Understanding the scholarly development of Ph.D. students in science, technology, engineering, and mathematics (STEM) is vital to the preparation of the scientific workforce. During doctoral study, students learn to be professional scientists and acquire the competencies to succeed in those roles. However, this complex process is not well studied. Research to date suffers from overreliance on a narrow range of methods that cannot provide data appropriate for addressing questions of causality or effectiveness of specific practices in doctoral education. We advocate a shift in focus from student and instructor self-report toward the use of actual performance data as a remedy that can ultimately contribute to improved student outcomes.

Developing causal models that account for individual differences in skill outcomes is especially challenging, as students enter Ph.D. programs with highly variable sets of prior experiences, expectations, and available means of support. Disciplinary research skills are perhaps the most critical outcomes, as they underlie the quality of scholarly work. Other factors, like creativity and motivation, do play a role, but they require a strong foundation in appropriate theoretical and methodological knowledge to yield useful products (1).

Although scholarly attention to doctoral education has increased substantially in the last decade, it remains a minor presence (2). This is especially problematic given concerns about the capacity of the STEM workforce and the current emphasis on accountability within higher education. Opinions differ regarding the merits of the accountability movement, but the increased demand for evidence of effectiveness warrants reexamination of assumptions regarding current Ph.D. training practices.

Methods of Doctoral Education Research

Accurately gauging the effectiveness of STEM doctoral research training is essential for informed decisions about its improvement. It requires unambiguous data that reflect the development of key disciplinary competencies for students at all stages of their degree programs. However, there is a startling paucity of data on the skills that students have at program entry, the trajectories of skill development during their programs, and the extent to which skills acquired during doctoral study are applicable in interdisciplinary contexts (3).

Current knowledge of effectiveness in the doctoral education process relies almost entirely on anecdotal and self-report data: individual reflection [e.g., (4)]; interviews [e.g., (5, 6)]; and surveys [e.g., (7)]. Although self-reports can provide valid data regarding participants’ attitudes, values, beliefs, and past behaviors, they are not necessarily accurate or objective measures of performance or underlying mental processes [e.g., (8, 9)]. Similarly, professors’ estimates of student ability and performance are frequently biased (10).

Studies examining Ph.D. students’ research skills as an outcome of their doctoral training typically rely on opinion-based assessments of readiness to conduct research independently (5–7, 11); coarse-grained student funding or publication data (12); or broad reputational assessments of degree programs or departments (13), which provide no insight into individual development. Occasionally, dissertations serve as proxies for student performance, but mentor and peer involvement with students’ work and the use of variable standards in evaluation limit inferences about the relative strengths and weaknesses of an individual student’s skills other than through personal opinion (14, 15). Individual course grades and grade point averages are not effective metrics because of problems with validity and generalizability across institutions, programs, and instructors (16).

Other approaches examine employment following degree attainment (e.g., U.S. National Science Foundation Survey of Doctorate Recipients). However, academic hiring decisions are directly influenced by factors reflecting neither an individual’s skill level nor the quality of their training (e.g., the perceived fit of the research agenda, personality, and teaching ability to the needs of the hiring department and the qualifications of competing candidates). Further, STEM fields change as the central research questions evolve, the demand for experts shifts among specialties, and the availability of financial resources changes, limiting the conclusions that can be drawn from employment outcomes.

Education Performance and Processes

In contrast with other educational levels, such as kindergarten to high school (K–12) or undergraduate studies, performance-based behavioral and cognitive investigations of research skills among STEM Ph.D. students are almost nonexistent. Some empirical studies of scientific problem-solving skills include graduate student populations [e.g., (17)], but sample sizes are small, STEM-specific disciplinary skills are not the focus, and the impacts of doctoral training on skill development are largely ignored. Such investigations occur at one time point and cannot capture longitudinal skill development or knowledge integration. In contrast to medical education (18), for example, no research examines trends and individual differences in doctoral students’ developmental trajectories toward expertise within their respective disciplines. Thus, data cannot be leveraged to refine training processes and to evaluate students against robust baseline models. This difference may be due in part to the common use of competency standards in both K–12 and professional education, which represent stakeholder consensus on criterion-based assessments. Such agreement is not commonly explicit at the doctoral level, but several such projects have been initiated [e.g., (14)].

It is commonly assumed that doctoral
students acquire research expertise through direct tutelage from and collaboration with their research supervisors or major professors. Although a causal relationship is often asserted (19), solid empirical evidence is scarce. The frequently mentioned influence of peers as mentors (20), concerns about the limited accuracy with which experts can articulate their problem-solving strategies (21), experts’ difficulty accurately gauging the abilities and instructional needs of novices (17), and the importance of independent trial and error in students’ process of learning scientific practice (J2, 22) undermine the sufficiency of the traditional apprenticeship model. The perceived quality of mentoring correlates positively with the number of student publications and negatively with attrition (6, 22). However, none of these studies differentiate influences of the faculty mentor, research opportunities afforded individual students, and support for early publication efforts.

Context of Accountability
Across educational levels, accountability pressures require policy decisions to be data-driven. The most visible of these efforts have resulted in widespread standardized testing in K–12 education and an increased use of tests at the undergraduate level. However, such methods are inappropriate for assessing doctoral student outcomes for two reasons. First, research is inherently a generative activity for which quality is judged in part on novelty rather than conformity. Second, doctoral programs are typically highly specialized according to the research interests and expertise of the individual faculty members. Meaningful programmatic or institutional comparisons are limited by the availability of a given specialty across institutions. For these reasons, postsecondary accountability efforts may be most productive when the performance data collected are used for research and formative assessment (i.e., ongoing, internal, and low stakes), rather than for external comparison. Rough proxies of learning, like program completion rates or student satisfaction survey data (23), cannot provide comparable utility for programmatic improvement.

In this context, a need exists for valid, reliable, performance-based data to assess both students’ skill development and the educational practices that enhance it. Data collected through instruments such as criterion-based rubrics (24) for core competencies within or across STEM disciplines can enable doctoral programs to demonstrate students’ competencies empirically to accreditation groups and future employers. They can also identify which features of programs have the most impact [e.g., (14)]. Without this basic foundation of empirical assessment, we cannot reliably differentiate between effective and ineffective doctoral education practices nor determine which training investments will maximize the return of human capital.

Future Directions
Concerns about competitiveness in STEM fields and about educational systems’ capacity to meet anticipated demands for a qualified STEM workforce persist. The number of STEM degrees earned is often a focal issue that is coupled with demands for innovation to accelerate economic growth (25). However, the number of STEM degree holders does not inherently enhance capacity for innovation. Degree attainment is a function of completing academic program requirements, but groundbreaking research occurs as a function of knowledge, skill, and preparation for entry into rapidly changing fields, among other factors (26).

Although some factors influencing individuals’ entry into the STEM workforce cannot be directly manipulated, the training practices that shape knowledge and skill development during doctoral education are determined directly by the decisions of degree programs and individual faculty. Many practices are legacies of tradition, personal history, and disciplinary culture. However, basing these decisions on a foundation of performance-based empirical evidence for effectiveness permits the validation of techniques and a basis for reformulation of others.

To identify optimal mechanisms for enhancing research skills through doctoral training, we must accurately identify and measure the core competencies needed for expert performance within and possibly across STEM disciplines. This requires the development of carefully aligned performance-based rubrics or other instruments that permit the disaggregation and assessment of pertinent skills as they develop over time. Successful projects to date build on two key elements: sustained prioritization and support from university leadership, and core groups of faculty committed to distilling consensus around discrete criteria for levels of performance [e.g., (14, 22, 27)]. One useful approach may be to identify an existing rubric used at the undergraduate level and to expand its scope to accommodate doctoral student performance.

Without longitudinal or other sequential analyses of demonstrated research skill development during doctoral training, one cannot fully understand the phenomenon as either a connected set of cognitive events or an attained process-driven competency within a professional context. By conducting these studies and utilizing their results, STEM doctoral training can be refined to more effectively broaden participation and reduce barriers to entry in STEM professions, increase retention and graduation rates of qualified students, and maximize the impact that students’ doctoral training has on their research skill development and subsequent productivity.

References and Notes
5. A. Austin, J. Higher Educ. 73, 94 (2002).
24. A rubric identifies essential dimensions of performance for a task (e.g., research proposal or dissertation) and describes the criteria that delineate multiple (usually three to five) levels of performance for each. Validation requires consensus among disciplinary experts and stakeholders (e.g., faculty mentors, research supervisors, and instructors) for performance criteria, strong interrater reliability when used with diverse instances of the target performance, and convergent evidence with other indicators that target skills are accurately measured. Extended discussions are available in (24) and (27).
28. This work is supported by a grant from the NSF to the authors (NSF-0723686). The views expressed do not necessarily represent the views of the supporting funding agency.