frequency of the activity, the nature of involvement, nor the perceived quality of the experience. It essentially equates engaging in a specific activity once with engaging many times. In the latter case, respondents report the frequency of events on a truncated scale (i.e. "5 or more"), in which participating in the activity 5 times and 20 times are scored identically. Depending on the joint socialization contexts of discipline and research intensity of the institution, it may be that "5 or more" is an atypically high occurrence for an activity, which would have a low frequency of response and serve as an appropriate category. However, it is also possible that 5 occurrences is relatively

low, and that the majority of respondents from a given context might have upwards of 10 occurrences.

For example, Feldon, Peugh, Maher, Roksa, and Tofel-Grehl (2017) found that while only 20% of first-year doctoral students in cellular and molecular biology programs reported authorship on a published journal article, the number of authored publications ranged from 1 to 3. Given that the average duration of a Ph.D. program in this field is 5.5 years, it is readily apparent that a different scale would be necessary to capture the full variance in scholarly productivity over the course of a degree program. Indeed, after three years in their programs, the participants in that study who have published in journals report a mean of 2.9 articles (SD=1.5), with 15% having published 5 or more articles (maximum reported is 10 articles) about the time they are halfway to degree completion (unpublished data). In another field, with different publishing norms, the publication rates would likely be very different. Thus, it may be advisable to permit respondents to directly report numeric values, rather than provide preset options that may be inappropriate to their circumstances.

It is also the case that the experiences entailed in activity participation will vary, not just by individual, but by group. Feldon et al. (2017) found that the number of research hours invested to yield a publication was significantly greater for women than for men, with the likelihood of receiving authorship credit increasing by 15% for men over women for every 100 hours of laboratory time invested. Additionally, women were significantly more likely than men to serve as first author, which typically requires substantially more investment of time and effort than a lower authorship position. In this sense, even a count of published articles without considering authorship order does not represent the same underlying experiences for men and women. When viewed as an outcome, such group-based differences would be readily observed. However, if the experience of writing and bringing a manuscript to publication were considered as an aspect of the socialization process (i.e., independent variable), it could readily lead to invalid conclusions based on inappropriately aggregated response items.

Implications of Data Type

The surveys described above constitute self-report instruments, in which respondents are asked to describe their experiences and judgments from their perspectives through the range of response options provided. While the constraints of closed-ended items formats are evident in terms of restricting nuance, personal meaning, and experience, there are additional characteristics that are also vital to consider in understanding self-report data. Despite an understandable tendency to treat subscale scores at face value, there are well-established sources of bias that must be considered.

Acquiescence and social desirability effects

When asking respondents to assess their traits or experiences, there is a well-established tendency to skew responses toward answers that protect the respondent from possible negative judgments by those who read their responses. One way in which this manifests is acquiescence, in which respondents tend to agree with the statements offered in item stems far more often than they tend to disagree (Couch & Keniston, 1960). Consequently, there is a tendency for

items with agree-disagree response scales to correlate positively with one another, which can inflate reliability estimates at the expense of valid measurement (Messick, 1967). This trend toward acquiescence can be amplified when respondents do not find items to be clear or meaningful (Schuman & Presser, 1981). One way to guard against this phenomenon is to employ a mix of positively and negatively framed items, rather than presenting all items positively (Cronbach, 1946). Another strategy is to ensure that response items are directly meaningful to participants. As graduate students progress through various stages of socialization (Gardner, 2009; Weidman et al., 2001), certain experiences or the meanings constructed from them may not hold the same salience at all points in time. Thus, our understanding of the socialization process can be leveraged to enhance response validity by tailoring items to avoid acquiescence due to mismatch with participants' individual stages of socialization. For example, respondents who have entered the Informal or Personal stages of socialization, in which graduate students are conceptualizing themselves more as independent researchers, may be more likely to acquiesce to items asking about their

assessments of coursework, because such activities are typically more salient during the Formal stage (Gardner, 2008b).

The related effect of social desirability occurs when respondents skew their answers to reflect the anticipated preference of the anticipated reader or researcher (Edwards, 1957). For example, if participants blame themselves for a negative graduate school experience or feel that some aspect of that experience reflected badly on them, they would be more likely to offer a mitigated estimate of their experience to protect against an anticipated negative judgment. This phenomenon can be attributed to either "impression management," which focuses on the judgments of a third party, or "self-deceptive enhancement," which focuses on self-judgments (Paulhus, 1991).

Research on social desirability suggests several strategies that may be useful in guarding against its biasing impact on survey items. First, during item development and validation, respondents can be asked to estimate the desirability of the item itself on a Likert scale, which can inform modification decisions for extreme ratings of high or low desirability (Nederhoff, 1985). Second, items likely

to invoke socially undesirable responses may have biasing effects mitigated through the use of projective language, which asks the respondent to evaluate in relation to a hypothetical third party (e.g., "Students in my program are treated as colleagues by the faculty" rather than "I am treated as a colleague by the faculty" [Weidman & Stein, 2003, p. 650]) (Fisher, 1993). Items targeting the intent to pursue or desirability of research-related or faculty careers after earning the Ph.D. may also be appropriate items to use this strategy, as many students report social discomfort in disclosing an aversion to those positions (Nerad, 2015).

Weighing against expectation and experience

When respondents are asked to evaluate their experiences through assessments of sufficiency (e.g., the extent to which something occurs, satisfaction with an event), they necessarily call upon their own frame of reference to respond. For example, Paglis et al. (2006) asked participants to respond to the item "My adviser helps me to meet new colleagues" on a 1-5 Likert scale with anchors of "to a slight extent" and "to a great extent." However, the response to this

item inherently requires the participant to weigh his/her experiences against an expectation of the extent to which the event *should* happen. If the participant expects that such help ought occur monthly, but the advisor engages in the behavior every six months, then the participant's response would likely be a low score (e.g., 1 or 2). In contrast, another participant might expect that an advisor would only help with meeting new colleagues annually, in which case the likely response to the event occurring every six months would likely be positive (e.g., 4 or 5). Similarly, assessing the frequency of receiving feedback about academic progress (Nettles & Millett, 2006) on a scale from "never (1)" to "very often (5)" will depend on how often the respondent feels feedback should be provided. While "never" is directly observable, differentiating between "somewhat often" and "very often" depends upon the ideal frequency in the mind of the individual. Thus, identical responses cannot be assumed to represent the same underlying events. Given that different demographic groups may hold divergent expectations of their mentors (e.g., by gender; Rose, 2005), it is possible that such measurement strategies

may systematically conflate differences in expectation with differences in socialization opportunity.

Self-efficacy vs. performance

Both Nettles and Millett (2006) and Paglis and colleagues (2006) elicit self-efficacy information from participants, asking them to estimate the extent to which they are capable of performing various research tasks. While such perspectives may be valuable in their own right for inferring the extent to which respondents feel capable of success, accepting such estimates as proxy responses for actual skill levels as outcome variables is fundamentally flawed. Extensive research indicates that the correspondence between individuals' beliefs about their skills and their demonstrated skill levels is poor (e.g., Dunning, Johnson, Ehrlinger, & Kruger, 2003; Ehrlinger & Dunning, 2003; Falchikov & Boud, 1989). Further, a meta-analysis by Stajkovic and Luthans (1998) estimated that self-efficacy beliefs predict no more than 25% of the variance in participants' actual performance during low-complexity, artificial tasks. For complex, authentic tasks such as those involved in scholarly research,

self-efficacy account for only 4% of variation in performance. Indeed, Feldon, Maher, Hurst, and Timmerman (2015) found that graduate students in STEM disciplines were unable to estimate their own strengths and weaknesses in research skill at levels better than chance, when compared to both their advisors' assessments of their skills and their rubric-based scores on research proposals they had written. Advisors' assessments likewise failed to predict performance as evaluated by the rubric.

Although skill development is not frequently studied in graduate education (Feldon, Maher, & Timmerman, 2010), it is consistently identified as a fundamental aspect of the socialization process (Austin & McDaniels, 2006; Merton, Reader, & Kendall, 1957). As such, understanding its trajectories and effective strategies for enhancing them is an important aspect of understanding the socialization process. Although this strategy is resource-intensive, employing performance-based assessment strategies can provide direct insight into skill development using authentic disciplinary activities in context. Evaluating written products (e.g., reports of empirical findings, literature reviews, or research proposals) does not

require physical proximity, and a number of validated instruments exist that establish consistent metrics for assessing scholarly rigor and quality (e.g., Boote & Beile, 2005; Feldon et al., 2011; Hackett & Rhoten, 2009; Lovitts, 2007; Timmerman, Strickland, Johnson, & Payne, 2011). Other important skills that may require direct observation or recording, such as effective communication of research to lay audiences, likewise can be assessed reliably using performance-based rubrics (e.g., Sevian & Gonsalves, 2008).

Validation Strategies

Recognizing the importance of local meaning in measurement presents a challenge in validating instruments. It cannot be taken for granted that an instrument valid in one context maintains its validity in another. While there are many aspects of doctoral education that are consistent over time and from institution to institution, several basic influences on socialization processes are not. Both the nature of academic work and the composition of the student population engaged in it has changed dramatically over the past 20

years. For example, the expected pace of productivity has increased substantially over time (Anderson et al., 2011; Austin & McDaniels, 2006). Whereas it was once exceptional for a student to publish more than one journal article prior to commencing his dissertation work (Nettles & Millett, 2006), in many fields, several publications is now the norm expected for gaining access to desirable academic positions after completing the Ph.D. (Ehrenberg, Zuckerman, Groen, & Brucker, 2009). Likewise, in scientific fields, team-based endeavors are increasingly common (Cumming, 2009; Wuchty, Jones, & Uzzi, 2007), resulting in more complex collaboration and mentoring structures for graduate students. The importance of this shift is reflected in the assertion that "cascading mentorship" (Golde et al., 2009) has become a signature pedagogy in many science disciplines. In this context, mentorship now occurs between more varied roles than just between faculty advisor and student: "post-doctoral fellows mentor senior graduate students, senior graduate students mentor junior graduate students, and junior graduate students mentor undergraduates" (p. 57).

Further, the demographic distribution of doctoral students themselves is fundamentally different. Over the past 20 years, the proportion of female U.S. citizens and permanent residents earning doctorates has increased from 44% to 51%. Similarly, the number of doctorates awarded to black/African American students over the past 10 years has increased 31%, and the number awarded to Hispanic/Latino students has increased 71% (NCSES, 2017). While the rates of increase have not yet translated to equitable representation across race and ethnicity, it is nonetheless reflective of changing demographics that may be salient to understanding the dimensions of socialization. Thus, it may not be appropriate to assume that the items and response scales developed to examine socialization decades ago would be wholly appropriate to use in the investigation of socialization now without empirical validation, because socialization processes now take place amongst students and faculty of less homogenous backgrounds. As such, constructs previously conceptualized as unitary may now reflect greater nuance in the underlying constructs.

Sampling

One of the ways in which the validity of instruments might be enhanced is to make efforts to test them across a substantial number of doctoral students in a variety of institutions and disciplines. Some studies have used this approach, such as Nettles and Millett (2006), in which over 6,000 doctoral students across 14 disciplines from 19 institutions provided responses. Analyses of the responses both reflected appropriate reliability metrics and clearly defined factor structures, which delineated the relative strengths of specific items in reflecting underlying factors. Similarly, the instruments used by Paglis and colleagues (2006) were validated with a sample of n=357 incoming doctoral students across 24 departments within a single university and reflected adequate internal consistency ($\alpha >$ 0.70) and robust factor structures (i.e., one factor per subscale) (Green & Bauer, 1995). In contrast, Weidman and Stein (2003) validated their instrument using data from only 50 respondents across two departments within the same institution. Given that correlation coefficients—even when statistically significant—are typically unstable until the sample size exceeds 250 (Schönbrot & Perugini,

2013), reliability estimates (based on intercorrelations amongst items) from small samples are likely to be equally unstable. Because instruments with low reliability are by definition lacking in validity, the consequence of such instability is a low ability to be confident in the instrument's validity. It should be noted, however, that when Weidman and Stein's subscales were used in a study with a larger sample (n=336; within a single discipline across 53 institutions), internal consistency for each subscale (i.e., Cronbach's alpha) was exceptionally high, ranging between 0.883 and 0.976, with most above 0.93 (Feldon, Jeong, et al., 2017). Thus, the primary concern regarding the validation of Weidman and Stein's instrument based on their sampling strategy is a lack of confidence in the stability of the reliability estimates rather than an inherently problematic measure.

Given the strong performance across measures in terms of reliability, it might be argued that validity concerns are exclusively hypothetical in nature. However, several facets of the situation warrant further consideration. First, the psychometric strengths of the instruments do not negate the initial concern that studies performed

using them have not been consistent in supporting the expectations of socialization theory. Second, reliability/internal consistency is a necessary but not sufficient criterion for validity, as stability in measurement can be high while the instrument's conception of the underlying construct may be inadequate. Indeed, the strongest estimate of validity stems from an instrument's ability to predict outcomes for another variable in a manner consistent with the predictions of theory.

Differential item functioning (DIF)

Newer trends in measurement under the assumptions of item-response theory (IRT; Wright & Stone, 2004) have heightened focus on the ways in which measurement items may systematically assess different underlying traits or behaviors for different groups within a sample. Issues of DIF have received almost no attention in the scales commonly used with research in socialization, in part as an issue of historical timing and in part as a reflection of the relatively small role that quantitative research has played in socialization re-

search to data within graduate education. However, given the changing demographics of the doctoral student population, it is an idea that warrants substantial attention. Importantly, DIF does not assess the probability of members of different groups responding differently to items where their experiences or beliefs differ (e.g., members of underrepresented groups reporting less access to research opportunities). Instead, it estimates the probability that groups with the same underlying experience or belief (as might be represented by a total score on an instrument) are differentially likely to select a specific response to a survey item (Holland & Thayer, 1986). Thus, if undetected, the item would introduce bias into the measure by increasing the likelihood of a specific answer that was based on group membership, rather than the facet of socialization targeted. DIF may occur both in terms of the likelihood of eliciting a given response and in terms of the likelihood to choose not to respond to a given item (Dorans, Schmitt, & Bleistein, 1992).

The one exception to this pattern (to the best of the author's knowledge) is a measure of graduate advising experiences developed by Barnes, Chard, Wolfe, Stassen, and Williams (2011). Dur-

ing the validation of this instrument, DIF was assessed as a function of degree level (masters vs. Ph.D.), discipline (using Biglan's [1973] framework for academic disciplines), and gender. Most items with significant DIF reflected disciplinary differences, and a lesser number reflected differences in degree level. Gender was not associated with DIF for any item in the survey, which is noteworthy given the broader concerns about gender inequity in advising experiences (e.g., Noy & Ray, 2012; Rose, 2005; Zhao, Golde, & McCormick, 2007). In their study, Noy and Ray identified women of color as a notably divergent group in their study, but Barnes and colleagues did not assess DIF for race/ethnicity or race/ethnicity by gender interactions—possibly due to low representation of minority groups within their sample. As instrument development and validation efforts move forward in graduate socialization, examining DIF presents itself as a major priority as a way to understand differentiated experiences and perspectives by group, for its own sake and as a way to ensure that quantitative data do not ossify misunderstandings of socialization mechanisms due to undetected influences on item responses.

Future Directions

Although this chapter highlights a number of quantitative measurement challenges facing the study of graduate socialization and, by extension, the further development of the theory, these challenges are not insurmountable. Broadly, they fall into three categories: (1) presenting an appropriately nuanced set of meanings within instruments that correspond to the ways in which students within disciplinary and institutional contexts understand their experiences, (2) avoiding validity challenges introduced through item format, and (3) increasing the frequency of best practices in measurement development, such as checking for DIF. Addressing the latter two is fairly straightforward through increasing awareness within the field and encouraging collaboration with colleagues who specialize in psychometrics and measurement. The first issue, however, requires more fundamental consideration.

Finding the optimal balance between situativity and generalizability is an ongoing challenge across many social science fields.

One possible strategy to resolve this tension would be to move socialization research toward a fully descriptive stance consistent with the exclusive use of qualitative inquiry methods. An argument could be made that the complex nature of the interactions between institution, discipline, and individual give rise to a local causality too nuanced to be adequately investigated using standardized quantitative instruments. However, curtailing the range of inquiry strategies would inherently constrain the potential to expand and refine socialization theory more broadly. It would also limit our ability to make generalizable claims about the mechanisms of socialization, which would limit both the ability to observe system-level patterns linking graduate education experiences to outcomes and the opportunities to make robust recommendations for practice and policy across contexts.

An alternative approach would be to shift our approach to the development of survey instruments to orient more specifically around the constructed meanings of respondents. Deliberate efforts to leverage the strengths of qualitative inquiry in the framing and construction of items may enhance not only the validity of instru-

ments in eliciting underlying constructs, but also advance socialization theory through a stronger integration of quantitative and qualitative scholarship in the field. While a full discussion of mixed methods research approaches is beyond the scope of this chapter, readers are encouraged to examine literature that deliberately engages strategies to retain the richness of qualitative data during instrument development and subsequent quantitative analyses (e.g., Creamer, 2018; Hesse-Biber, 2010).

A potentially productive example of this sort of integration in higher education research lies in the phenomenographic tradition (Åkerlind, 2005; Feldon & Tofel-Grehl, 2018; Marton & Pong, 2005). Initially developed as a qualitative paradigm compatible with critical realism, phenomenography assumes that individuals' conceptions of their experiences can be understood both within personal and collective frames. Thus, constructed meanings are considered within the contexts of the individual interview from which they came, the structural nature of the relevant social relationships, and the broader pool of meanings. Further, phenomenography posits that while there may be a very wide range of personal conceptions

held across individuals, the range is not infinite (Marton, 1994). The relationships that exist between individuals' conceptions and socializing structural influences drive *predictable* variation in individuals' conceptions based on systematic physical and social experiences (Entwistle, 1997). Thus, as qualitative inquiry yields saturation (i.e., no new categories emerging from new data collection), the number of distinct conceptions identified can serve as the foundation of the range of responses offered for closed-ended survey items. The resulting instrument is then conducive to identifying trends generalizable to the natural population under the presumption that the distribution of conceptions encapsulating local meanings identified through qualitative analyses represents the natural range of responses generalizable to the whole population.

In their review of instruments developed using phenomenographic approaches, Micari, Light, Calkins, and Streitweiser (2007) suggest that their value is enhanced by the emphasis on measuring changes in how respondents conceptualize or approach their experiences, beyond behavioral or performance-based metrics. Some of these instruments include the Approaches to Studying Inventory

(Entwistle & Ramsden, 1983; Entwistle & Tait, 1994) , the Approaches to Teaching Inventory (Prosser & Trigwell, 1999; Trigwell & Prosser, 2004), the Approaches and Study Skills Inventory for Students (Tait, Entwistle, & McCune, 1997), the Reflections on Learning Inventory (Meyer, 2000), and the Conceptions of Learning Inventory (Purdie & Hattie, 2002). In addition to providing indicators of frequency and magnitude for specific conceptions held by respondents, these instruments are also valuable in their ability to inform understanding of constellations and predictors of conceptions through statistical relationships. For example, examining the relationships amongst conceptions within their instrument's constructs, Trigwell & Prosser (1996) identified a significant and unexpected correlation between conceptions that they had initially grouped differently based on their inductive qualitative analyses. As a result, they restructured the items to reflect a different factor structure and enhance the ability of the survey to capture respondents' underlying conceptions, further improving the instrument's validity. Similarly, Crawford, Gordon, Nicholas, and Prosser (1998a, 1998b) correlated data from a phenomenographic instrument assessing students' con-

ceptions of mathematics with scores on an approach-to-learning questionnaire to identify underlying structural relationships that may not have been immediately evident through exclusively qualitative inquiry.

Another promising approach entails an interactive mixed methods embedded design (Tashakkori & Newman, 2010), in which item development begins with a rigorous sequence of qualitative research strategies to elicit not only individual meanings regarding target constructs, but also iterative member checking strategies and intentional focus on contextual influences on interpretation (David, Hitchcock, Ragan, Brooks, & Starkey, 2018). Cole, Kitchen, and Kezar (2018) engaged this approach using an 8-step process primarily through focus groups conducted with participants in the comprehensive college transition program during site visits at participating campuses. During the iterative process of exchange between researchers with respective quantitative and qualitative expertise over the course of the project, Cole and colleagues sought opportunities to formulate new survey items for validation that emerged unexpectedly during focus group discussions. They also closely examined

underlying meanings expressed during those conversations to understand unexpected or surprising variance or limited response ranges from piloted survey items using an IRT framework.

Using a similar approach with more varied modes of qualitative data collection, David and colleagues (2018) describe in depth their process as they developed and validated a new instrument to measure athlete-trainer trust via Rasch modeling (Bond & Fox, 2015). In the initial phase of development, the authors conducted semi-structured interviews with a number of participants, which is a common practice. However, they went on to engage multiple trustworthiness strategies, including member checking, engaging an external auditor, and reflexive journaling to identify potential interpretational bias on the part of the research team. Identified themes were discussed with a subset of interviewees subsequent to their development during member checking, and raw transcripts were provided to the auditor for an independent coding scheme to be developed and compared with the original findings of the team.

In the next phase, items were developed and piloted with 75 participants for validity and reliability analysis using both Rasch

modeling and classical test theory techniques. Items with suboptimal statistical characteristics were then interrogated with participants using highly focused "rapid reconnaissance" cognitive interviews to determine potential nuances in meaning that limited the ability of problematic items to optimally measure target constructs. Thus, David and colleagues (2018, p. 86) engaged a mixed methods approach that extended beyond a conventional "QUAL→QUAN" approach to one they describe as "QUAL→QUAN↔QUAL↔QUAN," yielding highly robust and nuanced items grounded in the situated understanding of their participants.

Conclusion

Theory and measurement have an intrinsic reciprocal relationship: Theory asserts the structure and function of mechanisms that give rise to specific outcomes, and measures reflect the nature of those mechanisms and capture the range of possible outcomes suing metrics that are meaningful within the interpretive context of the

theory. Conversely, analysis of the data collected through measurement further informs the development of the theory by making evident the ways in which outcomes were or were not consistent with the predictions of the theory. In this way, these two components of scholarly inquiry shape each other.

In the case of graduate socialization, the causal explanations offered through extensive qualitative research have outpaced the capacity for causal description offered by quantitative inquiry. As a natural consequence, various well-known instruments developed to facilitate such description do not fully reflect the insights that have refined our understanding of the mechanisms impacting graduate education outcomes. Thus, the ability of these instruments to support tests of the theory and new contributions to its development is constrained.

As graduate education contexts and populations continue to evolve, understanding its mechanisms and products becomes ever more important to ensure more effective and more equitable outcomes. To aid students in pursuing graduate degrees as a means to enhance their upward economic mobility (Posselt & Grodsky, 2017)

and address societal challenges (Cherwitz & Sullivan, 2002), socialization theory has a vital role to play through its ability to inform the shaping of both structures and functions. Enhancing and refining its contributions requires both further understanding of socialization mechanisms and sustained development of tools that can contribute to it.

Acknowledgments

The author gratefully acknowledges the support of the National Science Foundation. This material is based upon work supported under Award 1431234. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Åkerlind, G. S. (2005). Variation and commonality in phenomenographic research methods. *Higher Education Research and Development, 24*, 321-334.

Anderson, W. A., Banerjee, U., Drennan, C. L., Elgin, S. C. R., Epstein, I. R., et al. (2011). Education at research universities. *Science, 331*, 153-153.

Austin, A. E. (2002). Preparing the next generation of faculty: Graduate school as socialization to the academic career. *The Journal of Higher Education, 73*, 94-122.

Austin, A. E., & McDaniels, M. (2006). Preparing the professoriate of the future: Graduate student socialization for faculty roles. *Higher Education: Handbook of Theory and Research, 21*, 397-456.

Barnes, B. J., Chard, L. A., Wolfe, E. W., Stassen, M., & Williams, E. A. (2011). An evaluation of the psychometric properties of the Graduate Advising Survey for Doctoral Students. *International Journal of Doctoral Studies, 6*, 1-17.

Biglan, A. (1973). Relationships between subject matter characteristics and the structure and outputs of university departments. *Journal of Applied Psychology, 57,* 204-213.

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly, 50,* 240-250.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences (3rd ed.).* New York: Routledge.

Boote, D. N., & Beile, P. (2005). Schoalrs before researchers: On the centrality of the dissertation literature review in research preparation. *Educational Researcher, 34,* 3-15.

Bragg, A. K. (1976). *The socialization process in higher education.* ERIC/AAHE Research Report, no. 7. Washington, DC: American Association for Higher Education.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist, 32,* 513-531.

Burrell, M., & Morgan, G. (2016). *Sociological paradigms and organizational analysis: Elements of the sociology of corporate life.* New York: Routledge.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42,* 1150-1152.

Cherwitz, R. A., & Sullivan, C. A. (2002). Intellectual entrepreneurship: A vision for graduate education. *Change,* 34(6), 23-27.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

Cole, D., Kitchen, J. A., & Kezar, A. (2018). Examining a comprehensive college transition program: An account of iterative mixed methods longitudinal survey design. *Research in Higher Education.* https://doi.org/10.1007/s11162-018-9515-1.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60,* 151-174.

Crawford, K., Gordon, S., Nicholas, J., & Prosser, M. (1998a). University mathematics students' conceptions of mathematics. *Studies in Higher Education, 23,* 87-94.

Crawford, K., Gordon, S., Nicholas, J., & Prosser, M. (1998b). Qualitatively different experiences of learning mathematics at university. *Learning & Instruction, 8,* 455-468.

Creamer, E. G. (2018). Enlarging the conceptualization of mixed method approaches to grounded theory with intervention research. *American Behavioral Scientist, 62,* 919-934.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494.

Cumming, J. (2009). The doctoral experience in science: Challenging the current orthodoxy. *British Educational Research Journal, 35,* 877-890.

Cummins, R. A., & Gullone, E. (2000). *Why we should not use 5-point Likert scales: The case for subjective quality of life measurement*. Paper presented at the Second International Conference on Quality of Life in Cities, Singapore: National University of Singapore.

David, S. L., Hitchcock, J. H., Ragan, B., Brooks, G., & Starkey, C. (2018). Mixing interviews and Rasch modeling: Demonstrating a procedure to develop an instrument that measures trust. *Journal of Mixed Methods Research, 12,* 75-94.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50,* 61-77.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29,* 309-319.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83-87.

Edwards, A. (1957). *The social desirability variable in personality assessment and research.* New York: The Dryden Press.

Ehrenberg, R., Zuckerman, H., Groen, J., & Brucker, S. (2009). *Educating scholars: Doctoral education in the humanities.* Princeton University Press.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*, 5-17.

Entwistle, N. (1997). Introduction: Phenomenography in higher education. *Higher Education Research & Development, 16,* 127-134.

Entwistle, N.J., & Ramsden, P. (1983). *Understanding student learning.* New York: Nichols.

Enwistle, N.J., & Tait, H. (1994). *The revised approaches to studying inventory*. Edinburgh, Scotland: University of Edinburgh, Centre for Research into Learning and Instruction.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59,* 395-430.

Feldon, D. F., Jeong, S., Peugh, J., Roksa, J., Maahs-Fladung, C., Shenoy, A., & Oliva, M. (2017). Null effects of boot camps

and short-format training for Ph.D. students in life sciences. *Proceedings of the National Academy of Sciences, 114*(37), 9854-9858.

Feldon, D. F., Maher, M. A., Hurst, M., & Timmerman, B. (2015). Faculty mentors', graduate students', and performance-based assessments of students' research skill development. *American Educational Research Journal, 52,* 334-370.

Feldon, D. F., Maher, M., Roksa, J., & Peugh, J. (2016). Cumulative advantage in the skill development of STEM graduate students: A mixed methods study. *American Educational Research Journal, 53,* 132-161.

Feldon, D. F., Maher, M., & Timmerman, B. (2010). Performance-based data in the study of STEM graduate education. *Science, 329,* 282-283.

Feldon, D. F., Peugh, J., Maher, M. A., Roksa, J., & Tofel-Grehl, C. (2017). Effort-to-credit gender inequities of first-year PhD students in the biological sciences. *CBE-Life Sciences Education, 16*(1), ar4.

Feldon, D. F., Peugh, J., Timmerman, B. E., Maher, M. A., Hurst, M., Strickland, D., Gilmore, J. A., & Stiegelmeyer, C. (2011). Graduate students' teaching experiences improve their methodological research skills. *Science, 333*(6045), 1037-1039.

Feldon, D. F., & Tofel-Grehl, C. (2018). Phenomenography as a foundation for mixed models research. *American Behavioral Scientist, 62,* 887-899.

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20,* 303-315.

Fraenkel, J. R., & Wallen, N. E. (1996). *How to design and evaluate research in education.* New York: McGraw Hill.

Gardner, S. K. (2008a). What's too much and what's too little?: The process of becoming an independent researcher in doctoral education. *The Journal of Higher Education, 79,* 326-350.

Gardner, S. K. (2008b). Fitting the mold of graduate school: A qualitative study of socialization in doctoral education. *Innovative Higher Education, 33,* 125-138.

Gardner, S. K. (2009). *The development of doctoral students: Phases of challenge and support*. ASHE Higher Education Report 34, no. 6. Washington, D.C.: Association of the Study of Higher Education.

Gardner, S. K. (2010). Contrasting the socialization experiences of doctoral students in high- and low-completing departments: A qualitative analysis of disciplinary contexts at one institution. *The Journal of Higher Education, 81*, 61-81.

Gardner, S. K., & Mendoza, P. (2010). *On becoming a scholar: Socialization and development in doctoral education*. Sterling, VA: Stylus Publishing, Inc.

Glaser, R., & Strauss, R. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.

Golde, C. M., Conklin Bueschel, A., Jones, L., & Walker, G. E. (2009). Advocating apprenticeship and intellectual community: Lessons from the Carnegie Initiative on the Doctorate. In R. G. Ehrenberg & C. V. Kuh (Eds.), *Doctoral education and faculty of the future* (pp. 53–64). Ithaca, NY: Cornell University Press.

Green, S. G., & Bauer, T. N. (1995). Supervisory mentoring by advisers: Relationships with doctoral student potential, productivity, and commitment. *Personnel Psychology, 48,* 537-562.

Hackett, E. J., & Rhoten, D. R. (2009). The Snowbird charrette: Integrative interdisciplinary collaboration in environmental research design. *Minerva, 47,* 407-440.

Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice.* New York: Guilford Press.

Holland, P.W., & Thayer, D.T. (1986) *Differential item performance and the Mantel-Haenszel procedure (Technical Report No. 86-69).* Princeton, NJ: Educational Testing Service.

Holley, K. A. (2009). The challenge of an interdisciplinary curriculum: A cultural analysis of a doctoral-degree program in neuroscience. *Higher Education, 58,* 241-255.

Holley, K. A. (2018). The longitudinal career experiences of interdisciplinary neuroscience PhD recipients. *The Journal of Higher Education, 89,* 106-127.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38,* 1217-1218.

Kelly, S. (2017). Shared principles of causal inference in qualitative and quantitative research. In D. Wyse, N. Selwyn, E. Smith, & L. Suter (Eds.), *The BERA/SAGE Handbook of Educational Research* (pp. 90-115). Los Angeles, CA: Sage.

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., et al. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 66,* 371-403.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Lovitts, B. E. (2001). *Leaving the ivory tower: The causes and consequences of departure from doctoral study.* Lanham, MD: Rowman & Littlefield.

Lovitts, B. E. (2007). *Making the implicit explicit: Creating performance expectations for the dissertation.* Sterling, VA: Stylus Publishing, LLC.

Lovitts, B. E. (2008). The transition to independent research: Who makes it, who doesn't, and why. *The Journal of Higher Education, 79,* 296-325.

Marton, F. (1994). Phenomenography. In T. Husén, & T. N. Postlethwaite (Eds.), *The international encyclopedia of education (2nd ed., Vol. 8).* Oxford, UK: Pergamon.

Marton, F., & Pong, W. Y. (2005). On the unit of description in phenomenography. *Higher Education Research & Development, 24*(4), 335-348.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review, 62,* 279-299.

Maxwell, J. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33,* 3-11.

Maxwell, J. A., & Mittapalli, K. (2010). Realism as a stance for mixed method research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (2nd ed., pp. 145–168). Thousand Oaks, CA: Sage.

Merton, R. K., Reader, G., & Kendall, P. L. (1957). *The student physician.* Cambridge, MA: Harvard University Press.

Messick, S. (1967). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set*

in personality assessment (pp. 115-145). Chicago, IL: Aldine.

Meyer, J.H.F. (2000, September). *An overview of the development and application of the Reflections on Learning Inventory (RoLI)*. Paper presented at the 1st RoLI Symposium, Imperial College, University of London.

Micari, M., Light, G., Calkins, S., & Streitwesier, B. (2007). Assessment beyond performance: Phenomenography in educational evaluation. *American Journal of Evaluation, 28,* 458-476.

Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: A sourcebook of new methods.* Newbury Park, CA: Sage.

National Center for Science and Engineering Statistics. (2017). *2015 doctorate recipients from U.S. universities.* Washington, DC: National Science Foundation. https://www.nsf.gov/statistics/2017/nsf17306/static/report/nsf17306.pdf

Nederhoff, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15,* 263-280.

Nerad, M. (2015). Searching for taboos in doctoral education: An exploratory journey. *Die Hochschule [German Journal for Science and Education], 24*, 10-25.

Nettles, M. T., & Millet, C. M. (2006). *Three magic letters: Getting to Ph.D.* Baltimore, MD: Johns Hopkins University Press.

Noy, S., & Ray, R. (2012). Graduate students' perceptions of their advisors: Is there systematic disadvantage in mentorship?. *The Journal of Higher Education, 83*, 876-914.

Paglis, L.L., Green, S.G., & Bauer, T. N. (2006). Does adviser mentoring add value? A longitudinal study of mentoring and doctoral student outcomes. *Research in Higher Education, 47,* 451-476.

Patton, M. Q. (2002). *Qualitative research and evaluation (3rd ed.).* Thousand Oaks, CA: Sage.

Paulhus, D.L. (1991). Measurement and control of response biases. In J.P. Robinson et al. (Eds.), *Measures of personality and*

social psychological attitudes. San Diego, CA: Academic Press.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement design and analysis: an integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pole, C. (2000). Technicians and scholars in pursuit of the PhD: Some reflections on doctoral study. *Research Papers in Education, 15*, 95-111.

Posselt, J. R., & Grodsky, E. (2017). Graduate education and social stratification. *Annual Review of Sociology, 43,* 353-378.

Prosser, M., & Trigwell, K. (1999). *Understanding learning and teaching: The experience in higher education.* Buckingham, UK: Society for Research into Higher Education and Open University Press.

Purdie, N., & Hattie, J. (2002). Assessing students' conceptions of learning. *Australian Journal of Educational & Developmental Psychology, 2,* 17-32.

Reybold, L. E. (2003). Pathways to the professoriate: The development of faculty identity in education. *Innovative Higher Education, 27,* 235-252.

Rojas, F. (2017). *Theory for the working sociologist.* New York: Columbia University Press.

Roksa, J., Feldon, D. F., & Maher, M. (in press). First-generation students in pursuit of the Ph.D.: Comparing socialization experiences and outcomes to continuing-generation peers. *Journal of Higher Education.*

Roksa, J., Jeong, S., Feldon, D., & Maher, M. (in press). Socialization experiences and research productivity of Asians and Pacific Islanders: 'Model minority' stereotype and domestic vs. international comparison. *Research in Sociology of Education.*

Rose, G. L. (2005). Group differences in graduate students' concepts of the ideal mentor. *Research in Higher Education, 46,* 53-80.

Schönbrot, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47,* 609-612.

Schuman, H., & Presser, S. (1981). *Questions and answers: Experiments on question form, wording, and context in attitude surveys.* New York: Academic Press.

Schwandt, T. A. (1997). *Qualitative inquiry: A dictionary of terms.* Thousand Oaks, CA: Sage.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 95-105.

Sevian, H., & Gonsalves, L. (2008). Analysing how scientists explain their research: A rubric for measuring the effectiveness of scientific explanations. *International Journal of Science Education, 30,* 1441-1467.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* New York: Houghton Mifflin Company.

Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin, 124*, 240-261.

Tait, H., Entwistle, N.J., & McCune, V. (1997). ASSIST: A reconceptualisation of the Approaches to Studying Inventory. In C. Rust (Ed.), *Improving student learning: Improving students as learners* (pp. 262-271). Oxford, UK: Oxford Brookes University, Oxford Centre for Staff and Learning Development.

Tashakkori, A., & Newman, I. (2010). Mixed methods: Integrating quantitative and qualitative approaches to research. In B. McGaw, E. Baker, & P. P. Peterson (Eds.), *International encyclopedia of education* (3rd ed., pp. 514-520). Oxford, UK: Elsevier.

Thornton, R., & Nardi, P. M. (1975). The dynamics of role acquisition. *American Journal of Sociology, 80*, 870-885.

Tierney, W. G., & Rhoads, R. A. (1994). *Faculty socialization as cultural process: A mirror of institutional commitment.*

ASHE-ERIC Higher Education Report, no. 93-6. Washington, DC: The George Washington University.

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a 'universal's rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education, 36*, 509-547.

Trigwell, K., & Prosser, M. (1996). 'Changing approaches to teaching: A relational perspective', *Studies in Higher Education, 21*, 275-284.

Trigwell, K., & Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educational Psychology Review, 16*, 409-424.

Van Maanen, J., & Schein, E. (1979). Toward a theory of organizational socialization. In B. M. Straw (Ed.), *Research in organizational behavior,* (pp. 209-264). Greenwich, CT: JAI Press.

Weidman, J. C. & Stein, E. L. (2003). Socialization of doctoral students to academic norms. *Research in Higher Education, 44*, 641-656.

Weidman, J. C., Twale, D. J., & Stein, E. L. (2001). *Socialization of graduate and professional students: A perilous passage?* ASHE-ERIC Higher Education Report 28, no. 3. Washington, D.C.: Association of the Study of Higher Education.

Wright, B. D., & Stone, M. H. (2004). *Making measures.* Chicago, IL: The Phaneron Press.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039.

Wulff, D. H., Austin, A. E., Nyquist, J. D., & Sprague, J. (2004). The development of graduate students as teaching scholars: a four-year longitudinal study. In D. Wulff & A. Austin (Eds.), *Paths to the professoriate: strategies for enriching the preparation of future faculty*. San Francisco: Jossey-Bass

Zhao, C. M., Golde, C. M., & McCormick, A. C. (2007). More than a signature: How advisor choice and advisor behaviour affect

doctoral student satisfaction. *Journal of Further and Higher Education, 31*, 263-281.