Studies in Higher Education, 2013
Vol. 38, No. 5, 693–710, http://dx.doi.org/10.1080/03075079.2011.590971

Routledge
Taylor & Francis Group

Performance-based assessment of graduate student research skills: timing, trajectory, and potential thresholds

Briana Crotwell Timmerman[a*], David Feldon[b], Michelle Maher[c], Denise Strickland[d,e] and Joanna Gilmore[f]

[a]*Office of Research and Graduate Education and Department of Biological Sciences, University of South Carolina, Columbia, SC, USA;* [b]*Department of Curriculum, Instruction and Special Education, University of Virginia, Charlottesville, VA, USA;* [c]*Department of Educational Leadership and Policies, University of South Carolina, Columbia, SC, USA;* [d]*Center for the Advanced Study of Teaching and Learning in Higher Education, University of Virginia, Charlottesville, VA, USA;* [e]*Department of Educational Studies, University of South Carolina, Columbia, SC, USA;* [f]*Center for Teaching and Learning, University of Texas at Austin, Austin, TX, USA*

The development of research skills and scientific reasoning underpins the mission of graduate education in science, technology, engineering and mathematics (STEM) fields, yet our understanding of this process is mainly drawn from self-report and faculty survey data. In this study, we empirically investigate the pattern of research skill development using STEM graduate students' written research proposals. Analyses of proposal performance data suggest a potential developmental trajectory of research skills, in which the ability to effectively situate work in context using primary literature, and to generate testable hypotheses, emerge early in students' careers, while other skills, such as data analysis and forming conclusions from data, appear to develop later. We discuss these findings in relation to threshold concepts theory, a framework which posits that intellectual growth occurs in transformative leaps rather than a gradual progression, especially as it applies to graduate student research skill development.

Keywords: graduate education; higher degree research; performance assessment; threshold concepts

Introduction

The development of scientific research skills is a core goal of science education across secondary and post-graduate educational levels in the United States (US), United Kingdom, and Australia (Kiley 2008; National Research Council 1996; Roberts 2002). Variously termed 'inquiry skills' (Lawson 2008; Marx et al. 2004), 'scientific reasoning' (Keys 1994; Zimmerman 2000) or 'scientific thinking' (Dunbar 2000), most science programs are concerned with helping students develop the cognitive skills and abilities employed by practising scientists (Roberts 2002; Seymour 2001).

Extensive literature exists on how to encourage such thinking at the primary and secondary levels (Hofstein and Lunetta 2004; Sandoval and Reiser 2004; Schroeder et al. 2007), and a nascent literature base exists at the undergraduate level (DeHaan

*Corresponding author. Email: timmerman@sc.edu

2005; Seymour 2001; Willison and O'Regan 2007). Previous studies also examine the research skills of practising scientists (Dunbar 2000; Feldon 2007, 2010; Feldon et al. 2010; Schunn and Anderson 1999; Zimmerman 2000). However, examination of skill development at the level of graduate education is extremely limited. The only published *performance-based* assessments of graduate research skills known to the authors are either a measure of collaborative performance rather than individual skill (Hackett and Rhoten 2009), or measure either graduate students' ability to explain their research to middle-school students (Sevian and Gonsalves 2008) or their acquisition of statistical skills (Onwuegbuzie 2003). Thus, our understanding of graduate students' professional development is based on perceptions or indirect measures of program quality (Mervis 2000). For example, graduate student development is currently assessed using student or faculty perceptions of student ability as reported in interviews and surveys (Carnegie Initiative on the Doctorate 2001; Golde 2001; Halonen et al. 2003), or is based on perceptions of university or program reputation and/or faculty research productivity (Davis and Fiske 2001; Ostriker and Kuh 2003). We lack detailed, direct measures of the research skills commonly possessed by students when they enter graduate school, how those skills change over time, and the relative strength of those skills upon graduation (Feldon, Maher, and Timmerman 2010).

Recently, Kiley and Wisker (2009; Kiley 2009) attempted to characterize the cognitive development of graduate students as they transition from disciplinary novices into independent scientists using faculty mentors' perceptions of graduate student development. Invoking the notion of threshold concepts previously used to characterize undergraduate development (Land, Meyer, and Smith 2008; Meyer and Land 2006), they suggest that graduate students' research skills develop in a similar fashion. Threshold concepts are described as 'concepts that are so critical to an understanding of the discipline that advanced disciplinary learning is not possible … [until] the threshold of understanding for that concept [is crossed]' (Kiley 2009, 297). Attainment of these skills is considered to be neither gradual nor linear, and requires mastery of certain knowledge or skills prior to the attainment of others. Specifically, Land, Meyer and Smith (2008, x, emphasis in original) define thresholds as concepts which are:

> *transformative* (occasioning a significant shift in the perception of a subject), *integrative* (exposing the previously hidden inter-relatedness of something), likely to be *irreversible* (unlikely to be forgotten, or unlearned only through considerable effort), and frequently *troublesome*, for a variety of reasons . . . Threshold concepts also tend to be *bounded* in that they serve as boundary-markers for the conceptual spaces that constitute disciplinary terrain.

Threshold concepts, therefore, are both obstacles and significant opportunities for intellectual gain. Land, Meyer, and Smith (2008, xi) suggest that much of the value in threshold concepts as a theoretical framework is that they are 'both explanatory and "actionable" (capable of translation into action)'. As such, they offer a potentially valuable basis for making instructional decisions that can enhance graduate education in science, technology, engineering and mathematics (STEM) disciplines.

Through interviews with faculty mentors, Kiley (2009) identifies several threshold concepts marking the development of doctoral students' scientific reasoning skills. She suggests that before becoming effective scholars, graduate students must realize the need to: (1) construct an '*argument* or thesis supported by defensible evidence' (298) that accounts for conflicting views and data, (2) generate a *theoretical model* that allows the findings to be applicable in other situations, and (3) articulate an

awareness of the *conceptual framework* or intellectual context from which the work arises (emphasis in original, but paraphrased). Kiley's conclusions are based solely on interviews with faculty mentors, however, and lack assessment of students' actual performance or growth over time. Given experts' limited success in estimating the learning needs and processes of novices (e.g. Hinds 1999), we examine these proposed threshold concepts through an empirical (rather than phenomenological) lens.

We investigate the developmental trajectory of research skills of STEM graduate students in the early stages of their programs by evaluating changes in the quality of written research proposals over the course of an academic year. Written research proposals are an authentic source of data that align directly with students' academic and professional goals, and provide clear demonstrations of how students approach and frame problems (Hackett and Rhoten 2009).

Research questions

Research proposals used in this study provide an empirical assessment of graduate students' abilities to conceptualize and communicate their research. Proposal data allow inferences regarding initial strengths and weaknesses in scientific reasoning, and normative progressions of learning for this population. Because the sample predominantly includes first- and second-year graduate students (masters and doctoral), this study provides insights into the initial stages of research skill development during graduate school. Our investigation, therefore, targets those threshold concepts and skills which evidence change in these early stages. We address the following questions:

(1) Do the research skills of early stage graduate students differ as a function of previous research experiences?
(2) Do graduate students' research skills improve concurrently (all skills increase similarly over time) or asynchronously (some skills show gains before others)?
(3) If skills improve asynchronously, do the developmental patterns align with previously identified threshold concepts?

Methodology

Graduate student sample

Participants (n = 100) were enrolled in research-intensive graduate programs in STEM disciplines. Seventy-three were enrolled in doctoral or masters level programs at a large research-intensive university in the southeastern US. The remaining 27 participants were enrolled in master's degree programs at one of two smaller masters level institutions in the eastern US (Table 1). All were participating in a larger study examining the development of teaching and research skills of STEM graduate students, supported by the US National Science Foundation award #0723686 (2007–2010). For their participation in the larger study, participants received a financial stipend.

Participants varied in the amount of prior research experience they had upon entering the study. The extent of that experience was assessed during semi-structured interviews conducted as part of the larger study. Research experiences occurred mostly during students' undergraduate careers and graduate careers, but also included relevant experiences during high school or in industry. Participants' reported durations of research experience varied from none to more than 12 semesters (semesters are equivalent to

Table 1. Distribution of participants by area of study and graduate program educational level.

	Doctoral (R)	Masters (R)	Masters (M)	Total
Science	36	10	27	73
Engineering	16	8		24
Mathematics/Statistics	3			3
Total	55	18	27	100

Note: (R) indicates the large, research-intensive institution at which participants were enrolled in either PhD or masters degrees. (M) indicates the smaller masters-granting institutions at which participants were enrolled in research-focused masters degree programs. Academic programs categorized as science include biology, biotechnology, chemistry, geography, geology and marine sciences. Academic programs categorized as engineering include chemical, civil, mechanical, or nuclear engineering, as well as computer science.

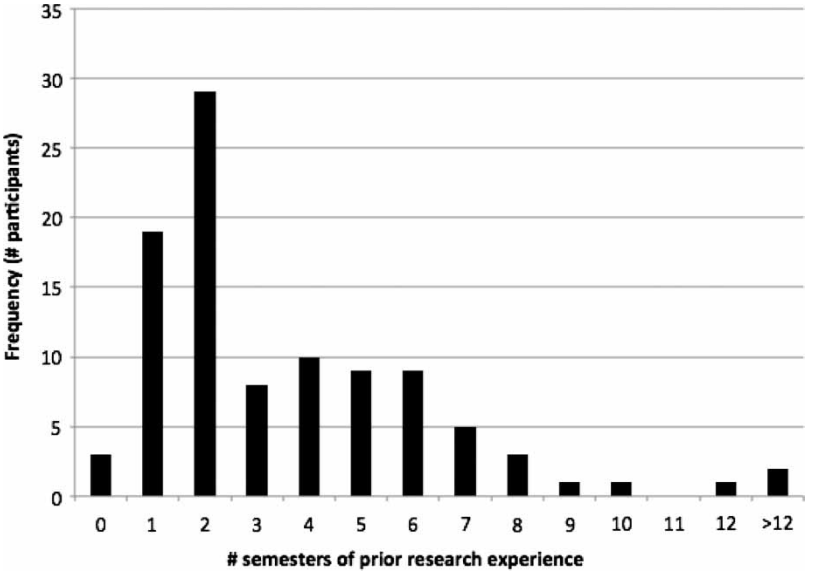

Figure 1. Duration of previous research experience reported by STEM graduate students at onset of participation in the study (n = 100 participants).

five months of experience). Given the expectation that students with more experience would have acquired more skills (Ericsson and Charness 1994; Zimmerman 2000), and given the skewed distribution of experience in our sample (Figure 1), participants were split at the median into two groups of roughly equal size: students with two semesters or less of prior experience (n = 51) and students with greater than two semesters (n = 49).

Performance data

Similar to the methodology used by Hacket and Rhoten (2009), participants wrote research proposals on a topic of their choosing, and were encouraged to conceptualize the proposal as a draft of their dissertation proposal, comprehensive examination

proposal or some other authentic activity, such as a potential submission to a funding agency. Personal relevance encourages greater cognitive investment (Schroeder et al. 2007), and, therefore, provides a more meaningful assessment. Informal communications with participants indicated many did use their proposals in this manner, reflecting the validity of this assumption. Participants were encouraged to use resources they would normally use in the formulation of their revisions to capture the processes and changes resulting from a year in a graduate program, but only individually written text was accepted for the purposes of this study (i.e. no co-authoring).

Prior to writing their proposals, participants were provided with a list of evaluation criteria, including definitions and writing prompts (see Table 2). Criteria include the following research-related skills: setting the proposed research in context, framing testable hypotheses, appropriately integrating primary literature, addressing validity and reliability of data, experimental design, selecting, presenting, and analyzing data, basing conclusions on data, and identifying alternative explanations and limitations of the proposed study. Successful scholarship also requires skills not captured in a written research proposal (e.g. collaborative skills), but these are beyond the scope of the current study. Researchers assessed the research proposals using a modified form of a previously validated rubric (Feldon et al. 2010; Timmerman et al. 2011). Wording revisions were made to the rubric to allow for discussion of predicted rather than actual results. These wording modifications occurred through discussion and consensus within the rating team.

Participants' proposals were assigned to raters based on alignment of rater expertise to subject matter. Raters were blind to participants' identities and demographics, including duration of prior research experience, type of graduate program, institution, mentor, etc. Raters possessed graduate-level backgrounds in appropriate STEM disciplines, as well as either at least two years of experience with coding written work using our rubric (Timmerman et al. 2011) or a related rubric (Caicedo et al. forthcoming), or they were tenured social science faculty who specialized in the coding of qualitative data. For each criterion, a proposal was rated as to whether evidence for that criterion was: absent (0), at a novice (1), intermediate (2) or proficient (3) level of performance. Before rating submitted proposals, the rating team first engaged in four rounds of calibration, wherein each rater individually scored proposals, and all discrepant scores were discussed among the entire rating team until consensus was achieved for both the score for the specific paper and the finer points of interpretation for each criterion. Thereafter, at least two raters scored each paper. Discrepant scores were resolved by discussion until consensus was achieved (Johnson, Penny, and Gordon 2000; Johnson et al. 2005).

Data analyses

Data analysis occurred in several stages. First, we assessed cross-sectional differences in performance on the research proposals as a function of the quantity of prior research experience using one-way analyses of variance (ANOVAs). Second, using paired t-tests, we assessed trends in individual growth from the first proposal submission to the second within the two experience groups. Criteria that changed significantly from one time point to the next within one experience group were compared with those in the other to detect differential growth patterns. To ascertain if performance on individual rubric criteria evidenced stable relationships with performance on other criteria, we constructed 4 x 4 joint frequency tables of rated scores for each possible pair of rubric

Table 2. Criteria, writing prompts and definitions provided to participants and used for rating research proposals (modified from Timmerman et al. forthcoming).

Criterion Name	Writing prompt	Definition
Context	Why should other people in your field care about your project?	Demonstrates a clear understanding of why this research question is important in this field. Background information is accurate, relevant and provides a clear rationale for the objectives.
Hypothesis(es)	What question(s) will you try to answer?	Research objectives and expected findings are clearly stated, plausible and testable. As appropriate to the field, specific hypotheses should be stated where possible. Plausible alternative explanations should be explained and the proposed research design will allow investigators to distinguish among them.
Primary Literature	What do we already know about your research questions? What are the gaps in our knowledge and how will your project help to fill them?	Relevant literature is reasonably complete and present in both the introduction and discussion sections. Use of the literature demonstrates the intellectual merit of the proposed research and specifies how it relates to other work in the field. Citations follow an accepted format for the field and are accurate (please indicate your citation style).
Validity/ Reliability	How will you know that your data are trustworthy and meaningful?	Appropriate controls and/or mechanisms to ensure validity and reliability are present and explained. Degree of replication and sample size are explained and appropriate for research area.
Experimental Design	How will you collect your data?	Data collection plan and experimental design are likely to produce salient and fruitful results (i.e. addresses the research objectives posed).
Data Selection	What data will answer your question? What do you expect the data to look like or what are the initial data?	Data produced by the research will be comprehensive, informative and relevant to the questions or hypotheses posed.

(*Continued*.)

Table 2. (*Continued*.)

Criterion Name	Writing prompt	Definition
Data Presentation	Are your figures/ diagrams/ tables clear and intelligible?	Expected or preliminary data are summarized and presented in a logical format. A rough picture of the anticipated results should be provided that includes, as appropriate, tables or graphs to show anticipated trends. Informative captions are present. If graphs are used, axes are appropriately labeled and scaled. Quantitative data should be presented using appropriate units.
Data Analysis	How will you make sense of your data? How will you know if the results are significant?	Proposed interpretive framework and/or statistical methods are appropriate for research objectives. Rationale for the choice of method is explained clearly. Expected evidence for data validity, reliability, and/or statistical significance (as appropriate to the proposed study) are indicated.
Conclusions based on data selected	If you have preliminary data:	Conclusion could be clearly and logically drawn from data. A logical chain of reasoning from research question or hypothesis to predicted data to conclusions is clearly and persuasively explained.
	If you are predicting your results:	If you get the data you predicted, what will they mean? What will you be able to say? What answers will you have to your research questions? If you get results different from what you predicted, what would that mean?
Alternative Explanations/ Limitations	How confident will you be about your conclusions? Are there other interpretations or factors that should be considered? What questions will still remain (are not addressed by your project?)	Limitations of findings and remaining questions to be answered in relation to the phenomenon of interest are discussed. Alternative explanations of the predicted data are considered and weighed against conclusions. How this study relates to other knowledge in the field is clearly discussed.

criteria. Cells reflected the frequencies with which scores on one criterion (i.e. 0, 1, 2, or 3) co-occurred with scores on another. Using a Pearson χ^2 test, we determined if the joint frequency distributions in each table differed significantly from chance. For those that did, we examined the adjusted residuals in each cell of the table to identify which value pairs were significantly over- or under-represented. Significant, positive adjusted residual values along the diagonal indicated that the two criteria scored at the same level for a given proposal significantly more often than would be predicted by chance (e.g. a score of 2 on one criterion disproportionately co-occurred with a score of 2 on another criterion). However, significant, positive residual values off the diagonal suggested that performance on one criterion differentially exceeded or led performance on the other (e.g. a score of 3 on one criterion disproportionately co-occurs with a score of 2 on another criterion) more frequently than predicted by chance. Non-significant adjusted residual values within individual cells reflected a frequency count consistent with a random distribution for the matrix.

Results

Beginning skills of graduate students and the effects of prior research experiences

Cross-sectional comparison of cohorts

Unsurprisingly, scores from less experienced participants in their fall (pre) proposal submissions have substantial room for improvement. The mean score for each criterion for students with two semesters or less of prior experience does not reach higher than a novice level (score of 1.0) performance. When only the students with three or more semesters of prior research experience are considered, the mean score reaches intermediate levels of performance (score of 2) for two criteria: Primary Literature and Context.

Despite the overall novice level of performance, more experienced students significantly outperform students with two semesters or less of research experience on four criteria: setting their research in Context ($F = 22.095$, $p = .000$), use of Primary Literature ($F = 20.63$, $p = .000$), posing testable Hypotheses ($F = 9.565$, $p = .003$) and Data Presentation ($F = 10.445$, $p = .002$) (Figure 2).

Longitudinal comparison

Using paired t-tests of pre-post measures, we evaluate gains in students' abilities over an academic year. Results indicate that less experienced students gain significantly on four criteria from fall to spring: setting one's research in Context ($p = .009$), generating testable Hypotheses ($p = .026$), Data Analysis ($p = .003$) and Primary Literature ($p = .002$) (Figure 3). Of these, all but the Data Analysis criterion are areas of significant difference between more and less experienced participants on their fall proposals.

In contrast to their less experienced counterparts, more experienced students do not show further gains in setting work in Context, generating testable Hypotheses or use of Primary Literature. Although there is no significant increase in performance on Primary Literature or setting work in Context, these two criteria remain the two highest mean scores for this group. Similar to their less experienced counterparts, the Data Analysis scores of the more experienced students increase significantly ($p = .020$). Additionally, more experienced students gain significantly in basing Conclusions on data ($p = .047$) (Figure 4).

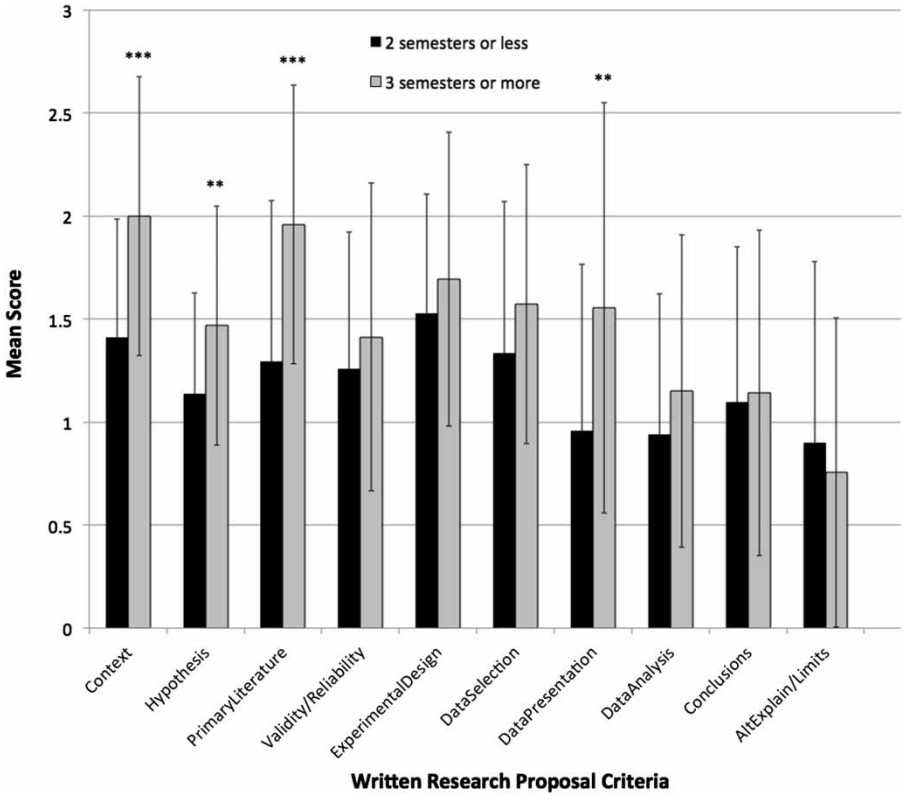

Figure 2. Participants' performance in the fall as a function of their prior research experience. Participants were categorized as having either two semesters or less of research experience (n = 51) or three semesters or more of research experience (n = 49). A score of zero indicates that aspect was missing from the proposal, a 1 represents novice level performance, 2 is intermediate, and 3 denotes competency. Bars represent mean scores $\pm$ SD. $^{***}p < .001$, $^{**}p < .01$.

Research skill development: isolated or linked?

To evaluate the extent to which performance on individual criteria are linked, χ^2 analyses against an assumption of equal cell counts reflect significant co-occurrence between scores among most criteria (see Table 3). The joint frequency distributions for some criteria are significantly different from chance with all other criteria (e.g. Primary Literature and Validity/Reliability). Others have significant χ^2 values for fewer than half of the other criteria (e.g. Limitations), and the remainder fall somewhere in between.

For those pairs of criteria with significant χ^2 values ($p < .05$), the frequency of significant positive adjusted residuals within each cell are displayed in Table 4. When scores on two criteria are the same more often than predicted by chance, the adjusted residuals are positive and significant on the diagonal (italicized values, Table 4). However, when the score on one criterion is significantly more likely to co-occur with a lower score on another, then significant positive adjusted residuals appear *below* the diagonal. For example, if students score a '3' for setting one's work in Context, but only a '2' in other criteria more often then would be expected by chance, then a positive frequency count occurs below the diagonal for the

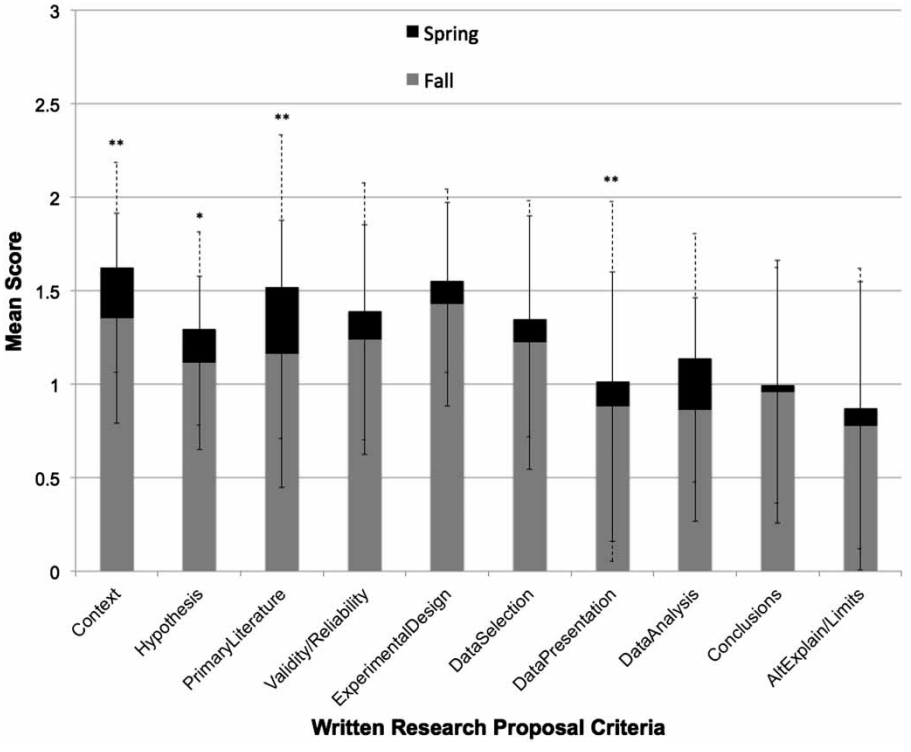

Figure 3. Longitudinal gains for students who had two semesters or less of research experience when the fall pre-measures were made (n = 51). Bars are standard deviation around the mean of each time point. Fall error bars are solid; spring error bars are dashed. $^{**}p < .01$ level, $^{*}p < .05$ level.

Context criterion in Table 4. In other words, students' scores in this area of setting work in Context are consistently higher then their scores in three other areas. Conversely, if the performance level in one criterion is lower than the quality of performance represented by another, positive adjusted residuals are found *above* the diagonal (e.g. see the Data Analysis criterion in Table 4 where scores of zero for Data Analysis occur frequently with scores of '1' for one other criterion, or Data Presentation where scores of zero occur frequently with scores of '1' for two other criteria).

Performance in Experimental Design, Primary Literature and setting work in: Context (to a lesser extent) commonly exceeds scores for other criteria. Specifically, Experimental Design has a total of ten positive significant adjusted residuals below the diagonal, Primary Literature has seven and Context has four (Table 4). In contrast, Data Analysis has six significant positive adjusted residuals *above* the diagonal, and Conclusions based on data and Limitations both have four significant positive adjusted residuals *above* the diagonal as well, indicating that scores in these areas often lag scores for other criteria.

Discussion

Our findings indicate that research skills develop neither synchronously nor independently. The amount of participants' prior research experience predicts differences in

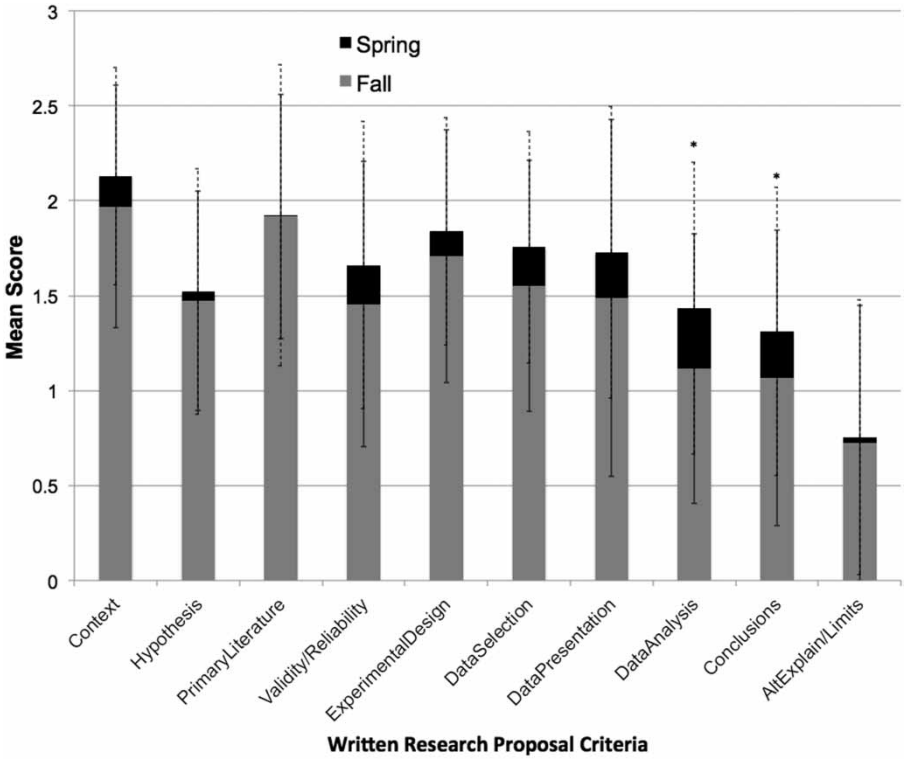

Figure 4. Longitudinal gains for students who had more than two semesters of research experience when the fall pre-measures were made (n = 49). Bars are standard deviation around the mean of each time point. Fall error bars are solid; spring error bars are dashed. $^{*}p < .05$.

some, but not all, of the skills assessed, and the longitudinal gains within experience groups reflect different profiles of growth. Specifically, significant differences between experience levels exist in the fall for setting work in Context, use of Primary Literature, and generating testable Hypotheses. Assessment of growth in these areas over an academic year reflects significant improvement in each of these areas for less experienced students and no significant improvement for more experienced students. Further, the scores of less experienced students at the end of the year (i.e. after two semesters) are similar to those of the more experienced students at the beginning of the year. These data are consistent with the existence of a developmental trajectory for research skills, wherein STEM graduate students first learn to read and apply primary literature to situate their work and appropriately frame research problems (i.e. establish testable hypotheses) within a disciplinary field. With more experience, participants develop the skills to analyze their data and draw valid conclusions from those data. The ability to consider alternative explanations and limitations of one's work, the ability to select appropriate data types to answer one's question, and the ability to present one's data appropriately showed little, if any, improvement during the course of this study, suggesting that these skills are acquired later in students' graduate careers than our early-career sample.

Examination of the data within individual proposals provides additional support for the conclusion that certain skills develop prior to others. The significant associations

Table 3. χ^2 p values for each pairwise combination of criteria.

	Context	Hypoth.	Validity Reliab.	Exp Design	Data Select	Data Present	Data Analysis	Concl.	Limits
Context									
Hypothesis(es)	***								
Validity/ Reliability	***	***							
Experimental Design	***	***	***						
Data Selection	**	***	***	***					
Data Presentation	**	.07	**	.08	***				
Data Analysis	***	***	***	***	***	***			
Conclusions based on data	**	.35	***	***	***	***	***		
Limitations	.43	.64	**	.06	.09	.54	.38	***	
Primary Literature	***	***	***	***	**	***	***	**	**

Note: *** $p < .01$; ** $.01 < p < .04$; * $p < .05$. df = 9 for all analyses.

Table 4. Cumulative counts of significant adjusted residuals produced by one criterion when summed over all other criteria.

Criterion	Performance level	All other criteria performance levels			
		0	1	2	3
Context	0	*3*	0	0	0
	1	1	*9*	−8	−2
	2	−1	−4	*3*	0
	3	−1	−4	3	*8*
Hypothesis(es)	0	*5*	0	0	0
	1	1	*5*	−2	−6
	2	−3	−4	*4*	3
	3	0	0	1	*3*
Experimental Design	0	*2*	1	0	0
	1	5	*6*	−9	−4
	2	−3	−5	*7*	0
	3	0	−4	5	*6*
Validity / Reliability	0	*6*	2	−4	0
	1	0	*8*	−4	−5
	2	0	−6	*8*	0
	3	0	−4	2	*7*
Data Selection	0	*7*	(1)−2	−3	0
	1	2	*5*	−6	−3
	2	−3	0	*5*	2
	3				*3*
Data Presentation	0	*4*	2	−3	−1
	1	0	*2*	0	0
	2	−4	−3	*4*	0
	3	0	−3	1	*6*
Data Analysis	0	*7*	1	−4	0
	1	−1	*3*	−1	−5
	2	−3	−7	*6*	5
	3	0	−1	0	*8*
Conclusions based on data	0	*6*	0	−1	−1
	1	−1	*3*	−1	−3
	2	−5	(1)−2	*4*	4
	3	−1	0	1	*3*
Limitations	0	*3*	0	−1	−1
	1	0	*2*	(1)−1	0
	2	−1	−1	*2*	3
	3	0	0	0	0
Primary Literature	0	*3*	2	−4	0
	1	2	*5*	−6	−2
	2	0	−2	*5*	0
	3	−4	−6	5	*6*

Table 5. Rubric criteria that align with Threshold Concepts identified in Kiley and Wisker (2009).

Threshold Concept	Authors' Criteria
Conceptual framework: '…through engagement with the literature… students begin to contribute to the debates in the field' (436)	Primary Literature. Literature review is relevant with sufficient depth for both introduction and discussion sections and follows citation conventions for the student's area of graduate study.
'Student understands where their work fits into the context of the entire field.' (437)	Introduction: Context. Writer provides a clear sense of why this knowledge may be of interest to other researchers in that field (describes the current gaps in understanding).
Argument/ significance: 'developing and sustaining a coherent story or line of argument that student could support with evidence from the literature or their own findings' (435)	Primary Literature. (again) Hypotheses Testable. Hypotheses are testable and distinguish among possible explanations for a phenomenon. Experimental Design. Data collection plan, experimental design, or solution approach is likely to produce salient and fruitful results (i.e. addresses the research objectives posed). Limitations. Limitations of the study and alternative explanations / potential future directions are considered and discussed.
'student can explain the significance or relevance of the project' (435)	Conclusions Based on Data. Conclusions are justified by data; connections between hypothesis, data and conclusions are clear and persuasive and conflicting or ambiguous data are discussed.
Students present and address (rather than ignore) conflicting data or evidence which would refute their work	

between specific performance levels across pairs of criteria (significant positive residuals below the diagonal in Table 4) suggest that some skills may be prerequisites for others. For example, scores of '3' in Experimental Design co-occur significantly more often than chance with scores of '2' in five other criteria, and students score '1' in Experimental Design while still scoring zero on five other criteria. This asynchronous development of skills is supportive of the notion of threshold concepts as first elucidated by Land and Meyer (Land, Meyer, and Smith 2008; Meyer and Land 2006), and later refined for graduate students by Kiley and Wisker (Kiley 2009; Kiley and Wisker 2009).

Alignment of skills assessed and threshold concepts

The assessment criteria used in our study predate the publication of Kiley's (2009) and Kiley and Wisker's (2009) findings regarding threshold concepts. However, their work provides the opportunity to compare proposed threshold concepts from the graduate education literature to our performance criteria to determine the extent to which our findings converge.

As shown in Table 5, our criteria for use of Primary Literature and ability to set one's work in Context align with the description of Kiley's threshold concept of

'conceptual framework.' Our data are consistent with a conceptual framework/primary literature/context threshold; more experienced participants earned higher scores than their less experienced counterparts in Primary Literature and Context, while only the less experienced students showed significant growth in those areas during their participation in the study. Additionally, both criteria had significant positive adjusted residuals *below* the diagonal, indicating that scores on these criteria are commonly higher than those on other criteria. While our data are correlative, not causal, higher scores in these areas preceding development of scores in other areas are supportive of the existence of a threshold concept centered on primary literature and contextual frameworks.

Another of Kiley and Wisker's (2009) threshold concepts, 'argument/significance,' encompasses five of our criteria: Primary Literature (again), Hypotheses, Experimental Design, Limitations, and Conclusions based on data. In our data, both Primary Literature and Experimental Design have substantial counts of significant positive residuals below the diagonal, reflecting a trend of stronger performance on those criteria prior to equivalent performance on other criteria. Additionally, only less experienced participants demonstrate significant growth on the Primary Literature and testable Hypotheses criteria, suggesting an earlier positioning within a developmental trajectory. In contrast, only more experienced participants demonstrated a significant gain in the quality of conclusions based on data, and limitations did not improve significantly for either set of participants.

We interpret the improvement and below-diagonal positive residuals of only some of the criteria aligned with the argument/significance threshold to indicate that there may be more refined, differentiable threshold concepts within Kiley and Wisker's (2009) argument/significance category. In STEM disciplines, the thesis of an argument is most directly represented by a hypothesis as defined in our criterion. Not all STEM disciplines utilize formal hypothesis testing as conceptualized by Popper (1959), so the operational definition of this term as used in our performance criterion is a broader conception that also encompasses design objectives as used in engineering and mathematical conjectures. Positioning the objective of a research study within disciplinary context using a disciplinarily appropriate form represents the crux of the argument component. However, the significance component aligns most directly with our Limitations and Conclusions from data criteria. Differentiating between these two facets of the argument/significance threshold permits a cleaner alignment with developmental trajectory identified through our performance data.

Limitations

Given that our sample of graduate students was in the early stages of their education, we are neither able to empirically assess all of Kiley and Wisker's (2009) proposed threshold concepts, nor to evaluate the full scope of graduate students' developmental trajectory. Subsequent work with students at later stages of development may identify operational indicators for other threshold concepts and patterns of development. Further, it is not known if the use of anticipated data by students influenced their performance on some criteria where obtained results are normally used.

Another important limitation is the correlative nature of the data used for the analysis of residuals in the joint frequency matrices. Because performance on pairs of criteria inherently occurs simultaneously, it is only possible to assess relative strength as a function of score rather than a temporal ordering of skill development.

We cannot determine the extent to which two skill areas are causally linked and to what extent development is concurrent, but some are clearly developing while others are still nascent. The developmental trajectory we suggest is compatible with the data, but it cannot be tested directly without performance-based assessments throughout the full scope of students' graduate studies.

Implications

The significance of our findings lie both in the insight they provide for the development of research skills in general and their potential to provide evidence-based recommendations for the practice of graduate education (Feldon, Maher, and Timmerman 2010). Identification of normative differences in initial levels and patterns of skill development could be used to improve graduate training and possibly accelerate skill development. A fully identified developmental trajectory could optimize the development and implementation of training experiences to maximize students' skill development. Further, replication through subsequent research can establish benchmarks for performance that could be informative for detecting students who are experiencing atypical difficulties and allow targeted remediation.

Based on the current findings, students may benefit from an increased emphasis on the role of primary literature in scientific practice, and receive ample opportunities for students to evaluate the quality of research using primary literature. Once students gain baseline competency in the primary literature relevant to their field, their readiness to engage in other facets of research appears to follow.

Acknowledgements

This work was supported by a grant from the US National Science Foundation to David Feldon, Michelle Maher, Briana Timmerman, Jed Lyons, and Stephen Thompson (NSF-0723686). The views expressed do not necessarily represent the views of the supporting funding agency. The authors would also like to express their appreciation to the graduate programs, faculty and students whose cooperation made this work possible.

References

Caicedo, J.M., C.E. Pierce, J. Flora, B. Timmerman, A.P. Nichols, W. Graf, and T. Ray. Forthcoming. Instructional environment to stimulate critical thought of freshmen civil engineering students. *Advances in Engineering Education*.

Carnegie Initiative on the Doctorate. 2001. *Overview of doctoral educational studies and reports: 1990–present*. Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Davis, G., and P. Fiske. 2001. *The 2000 national doctoral program survey*. University of Missouri-Columbia: National Association of Graduate-Professional Students.

Dehaan, R.L. 2005. The impending revolution in undergraduate science education. *Journal of Science Education and Technology* 14, no. 2: 253–70.

Dunbar, K. 2000. How scientists think in the real world: Implications for science education. *Journal of Applied Developmental Psychology* 21, no. 1: 49–58.

Ericsson, K.A., and N. Charness. 1994. Expert performance: Its structure and acquisition. *American Psychologist* 49, no. 8: 725–47.

Feldon, D. 2007. The implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review* 19, no. 2: 91–110.

Feldon, D. 2010. Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy along a continuum of expertise. *Instructional Science* 38, no. 4: 395–415.

Feldon, D.F., M.M. Maher, and B.E. Timmerman. 2010. Performance-based data in the study of STEM PhD education. *Science* 329, no. 5989: 282–3.

Feldon, D.F., B.E. Timmerman, R. Showman, and K.A. Stowe. 2010. Translating expertise into effective instruction: The impacts of cognitive task analysis (CTA) on laboratory report quality and student retention in the biological sciences. *Journal of Research in Science Teaching* 47, no. 10: 1165–85.

Golde, C.M. 2001. *Findings of the survey of doctoral education and career preparation: A report to the preparing future faculty program*. Madison: University of Wisconsin.

Hackett, E., and D. Rhoten. 2009. The snowbird charrette: Integrative interdisciplinary collaboration in environmental research design. *Minerva* 47, no. 4: 407–40.

Halonen, J.S., T. Bosack, S. Clay, M. McCarthy, D.S. Dunn, G.W. Hill, R. McEntarffer, C. Mehrotra, R. Nesmith, K.A. Weaver, and K. Whitlock. 2003. A rubric for learning, teaching, and assessing scientific inquiry in psychology. *Teaching of Psychology* 30, no. 3: 196–208.

Hinds, P.J. 1999. The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied* 5, no. 2: 205–21.

Hofstein, A., and V.N. Lunetta. 2004. The laboratory in science education: Foundations for the twenty-first century. *Science Education* 88, no. 1: 28–54.

Johnson, R.L., J. Penny, and B. Gordon. 2000. The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education* 13, no. 2: 121–38.

Johnson, R.L., J. Penny, B. Gordon, S.R. Shumate, and S.P. Fisher. 2005. Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores?. *Language Assessment Quarterly: An International Journal* 2, no. 2: 117–46.

Keys, C.W. 1994. The development of scientific reasoning skills in conjunction with collaborative writing assignments: An interpretive study of six ninth-grade students. *Journal of Research in Science Teaching* 31, no. 9: 1003–22.

Kiley, M. 2008. *Quality in postgraduate education*. Canberra: Australian National University, Centre for Educational Development and Academic Methods 2008. http://www.qpr.edu.au/overview.html.

Kiley, M. 2009. Identifying threshold concepts and proposing strategies to support doctoral candidates. *Innovations in Education and Teaching International* 46, no. 3: 293–304.

Kiley, M., and G. Wisker. 2009. Threshold concepts in research education and evidence of threshold crossing. *Higher Education Research & Development* 28, no. 4: 431–41.

Land, R., J.H.F. Meyer, and J. Smith, eds. 2008. *Threshold concepts within the disciplines*. Rotterdam: Sense.

Lawson, A.E. 2008. *Biology: An inquiry approach*. 2nd ed. Dubuque, IA: Kendall/Hunt Publishing Company.

Marx, R.W., P.C. Blumenfeld, J.S. Krajcik, B. Fishman, E. Soloway, R. Geier, and R.T. Tal. 2004. Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching* 41, no. 10: 1063–80.

Mervis, J. 2000. Graduate educators struggle to grade themselves. *Science* 287, no. 5453: 568–70.

Meyer, J.H.F., and R. Land, eds. 2006. *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge*. Abingdon, UK: Routledge.

National Research Council. 1996. *National science education standards*. Washington, DC: National Academy Press.

Onwuegbuzie, A.J. 2003. Modeling statistics achievement among graduate students. *Educational and Psychological Measurement* 63, no. 6: 1020–38.

Ostriker, J.P., and C.V. Kuh, eds. 2003. *Assessing research-doctorate programs: A methodology study*. Washington, DC: National Academies Press.

Popper, K. 1959. *The logic of scientific discovery*. New York: Routledge.

Roberts, G. 2002. *Set for success: The supply of people with science, technology, engineering and mathematical skills*. HM Treasury, April 2002. http://www.employment-studies.co.uk/pubs/report.php?id=1440robert.

Sandoval, W.A., and B.J. Reiser. 2004. Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education* 88, no. 3: 345–72.

Schroeder, C.M., T.P. Scott, H. Tolson, T.-Y. Huang, and Y.-H. Lee. 2007. A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching* 44, no. 10: 1436–60.

Schunn, C.D., and J.R. Anderson. 1999. The generality/specificity of expertise in scientific reasoning. *Cogitive Science* 23, no. 3: 337–70.

Sevian, H., and L. Gonsalves. 2008. Analysing how scientists explain their research: A rubric for measuring the effectiveness of scientific explanations. *International Journal of Science Education* 30, no. 11: 1441–67.

Seymour, E. 2001. Tracking the processes of change in US undergraduate education in science, mathematics, engineering and technology. *Science Education* 86, no. 1: 79–105.

Timmerman, B.E., D.C. Strickland, R.L. Johnson, and J. Payne. 2011. Development of a 'universal' rubric for assessing undergraduates' science inquiry and reasoning skills using scientific writing across multiple courses. *Assessment & Evaluation in Higher Education* 36, no. 5: 509–47.

Willison, J., and K. O'Regan. 2007. Commonly known, commonly not known, totally unknown: A framework for students becoming researchers. *Higher Education Research & Development* 26, no. 4: 393–409.

Zimmerman, C. 2000. The development of scientific reasoning skills. *Developmental Review* 20, no. 1: 99–149.