Instructional Science
https://doi.org/10.1007/s11251-022-09578-5

ORIGINAL RESEARCH

Detecting threshold concepts through Bayesian knowledge tracing: examining research skill development in biological sciences at the doctoral level

Jina Kang[1] · Ryan Baker[2] · Zhang Feng[3] · Chungsoo Na[3] · Peter Granville[4] · David F. Feldon[3]

Received: 12 June 2020 / Accepted: 17 January 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract
Threshold concepts are transformative elements of domain knowledge that enable those who attain them to engage domain tasks in a more sophisticated way. Existing research tends to focus on the identification of threshold concepts within undergraduate curricula as challenging concepts that prevent attainment of subsequent content until mastered. Recently, threshold concepts have likewise become a research focus at the level of doctoral studies. However, such research faces several limitations. First, the generalizability of findings in past research has been limited due to the relatively small numbers of participants in available studies. Second, it is not clear which specific skills are contingent upon mastery of identified threshold concepts, making it difficult to identify appropriate times for possible intervention. Third, threshold concepts observed across disciplines may or may not mask important nuances that apply within specific disciplinary contexts. The current study therefore employs a novel Bayesian knowledge tracing (BKT) approach to identify possible threshold concepts using a large data set from the biological sciences. Using rubric-scored samples of doctoral students' sole-authored scholarly writing, we apply BKT as a strategy to identify potential threshold concepts by examining the ability of performance scores for specific research skills to predict score gains on other research skills. Findings demonstrate the effectiveness of this strategy, as well as convergence between results of the current study and more conventional, qualitative results identifying threshold concepts at the doctoral level.

Keywords Threshold concepts · Bayesian knowledge tracing · Research training · Doctoral education

✉ Jina Kang
jinakang@illinois.edu

[1] Department of Curriculum and Instruction, University of Illinois Urbana-Champaign, IL, Champaign, USA

[2] Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA

[3] Instructional Technology and Learning Sciences, Utah State University, Logan, UT, USA

[4] Penn Center for Learning Analytics, Philadelphia, PA, USA

Published online: 14 March 2022

Introduction

Transformation from disciplinary novices to independent researchers is a primary objective of doctoral education (Kiley & Wisker, 2009; Lovitts, 2008). A common experience for doctoral students during their academic journeys is the feeling of being "stuck" at certain points in their developmental trajectories as they acquire new research concepts and skills (Keefer, 2015; Land et al., 2005). These points represent both obstacles and significant opportunities for developing proficiency in conducting research (Feldon et al., 2017). These points, known as threshold concepts (Meyer et al., 2010), represent key knowledge elements that are challenging to attain and, once mastered, transform doctoral students' subsequent perspectives, enabling the further development of expertise. Thus, threshold concepts "can be akin to a portal, opening up a new and previously inaccessible way of thinking about something" (Meyer & Land, 2006, p. 3).

Previous research to identify threshold concepts in doctoral education has often focused on those that typically emerge across disciplines and from the perspective of faculty. For example, Kiley and Wisker (Kiley, 2009; Kiley & Wisker, 2009) found that faculty supervisors perceived fundamental transition points for doctoral students around the concepts of developing and supporting a scholarly argument, situating work within a theory, and applying a conceptual framework to structure the development of a study. Subsequent work, based on interviews with students, supported and elaborated on the *theory* and *framework* threshold concepts as especially challenging, but transformative to research knowledge, once attained (Kiley, 2015).

Other research has focused on threshold concepts within specific disciplines. For example, through interviews and focus groups with advanced and recently completed doctoral students in the biological sciences, Feldon and colleagues (2017) identified two basic threshold concepts: (1) the ability to effectively engage primary literature in a critical and constructive manner and (2) the ability to conceptualize appropriate controls when designing and interpreting experiments. While the former is a common barrier faced by students in many disciplines (e.g., Boote & Beile, 2005; Lovitts, 2007), the specific nature of controls in the design of biology experiments is unique in some ways and an essential concept to master to become an independent scholar in the discipline (Gross & Mantel, 1967).

At a practical level, identifying threshold concepts is a valuable activity. Doing so provides a means of identifying core competencies or necessary interim benchmarks. Further, because threshold concepts are recognized as being transformative and especially challenging to master, identifying them provides guidance for faculty mentors who may need to offer students additional instruction and support. However, research on threshold concepts at the doctoral level faces several limitations. First, the generalizability of past findings has been limited due to the relatively small numbers of participants in available studies. Second, it is not clear which specific skills are prerequisite to or contingent upon mastery of identified threshold concepts, making it difficult to identify appropriate times for possible intervention. Third, the observed commonalities across disciplines may or may not mask important nuances that apply within specific disciplinary contexts (e.g., Feldon et al., 2017).

For these reasons, the current study employs a novel approach to identifying possible threshold concepts on the basis of a large data set specific to the discipline of laboratory-based biological sciences (e.g., cellular and molecular biology, microbiology, developmental biology). Using rubric-scored samples of doctoral students'

sole-authored scholarly writing, we apply Bayesian Knowledge Tracing as a strategy to identify potential threshold concepts by examining the ability of performance scores for target research skills to predict score gains on other research skills.

Review of literature

The study of research skill development in the context of graduate education has historically relied on two discrete traditions: situated studies of academic development focused on graduate students and other early career researchers (e.g., Delamont & Atkinson, 2001; Lovitts, 2005; McAlpine & McKinnon, 2013), and cognitively oriented studies conceptualizing research proficiency as the acquisition of complex skills (e.g., Feldon et al., 2019; Schraagen, 1993; Schunn & Anderson, 1999). The purpose of the current study, to identify threshold concepts through the application of skill modeling strategies, represents an integration of these two traditions. Research skills and potential threshold concepts are situated within the authentic context of participants' scholarly writing generated as part of their ongoing pursuit of a Ph.D. in the biological sciences. At the same time, the skills examined are narrowly defined and scored according to performance criteria. Accordingly, in this section, we review central tenets and findings from both traditions in the following paragraphs and characterize the ways in which they can be reflected in a Bayesian Knowledge Tracing format.

Threshold concepts

Threshold concepts are typically described as essential knowledge for a domain that, once mastered, irreversibly transform an individual's perspective and enable performance and understanding that were not previously possible. They are often likewise characterized as being transformative, irreversible, integrated, bounded, and troublesome (Meyer & Land, 2003), though a comprehensive set of definitional criteria is still a topic of debate (Salwën, 2019). Threshold concepts reflect barriers to the attainment of expertise until they are mastered, because their attainment catalyzes a reorganization of knowledge that permits more sophisticated engagement with a broader array of tasks and concepts than is typical of an encapsulated concept (Roberts, 2016). Kiley and Wisker (2009) suggest that the mastery of threshold concepts, which are different from discipline-specific "core concepts" due to their transformative impact on subsequent understanding, will equip a learner with a qualitatively different view of oneself and the domain to which it is relevant. Further, until a threshold concept is attained, learners experience a liminal lack of progress during which broader progress in the domain stalls (Keefer, 2015). Accordingly, the process of developing into an independent scholar with a specific field of expertise is often experienced as a process of advancing from one liminal space to the next (Kiley, 2015; Leshem, 2020).

Based on interviews and surveys with experienced Ph.D. supervisors, Kiley (Kiley 2009; Kiley & Wisker, 2009) identified six generic threshold concepts addressing developmental trajectories of research concepts and skills: *argument*, the ability to mount a defensible argument; *theorizing*, the ability to generate theoretical models that make sense of findings and results; *framework*, the ability to explain and articulate based on a theoretical framework and methodological position; *knowledge creation and originality*, the ability to make original contributions to the field of academia; *analysis*, the ability to conduct rigorous analysis; *paradigms*, the ability to appreciate and understand the existing paradigm

and appropriate methodological approach. Consistent with these broad perspectives, which emphasize central facets of building a comprehensive scholarly argument, Chatterjee-Padmanabhan et al. (2018) argue that critically engaging in primary literature and developing a scholarly voice remained a major threshold concept for graduate students. Similarly, Feldon and colleagues (2017) identified through interviews and focus groups with graduate students and early career researchers that the ability to take a balanced perspective when evaluating scholarly literature was a notable threshold concept, preceded first by acceptance of published claims at face value and then by overly critical analysis that was counterproductive to understanding the state of developing knowledge in a given field. Wisker (2015) also suggested that the ability to write a good literature review can be regarded as the crossing of conceptual threshold, which demonstrates their interpretation of theoretical perspectives in the domain.

These prospective threshold concepts tend to be consistent across disciplines. However, the nature of learning during doctoral study—and in the development of expertise generally—is domain-specific, with epistemic framing, argumentation, understanding of relevant theory, and relevant problem-solving strategies unique to individual disciplines (Ericsson et al., 1993; Knorr-Cetina, 1999; Kuhn, 1962, 1977; Thagard, 2003). Accordingly, Feldon and colleagues (2017) also identified a threshold concept specific to laboratory biology, the design of experimental controls. The specific understanding and approach to design of these controls are unique to the building of disciplinary arguments in biology, so they reflect a fundamentally different scope than more generic thresholds described above.

Research skill development

In contrast to the identification of threshold concepts which are commonly identified on the basis of individuals' reports of transformative knowledge that is difficult to attain, the cognitive tradition of research in scientific problem solving identifies key research skills through the observation or measurement of their application to specific problems. For example, the use of think-aloud protocols has identified differences between expert and novice problem-solving strategies related to reasoning through analogy (Nersessian & Chandrasekharan, 2009) and use of mental models to visualize likely outcomes (Christensen & Schunn, 2009). Likewise, performance in simulated experimental design and analysis problems has identified specific skills such as selection of the number of variables manipulated or measured that can reflect varied levels of expertise (Feldon, 2010; Hmelo-Silver et al., 2002; Schraagen, 1993; Schunn & Anderson, 1999; Tschirgi, 1980). In more naturalistic contexts, studies have analyzed written artifacts such as research proposals or reports of empirical findings to assess the level of skill manifested in different aspects of the arguments presented (e.g., Feldon et al., 2011, 2019; Hackett & Rhoten, 2009; Timmerman, et al., 2011).

Studies examining sole-authored, written scholarly artifacts offer several advantages for evaluating manifestations of relevant skills. First, they are generated as an authentic task central to the work of scholars, which avoids concerns about the potential biasing of performance during simulated tasks (Feldon et al., 2010; Seashore Louis et al., 2007). Second, they do not rely on supervisors' reports or self-reports of scholarly skill, as each source rarely aligns with the other or with independently scored performance-based assessments (Cox & Andriot, 2009; Feldon et al., 2015; Goldstein et al., 2014; Kardash, 2000). Third, the structure and use of validated performance rubrics have proven effective in identifying both individual skills reflected in assessed writing samples and trajectories of skill

development over time (Feldon et al., 2019; Timmerman et al., 2013), including potential threshold concepts (Urquhart et al., 2016).

In their application of this strategy, Feldon and colleagues (2019) assessed 12 specific research skills over a period of four years of doctoral study in the biological sciences. These included: setting the study in context, integration of primary literature, establishing testable hypotheses, using appropriate controls and replication strategies, experimental design, selection of data for analysis, data analysis, presentation of results, basing conclusions on results, identifying potential alternative explanations for findings, identifying the limitations of the study, and generating implications from the conclusions generated. Longitudinal analysis of these skills did yield insights into the patterns of skill development evident over time in the form of latent profiles reflecting patterns of relative strengths amongst skills and latent growth models indicating common trajectories of skill growth. However, those statistical analyses were unable to determine the extent to which any specific skills acted as threshold concepts, limiting growth in other skills until mastered.

Understanding the relationship between prerequisite skills and threshold concepts is an underdeveloped aspect of the literature, largely due to the different methodological and theoretical traditions applied to their study. For the current research, we posit that a threshold concept would have multiple prerequisite skills that could also hold prerequisite relationships with one another. If a threshold concept had few or no prerequisite skills, it would be likely to demonstrate rapid early improvement, consistent with models of single skill development (Ackerman & Beier, 2018; Murre, 2014). However, attainment of threshold concepts typically requires both extended periods of time for mastery and reliance upon requisite prior knowledge (Shanahan et al., 2006). Threshold concepts differ from complex core concepts in their ability to transform learners' overall understanding or perspective (Meyer & Land, 2003), for which experiential or reflective data from learners or instructors would be required.

Synthesizing research skill through Bayesian knowledge tracing

Bayesian knowledge tracing (BKT; Corbett & Anderson, 1995) has been successfully applied to model students' knowledge in adaptive learning environments such as intelligent tutoring systems (Koedinger & Corbett, 2006) and simulation-based learning environments (Sao Pedro et al., 2013). BKT is widely used in a variety of learning systems, because it offers both a reasonably good fit to data and interpretable parameters. In particular, BKT's extensibility has enabled it to be used to answer a variety of research questions around the effectiveness of learning system features and content (Sao Pedro et al., 2014, Baker et al., 2018; Beck et al., 2008; Yudelson et al., 2008).

BKT is a distinct variation of a Hidden Markov Model—known as a two-state Hidden Markov Model. The key mechanism is to iteratively estimate the probability of skill acquisition based on the observed performance of that skill. To capture students' knowledge status, BKT is guided by four parameters in the model:

- L_0 is the initial probability that a student already learned the skill prior to practicing it.
- T is the probability that a student will learn the skill after practicing it.
- S is the probability that a student makes a mistake in regard to the skill despite possessing it.
- G is the probability that a student demonstrates the skill by chance despite actually not possessing it.

Basic BKT regards the relationship between skills as *complete transfer* (where performance on one skill predicts increases in performance on another skill equal to the extent that the latter skill predicts itself across time) or *complete skill independence* (where performance on one skill does not predict any increase in performance on another skill). Accordingly, it ignores the possibility of a partial transfer of skills (Singley & Anderson, 1989; Speelman & Kirsner, 1997) across different tasks (where performance on one skill predicts performance on another skill, but to a lesser extent than the latter skill predicts itself across time). In the current study, partial transfer manifests as (1) the dependence on attainment of a threshold concept to enable the development or improvement of a separate, new skill or (2) the dependence of threshold concept attainment for a given skill on the prior mastery of other skills. To address this, Sao Pedro and colleagues (2014) proposed Bayesian Knowledge Tracing-Partial Skill Transfer (BKT-PST), as an extension to basic BKT. This approach accounts for partial transfer between two states of mastery learning. Compared to basic BKT, BKT-PST adds the parameter *k*, which is the likelihood of maintaining skills when switching between topics or time points. It can therefore capture how competency in a latent skill might be transferable to another latent skill through the value *k*.

Despite this advantage of capturing the partial transfer of skills, established BKT-PST strategies assume that a skill is a binary variable (e.g., known and unknown), so it is constrained in detecting non-binary trajectories of transfer, which can be often found in advanced cognitive skills. To address such limitations, the current paper redefines skills as continuous variables rather than binary categorical values within the BKT framework. It does so by specifying the degree of observable competence that students actually demonstrate in a certain task (i.e., continuous numerical scores). Thus, the new model implemented in the current study is an extension of BKT-PST, which we label Bayesian Knowledge Tracing-Partial Skill Transfer as Continuous (BKT-PSTC). This model's goal is to capture the nuanced developmental process of acquisition and transfer across sets of different research skills. More details of BKT-PSTC are provided in the methods section.

Although BKT is most frequently applied within computer-based training or intelligent tutoring systems, its ability to model skill acquisition and performance can be applied in any learning context. Accordingly, we consider the BKT framework to be promising for the analysis of students' research skill development in authentic contexts, such as scholarly writing (e.g., Florence & Yore, 2004). Thus, the application of BKT-PSTC to performance data from authentic tasks in this area of research permits more detailed analyses of the relationships amongst skills as they develop than either performance on simulated research tasks (e.g., Hmelo-Silver et al., 2002) or in-depth interviews (e.g., Kiley, 2009; McAlpine & McKinnon, 2013).

Specifically, as this model attempts to capture the degree to which transfer of a latent skill is dependent on the observed performance of skills in an earlier stage, tracing students' research skills allows us to determine the extent to which the acquisition of a specific research skill (i.e., source skill) can contribute to the acquisition of another skill (i.e., destination skill). The BKT model can quantify the degree of transition from a source skill to a destination skill through a parameter value, *k*, and can be analyzed to examine the statistical significance of transfer from one research skill to other skills over time. In this way, we can examine which research skills act as precursors of the attainment of a threshold concept or demonstrate faster growth after attainment of a threshold concept. The parameter values of each source skill can provide a precise estimate of these relationships.

This work is related to—but not quite the same as—work that attempts to capture prerequisite structure among a set of skills (e.g., Chen et al., 2016; Scheines et al., 2014). Such prerequisite structures posit that a post-requisite skill cannot be learned without first

acquiring the prerequisite skill. The structure we instead posit in this paper is one where a threshold concept assists in the acquisition of another skill, but is not strictly required. Several papers have attempted to represent and leverage prerequisite connections between skills, using approaches ranging from modifications of BKT, to knowledge spaces, to neural networks (e.g., Adjei et al., 2014; Botelho et al., 2015; Chen et al., 2018; Doignon & Falmagne, 2012). In general, these studies have investigated whether prediction of student performance could be enhanced by using information on prerequisite relationships between skills, again with the operational definition that prerequisites imply the fact that a specific skill cannot be acquired without first acquiring another skill. Perhaps the closest to our approach (aside from Sao Pedro et al., 2014) is Botelho et al. (2015). Botelho and his colleagues proposed a prerequisite binning extension to BKT where students are grouped according to their mastery speed on prerequisite skills, and the prerequisite information is utilized to improve prediction of students' initial response on subsequent skills. Although this approach was referred to as involving prerequisites, it shared our paradigm where acquiring one skill is facilitated by having another skill, but that earlier skill is not seen as absolutely necessary.

In this paper, we therefore capitalize on the BKT framework as a vehicle to identify the contingent structure of research skills over time as an indicator of prospective threshold concepts, based on written scholarly artifacts. To capture the moments when transfer occurs among skills, we apply the BKT-PSTC as a new strategy to compute the probability of transfer from source skills to target skills. To the extent that many skills collectively transfer to one subsequent skill or a single skill transfers to many subsequent skills—especially when the subsequent skill(s) evidence no improvement until substantial gains are demonstrated for its predictor(s)—this framework enables the detection of prospective threshold concepts. To test this approach, we address the following research questions:

1. What are the sequential dependencies between individual research skills detected by BKT-PSTC in samples of students' scholarly writing?
2. Do detected dependencies reflect patterns consistent with threshold concepts identified through prior research?

Method

Participants

As part of a 4-year longitudinal study, we recruited 336 doctoral students in the laboratory-based biological sciences from 53 institutions across the United States. Forty-two institutions were classified as R1 (highest research activity), seven as R2 (higher research activity) institutions, and four as other Carnegie Foundation categories. All participants consisted of incoming Ph.D. students in Fall 2014, of whom a majority were female (n = 183), domestic students (n = 237), continuing-generation (n = 210), and from racial/ethnic majority groups (n = 240). Mean age was 24.9 years (SD = 3.6) at the outset of the study. Four participants did not provide their racial/ethnic information, and four participants did not provide information on their status as domestic or international students. All participants provided informed consent to participate in the research and received a participation incentive of $400 USD per year. Not all participants contributed data in every year of the study. 297 students provided at least one writing sample. However, the current study only

included the participants with data every year, which yielded a total of 84 students (see more details in 'Data source').

Data source

To measure individuals' research skills, we collected each participant's sole-authored writing sample (e.g., draft manuscripts, qualifying or comprehensive examinations, dissertation proposals) written between February and June of each year. These artifacts were written without contributions (e.g., co-authorship or editing) from others. The writing samples were all unpublished at the time they were submitted for the study.

Written scholarly artifacts, as a common research practice, represent how researchers can build scientific arguments, which is a key competency for successful research. Accordingly, capturing research skills through writing samples can allow us to determine whether individuals acquired the research skills and successfully applied them in an authentic research context. The following specific skills were assessed according to a rubric fully reported by Feldon et al. (2019):

- Data analysis (ANA; 0).
- Selecting data for analysis (SEL; 1).
- Basing conclusions on results (CON; 2).
- Identifying limitations of the study (LIM; 3).
- Identifying alternative explanations of findings (ALT; 4).
- Discussing the implications of the findings (IMP; 5).
- Establishing testable hypotheses (HYP; 6).
- Introducing/setting the study in context (INT; 7).
- Using appropriate experimental controls and replication (CTR; 8).
- Experimental design (EXP; 9).
- Presenting results (PRE; 10).
- Appropriately integrating primary literature (LIT; 11).
- Writing quality (WRT; 12).

Two blind raters scored each writing sample on these thirteen research skills originally identified from a thorough review of literature on the development of scientific arguments generally and from the biological sciences in particular (Timmerman et al., 2011). Each research skill had a scored range of 0 to 3.25, with 0 meaning 'Not Addressed' or providing a completely irrelevant statement (0 + 0.25), and 3 ± 0.25 meaning 'Proficient'. Scores were averaged across two raters and then used as a composite measure for each research skill. Interrater reliability as measured by intraclass correlations (ICCs; two-way random effects) was good ($0.818 \leq \text{ICC} \leq 0.969$; see exact ICC values in Feldon et al., 2019).

Of the students ($n = 297$) who were scored at any point from Year 1 to Year 4, 21.9% were missing data in Year 1, 22.9% were missing data in Year 2, 33.3% were missing data in Year 3, and 50.8% were missing data in Year 4. 28.3% of students ($n = 84$) were scored at all four time points. To handle missing data, we applied listwise deletion (Peugh & Enders, 2004) and removed any students that had one or more missing data from the entire dataset. This yielded a total of 84 students. According to Little's (1988) test, data were missing completely at random ($\chi^2[312] = 346.41$, $p = 0.09$), indicating that missingness should not have introduced bias into the analyses. Table 1 shows the means and standard deviations for each research skill score across time.

Table 1 Means and standard deviations for research skills

Skills	Year 1 ($n = 84$)		Year 2 ($n = 84$)		Year 3 ($n = 84$)		Year 4 ($n = 84$)	
	M	SD	*M*	SD	*M*	SD	*M*	SD
0 (ANA)	0.51	0.71	0.84	0.89	1.03	1.02	1.14	0.77
1 (SEL)	0.78	0.80	1.21	0.86	1.39	0.95	1.53	0.82
2 (CON)	1.12	0.94	1.18	0.89	1.38	0.92	1.53	0.89
3 (LIM)	0.92	0.93	1.19	1.00	0.90	0.80	1.06	0.98
4 (ALT)	0.87	0.91	1.07	0.95	0.98	0.90	1.06	0.96
5 (IMP)	1.36	0.93	1.45	0.86	1.29	0.81	1.47	0.90
6 (HYP)	1.48	0.92	1.73	0.89	1.53	0.86	1.53	0.85
7 (INT)	1.90	0.75	2.03	0.71	1.79	0.70	1.82	0.87
8 (CTR)	1.30	0.83	1.63	0.82	1.32	0.86	1.43	0.79
9 (EXP)	1.92	0.75	2.10	0.62	1.88	0.66	1.98	0.66
10 (PRE)	0.39	0.74	0.98	1.06	1.15	1.11	1.22	1.14
11 (LIT)	1.73	1.05	1.84	1.02	1.42	1.11	1.54	1.18
12 (WRT)	2.08	0.60	2.26	0.51	2.17	0.38	2.27	0.52

Data analysis

After rescaling scores from 0 to 1 for easier application within BKT, we created models to estimate the probability of students transferring one research skill to another research skill using BKT-PSTC, an extended version of BKT-PST model. Like BKT-PST, the proposed BKT-PSTC also added a *k* parameter [0.01–0.99] in the model to adjust the probability of the skill acquisition, in addition to the four parameters (i.e., L_0, G, S, T) of the basic BKT model. A *k* parameter indicated the percentage of a skill learned at one time point (i.e., source skill) that was transferred to another skill at the next available time point (i.e., destination skill).[1] That is, each BKT-PSTC model predicted a student's score on the destination skill by allowing a portion of its prediction to come from the probability of the student knowing the source skill at the previous time point. This portion was multiplied by the *k* parameter where a larger *k* indicates that the model weights the past source skill more in calculating the current destination skill.

We created 13 models for each skill. Each model investigates the transfer from a single skill to another single skill. Were we to fit all skill transfer simultaneously, it would be difficult to tease out the transfer between specific pairs of skills. In addition, given the limited amount of data available there would be significant concerns around over-fitting and identifiability, making it difficult to ascertain whether a specific skill–skill relationship improved model fit to a statistically significant degree.

For each combination of source skill and destination skill, the model finds the *k* parameter that maximizes the predictive accuracy of the BKT-PSTC model (using the sum of

[1] In our implementation of BKT-PSTC, we treat time as an ordinal value rather than continuous. For example, if a participant was missing data at the second time point, then data from the third time point was treated as if it were collected at the second time point within the model. This choice may have resulted in overestimation of the degree of learning per time point, but probably only to a small degree, given the limited amount of data missing. The alternative strategy, imputing estimated correctness and inputting it into the model, can produce extreme and unreliable estimates (cf., Author, in press).

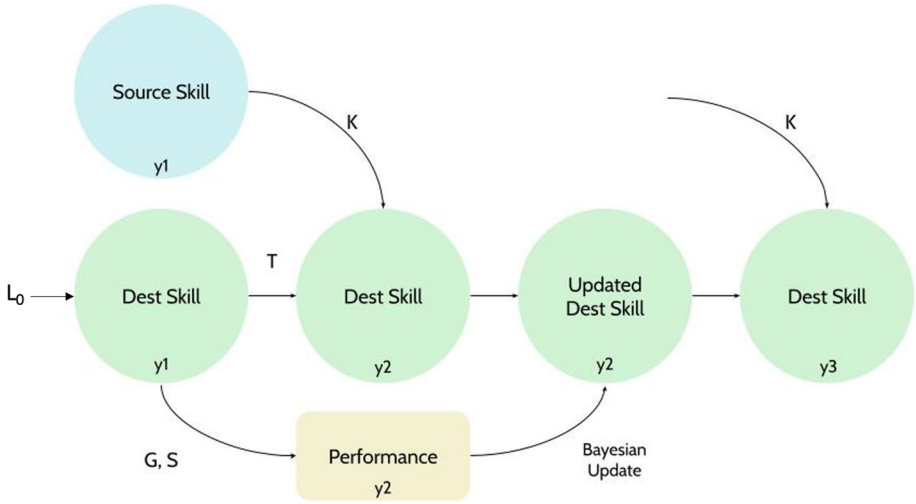

Fig. 1 Example of BKT-PSTC model

squared residuals as the goodness criterion during training). Figure 1 shows the example of our BKTC-PSTC models. When predicting the destination skill at a time point $y = n$, the probability of the student knowing the source skill at time $y = n - 1$ is used. When predicting the destination skill at $y = 1$, no source skill score is applied; this is to say that, when predicting the destination skill at $y = 1$, the model has no PSTC component (k). Each BKT-PSTC model was evaluated using fourfold cross validation at the student level. Specifically, each fold has a different set of students; that is, the same set of students are in a specific fold for each skill. The models were trained on three groups of students and tested on a fourth group of students. The values for the five parameters are the average values across the four folds. Student-level cross-validation attempts to estimate the model's goodness for unseen students (within the overall population), which is important if we want to understand how the model might generalize to students beyond our training sample. Fourfold cross-validation was selected in order to achieve a balance between having relatively large proportions of the sample included in training (compared to, say, twofold cross-validation), and overlap between training sets that is further from 100% (compared to, say, 10-fold cross-validation). It also had the benefit of tractable model training time (compared to, say, 10-fold cross-validation).

Specifically, the BKT-PSTC model works as follows:

1. Take a source skill and destination skill.
2. For each fold.
    a. Find 4 standard parameters ($P(L_0)$, G, S, T) for the source skill in fold.
    b. Apply the source skill model to the source skill, producing $P\left(L_{y\_source}\right)$ for the source skill at each step, and $P\left(L_{y\_source}\right) - P\left(L_{y-1\_source}\right)$ for one time point earlier.
    c. Find 5 parameters for the destination skill: $P(L_0)$, G, S, T, and k.
    d. Where k can take value [0.01, 0.99] with a step of 0.01.
    e. After computing formula, $P\left(L_{y\_destination}\right) = P\left(L_{y-1\_destination}\right) + \left(1 - P\left(L_{y-1\_destination}\right)\right) \times T$
    Use formula, $P\left(L_{y\_destination}\right)^* = P\left(L_{y\_destination}\right) + \left(\left(1 - P\left(L_{y\_destination}\right)\right) \times P\left(L_{y-1\_source}\right) \times k\right)$

And then use formula, $P\left(L_{y\_destination}\right)^{*}$, for knowledge going forward.

3. Repeat for each possible source skill and destination skill.

    a. Where a source skill is not equal to a destination skill.

We employed a brute force grid search approach to find the best fitting parameters (Baker et al., 2010), using a modified version of Baker's BKT-BF software package, for which open source was available. The values of G and S parameters were bounded to be below 0.3 to avoid model degeneracy issues (e.g., model's estimation of a lower probability of a student's knowing of each skill, $P(L_y)$, after observing the student's skill demonstration; Baker et al., 2008). All other parameters were allowed to range from 0.01 to 0.99. Within these bounds, we tried every set of parameters at a grain size of 0.01, using a brute force grid search approach (Baker et al., 2010). Brute force grid search was chosen as the fitting algorithm, due to speed limitations with other approaches (Thai-Nghe et al., 2012), and due to its ability to achieve comparable or better fit than other commonly used approaches, such as Expectation Maximization (Beck & Chang, 2007) and Iterative Gradient Descent (Baker et al., 2008). For each of the four folds of cross-validation, a k value was produced, and the overall estimate of the k value for the source and destination skills was computed from the average of the four values.

Findings

Table 2 shows the average k-value for each combination of destination skill and source skill. A higher value of k indicates that the model weights the past source skill scores more in calculating the current destination skill scores. We then used a Spearman's correlation (denoted by ρ) to investigate the predictive power of BKT-PSTC models (see the results in Appendix) in comparison with a non-PSTC BKT model. While the non-PSTC BKT assumes no transfer occurs among skills (equivalent to BKT-PSTC with $k = 0$), a BKT-PSTC model empirically estimates partial transfer (BKT-PSTC with $k > 0$). Specifically, in each pair of source skill and destination skill, we examined: (1) the correlation between students' actual skill scores and BKT-PSTC models' predicted scores, (2) the correlation between students' actual skill scores and non-PSTC BKT models' predicted scores, and (3) the correlation of predicted scores between the BKT-PSTC models and non-PSTC BKT models.

The BKT-PSTC models predict a student's score for a certain skill (a destination skill at one time point) by using the student's score on a source skill (at the previous time point) as an input. By comparison, the non-PSTC BKT model predicts a student's skill score (at one time point) using the students' same skill score at the previous time point. As shown in Appendix, for the non-PSTC model, the correlation coefficients (predictive performance) are the same for each destination skill since no transfer is occurring. We then used Hotelling's t-tests (Cohen et al., 2013) to compare the correlation coefficients of the BKT-PSTC and non-PSTC BKT, to determine if any cases had a statistically significant difference of the predictive power between the BKT-PSTC model and non-PSTC BKT model. If the BKT-PSTC models fit student skill scores better than the non-PSTC BKT models, it implies that partial transfer occurred among skills. For example, including Skill 9 as a source skill significantly improved our predictions of Skill 8 as a destination skill,

Table 2 Average *k*-value for each BKT-PSTC model fitted across four-folds

Destination skill	Source skill												
	0	1	2	3	4	5	6	7	8	9	10	11	12
0		0.24	0.24	0.20	0.28	0.19	0.20	0.13	0.17	0.15	0.05	0.17	0.14
1	0.06		0.25	0.30	0.32	0.23	0.23	0.17	0.23	0.17	0.01	0.19	0.17
2	0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
3	0.22	0.08	0.21		0.26	0.29	0.26	0.20	0.34	0.22	0.02	0.23	0.20
4	0.04	0.03	0.09	0.06		0.18	0.19	0.14	0.25	0.14	0.01	0.15	0.13
5	0.01	0.01	0.01	0.01	0.02		0.06	0.05	0.08	0.08	0.01	0.03	0.06
6	0.12	0.10	0.30	0.36	0.38	0.29		0.21	0.27	0.20	0.02	0.23	0.18
7	0.01	0.01	0.01	0.01	0.01	0.01	0.01		0.01	0.01	0.01	0.01	0.01
8	0.10	0.18*	0.35**	0.44**	0.47**	0.34**	0.30**	0.26**		0.25**	0.02	0.28**	0.22**
9	0.01	0.01	0.03	0.03	0.04	0.03	0.03	0.03	0.04		0.01	0.03	0.04
10	0.67	0.43	0.59	0.61	0.59	0.50	0.39	0.33	0.47	0.31		0.37	0.30
11	0.09	0.04	0.06	0.05	0.07	0.06	0.05	0.04	0.07	0.05	0.01		0.04
12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	

Significant results (Hotelling's two-tailed t-tests) are boldfaced (*$p < 0.05$, **$p < 0.01$)

compared to when we predicted Skill 8 without using any source skill ($\rho_{-PSTC} = 0.084$, $\rho_{-non-PSTC} = 0.014$, $t(333) = 3.373$, $p < 0.01$).

Interestingly, all the combinations of skills that returned p-values less than 0.05 are found only for the destination skill, CTR (8: Using appropriate experimental controls and replication). Nine source skills improved the model for this skill at a significant level, with most p values below 0.01. The nine source skills are.

- CON (2: Basing conclusions on results, $k = 0.353$; $t(333) = 2.862$, $p = 0.004$),
- LIM (3: Identifying limitations of the study, $k = 0.443$; $t(333) = 3.192$, $p = 0.002$),
- ALT (4: Identifying alternative explanations of findings, $k = 0.465$; $t(333) = 3.078$, $p = 0.002$),
- IMP (5: Discussing the implications of the findings, $k = 0.343$, $t(333) = 3.400$, $p = 0.001$),
- HYP (6: Establishing testable hypotheses, $k = 0.298$, $t(333) = 3.109$, $p = 0.002$),
- INT (7: Introducing/setting the study in context, $k = 0.255$; $t(333) = 3.327$, $p = 0.001$),
- EXP (9: Experimental design, $k = 0.253$; $t(333) = 3.373$, $p = 0.001$),
- LIT (11: Appropriately integrating primary literature, $k = 0.280$; $t(333) = 3.114$, $p = 0.002$), and
- WRT (12: Writing quality, $k = 0.223$; $t(333) = 3.617$, $p < 0.001$).

Individually, many of these tests would become non-significant, if subjected to a post-hoc test that treats tests as independent from one another, such as Bonferroni or Benjamini-Hochberg. However, the pattern seen here suggests that the tests are related—all of the significant tests involve CTR (8). To see if this pattern could have been due to chance, we ran a Monte Carlo simulation, where we assumed the null hypothesis that only chance p values were obtained, and that all tests were independent. We then ran the same number/structure of chance tests as in Table 2, asking the question of how often at least one destination skill (any of the 13 destination skills) has statistically significant results for at least 9 of the 12 tests. We ran 100,000 simulations, and obtained this pattern 0 times, suggesting that the pattern seen in this paper is highly unlikely ($p < 0.00001$) if due to chance.

This suggests that development of the CTR skill is dependent on development on a large range of other skills attained previously. For example, in the combination of CON (source skill) and CTR (destination skill), the average k value is 0.3525. BKT-PSTC assumes when predicting a CTR skill score at each time point a student has a certain amount of potential improvement in their knowledge of the skill (or, to be more precise, a certain amount of greater confidence that the model could have as to whether they know the skill). Considering how much of CTR skill still remains to be learned, the student gains 35.25% of their degree of knowledge of CON skill. In other words, assuming that a student has a 40% chance of knowing CTR skill, the model could still become 60% more confident in the student's knowledge. If the student showed 70% knowledge of CON skill gained at a previous time point, then the student gains 14.8% (60% × 70% × 35.25%), yielding a new estimate of 54.8% (40% + 14.8%) that they have attained the CTR skill.

The other source skills (ANA, SEL, and PRE) did not improve the CTR skill model at a significant level. As shown in Table 3, the models show a higher learning rate ($T > 0.01$), but overall a lower k value than the significant models. Despite such a higher probability that the CTR skill will be learned at each opportunity to use it, the relatively lower k values indicate that students may not readily transfer their competency of these source skills attained at a previous time point to the CTR skill.

Table 3 Parameter values for BKT-PSTC models for the destination skill, CTR

Source skill	$P(L_0)$	G	S	T	k
0 (ANA)	0.35	0.29	0.29	0.05	0.10
1 (SEL)	0.36	0.29	0.29	0.04	0.18
2 (CON)*	0.37	0.29	0.29	0.01	0.35
3 (LIM)**	0.36	0.29	0.29	0.01	0.44
4 (ALT)**	0.36	0.29	0.29	0.01	0.47
5 (IMP)**	0.36	0.29	0.29	0.01	0.34
6 (HYP)**	0.36	0.29	0.29	0.01	0.30
7 (INT)**	0.36	0.29	0.29	0.01	0.26
9 (EXP)**	0.35	0.29	0.29	0.01	0.25
10 (PRE)	0.35	0.29	0.29	0.06	0.02
11 (LIT)**	0.36	0.29	0.29	0.01	0.28
12 (WRT)**	0.36	0.29	0.29	0.01	0.22

Significant results (Hotelling's two-tailed t-test) are boldfaced (*$p < 0.05$, **$p < 0.01$)

Beyond this destination skill, the highest k value ($k = 0.6675$) was found for the combination of the source skill, ANA (0: Data analysis) and the destination skill, PRE (10: Presenting results). However, the associated BKT-PSTC models were not statistically significantly more predictive than the non-PSTC BKT model. It is worth asking why a high k may not be associated with a statistically significant finding. It may be that some non-significant findings, including this one, would have become significant with a larger sample. However, if the students' scores for these skills had a very high correlation, then a higher k value would do little to change the predictions because the source skill scores would not add much new information to the model. This is the more likely explanation for this case, as the scores for PRE and ANA have a fairly sizable correlation ($\rho = 0.5703$). Another possibility is that the non-PSTC BKT model had low predictive power to begin with, and adding information from the source skill did marginally improve the model, but not to a meaningful degree. This could happen if student performance is not stable across years. Indeed, even though the correlation between the non-PSTC BKT model and the students' actual scores for PRE is among the highest in the set ($\rho_{non\text{-}PSTC} = 0.251$), this correlation is fairly weak overall. We also note that for this ANA → PRE case, when k was included in the model (i.e., BKT-PSTC model) the model correlation was 0.2525, and when k was excluded (i.e., non-PSTC BKT) the model correlation was 0.2511. Because the predictions were almost identical ($\rho = 0.9791$), the value of k had limited impact on the actual predictions of the model, even though k was high. This can happen if performance of either skill is at floor or ceiling or if the two skills are highly correlated.

Discussion

The purpose of this study was to detect sequential dependencies between individual research skills in samples of students' scholarly writing using BKT-PSTC and determine if those identified dependencies were consistent with known threshold concepts. Our findings demonstrate the effectiveness of capturing the moments when the partial transfer occurs among a certain skill set (i.e., a source skill and destination skill) by

identifying explicit growth of research skill development over time through doctoral training experiences. Specifically, the identification of the control and replication skill was the only skill significantly predicted by any other skill acquired at previous time points. This pattern of data suggests that control and replication skills are highly contingent on the prior development of other skills necessary to conduct research in the biological sciences. The large number of contingencies further highlights the positioning of control and replication as a demanding concept that may not be attainable without prior mastery of other elements of research skill, consistent with the structure of threshold concepts. Coupled with results from previous qualitative research that identified the effective development and application of experimental controls as a key threshold concept in the biological sciences (Feldon et al., 2017), the findings of the current study point to control and replication as a threshold concept. This provides supporting evidence of the viability of BKT-PSTC as a means for detecting threshold concepts using quantified data at scale.

The identification of threshold concepts generally can be challenging, because their development is currently undertheorized. Despite extensive research, the field has yet to define clear indicators of progress during liminality that suggest a threshold concept might soon be attained. Likewise, the observable consequences of mastering a threshold concept are typically described in relation to the concept itself and not subsequent development that hinged upon its mastery (Nicola-Richmond et al., 2018; Salwën, 2019). However, based on the understanding that a threshold concept is both challenging to master and transformative (Meyer & Land, 2003), it likely relies on the development of a number of prerequisite skills and concepts before it can be attained. Thus, identifying a skill for which growth is predicted concurrently by prior growth in multiple other skills is generally consistent with an understanding of threshold concepts.

In the BKT-PSTC models, partial transfer between research skills was captured by the linear transfer factor, k. A high value of k suggests more transfer between a source skill and a destination skill, indicating specifically that the model weights the past source skill more in calculating the current destination skill. Our findings showed not all high k values were statistically significant in the comparison of the non-PSTC BKT models (e.g., $k = 0.6675$ between a source skill [data analysis] and a destination skill [presenting results]). Comparing coefficients of BKT-PSTC and non-PSTC BKT reveals a statistically significant difference in the predictive power between the models and further determines if high values of the parameter k actually implies that partial transfer occurred among skills. That is, comparing the predictive power between BKT-PSTC and non-PSTC BKT models, combined with observing high values obtained for the partial transfer parameter k can support the detection of threshold concept acquisition in a rigorous way.

It is worth noting that BKT-PSTC treats the degree of transfer between two skills (k) as being constant over time. It is possible that the transfer between skills may be higher at specific points in a student's academic career—for specific skill combinations, perhaps at the beginning of graduate study, or after the dissertation proposal. It may be a valuable area of future work to consider models where k is allowed to vary over time or by context, an extension of the contextual approach to parameter estimation seen in Baker et al. (2008).

Many studies of learning transfer typically use post-hoc tests to show any significant difference in performance scores between a treatment group and a control group. However, such tests only demonstrate the degree of transfer by using aggregated performance scores (e.g., pre-/post-test scores), instead of capturing the process of learning transfer. Drawing on educational data mining models such as BKT-PSTC in capturing transfer, this study suggests

that there may be benefits of the application of educational data mining models such as BKT-PSTC to yield a nuanced mechanism of how skills transfer dynamically to other skills.

Limitations

As a first effort at applying BKT-PSTC to detect threshold concepts at scale, the findings from the current study offer useful insight. However, the study as conducted has certain limitations. First, it is important to note that the sample was drawn from a single country and a single discipline. Accordingly, it is impossible to assert with confidence that the same results or patterns of performance would be obtained if either feature were different. The structure of Ph.D. training differs substantially across countries, with programs of study in the United States typically including at least one year of coursework, an emphasis on early publication, and interim benchmark assessments required for continuation to the dissertation phase of study (Gardner, 2009; Nerad & Heggelund, 2008). Accordingly, the type and sequencing of skills developed during graduate training may differ as a function of either or both of these features.

Second, despite the fact that the sample size is relatively large for the type of data collected in typical studies of graduate education, for the purpose of statistical analysis, it is somewhat small. Likewise, use of listwise deletion for missing data further reduced the operational data set. Accordingly, it is possible that lack of statistical significance for certain relationships might be due to limited statistical power to detect effects when standard errors are large. Limited sample size also prevented the disaggregation of data by gender or other demographic variables, as well as specialized subfields of the biological sciences (e.g., cellular and molecular biology, developmental biology, neurobiology). Therefore, it is possible that heterogeneity in relationships between skills differed in some way that was unobserved. However, other analyses of these data have not detected such effects (e.g., Feldon et al., 2019). Similarly, although our missing data meets Little's (1988) standard for missing completely at random (MCAR), it is possible that an undetected regularity in missingness across participants could introduce undetected bias into the current analyses.

Third, although the types of academic writing accepted for this study were constrained to either research proposals with discussion of anticipated findings or reports of obtained empirical findings, it is possible that the specific type of submission could introduce differences undetected during analysis. Even without systematic differences in mean score values between writing sample types, it is possible that standard errors or the magnitude of specific relationships between pairs of skills were impacted.

Lastly, the analytic strategy reported here employed only pairwise analyses of skills—i.e., whether a single source skill influenced later performance on a single destination skill. Thus, the analyses presented here could not detect unique contributions to destination skills by multiple source skills in combination (e.g., partial correlations). Likewise, it is possible that observed correlations could be affected by more complex multi-year skill development patterns. Our current data set and modeling approach could not investigate these questions but they represent relevant research questions for future research.

Implications

The results of this study hold several implications for both future research and practice in graduate education. First, the identification of a discipline-specific threshold concept that is robust across studies suggests that such research is both viable and necessary to understand the specific ways in which scholarly skills develop within individual disciplines. Most current research on threshold concepts in graduate education has focused on pan-disciplinary concepts that often link to scholarly identity as well as the broader ability to frame general scholarly arguments (e.g., Keefer, 2015; Kiley, 2009, 2015).

Second, the ability to detect threshold concepts using quantitative analysis permits studies that engage larger samples which can be randomly selected, rather than the small, purposeful samples typical of qualitative threshold concept studies. In contrast to the few prior quantitative studies to detect threshold concepts, which relied on student responses to multiple-choice exam questions (e.g., Shanahan et al., 2006; Vidal et al., 2015), the strategy employed here permits a deeper and more extensive examination of the knowledge used. Further, the use of large, randomly selected samples will be better positioned to establish broadly generalizable conclusions and permit better estimation of population parameters. Likewise, such analyses can be conducted using data sets compiled through institutional data, as well as independent research. For example, a number of universities have taken steps toward more systematic scoring of dissertations, theses, and other major benchmarks within graduate degree programs (Lovitts, 2007; Williams & Kemp, 2019). Accordingly, such data might be used to understand the development and subsequent impacts of threshold concepts in relation to the structure of training at the intersection of disciplines and academic programs.

Third, calling upon some of the well-established practical uses of BKT, universities might explore the use of such analyses to monitor the development of graduate students as they progress toward their degrees and alert appropriate faculty or staff if early warning signs emerge (cf., Milliron et al., 2014). If individual students do not demonstrate the attainment of identified threshold concepts on a normative timeline for a given discipline and program, it would be possible to identify them and offer additional programmatic supports tailored to the specific skills predictive of threshold concept attainment.

Lastly, the identification of experimental control and replication skills as a threshold concept has concrete applications for both the sequencing of instruction and the focal efforts of faculty mentors in supporting developing Ph.D. students. As a skill area that relies on partial transfer from multiple other skills, it is intuitive that focused instruction intended to support students' research skill development would introduce and facilitate mastery of the contributing skills prior to focusing on control and replication. Doing so would be likely to reduce the duration or severity of liminality prior to students crossing that threshold to mastery of the concept by ensuring that the necessary contributing skills were developed prior. Likewise, individualized support of doctoral students by supervisors or other faculty mentors might prioritize focused interactions to bolster the development of control and replication skills as a threshold concept following the perceived attainment of contributing skills. Because learning in Ph.D. programs is often solitary (Keefer, 2015) and student access to supervisors can be limited as a function of the time-consuming and diverse responsibilities of research faculty (Gappa et al., 2007; Jones et al., 2008), evidence-supported principles of which transitions during learning are most challenging could guide the strategic engagement of Ph.D. supervisors in the allocation of their time as a limited resource.

Conclusions

This study employed a novel approach based on Bayesian knowledge tracing (BKT), BKT-PSTC, to identify sequential dependencies between individual research skills in samples of students' scholarly writing. Analysis of the dependencies identified indicate that one specific skill was dependent upon growth in most others over time. This pattern of dependency is consistent with the relationships expected of a threshold concept, because the integrative nature of threshold concepts (Meyer & Land, 2003) and the common delays in acquiring threshold concepts (Keefer, 2015) indicate a synthesis of multiple facets of prior knowledge (Shanahan et al., 2006).

Specific to the discipline of biological sciences, the current findings indicated that the control and replication skills in the design of biology experiments is highly contingent on the prior development of other research skills. Further, these findings converge robustly with qualitative findings from prior research (i.e., Feldon et al., 2017), providing evidence of convergent validity. The ability to identify dependency patterns that are consistent with the identification of a threshold concept articulated through interviews with an independent sample provides strong supporting evidence of the viability of BKT-PSTC as a means for detecting threshold concepts using quantified data at scale. We highlight implications for future research and practice in graduate education including the benefits of the identification of disciplinary-specific threshold concepts using quantitative analyses, which expands our understanding of students' research skill development and facilitates mentoring Ph.D. students.

Appendix

Spearman's Correlations between students' actual performance and predicted performance for the two BKT models.

Destination	Source	0	1	2	3	4	5	6	7	8	9	10	11	12
0	[1]S-P		0.301	0.299	0.292	0.292	0.290	0.298	0.288	0.289	0.288	0.298	0.296	0.288
	[2]S-NP		0.291	0.291	0.291	0.291	0.291	0.291	0.291	0.291	0.291	0.291	0.291	0.291
	[3]P-NP		0.982	0.978	0.979	0.971	0.968	0.952	0.964	0.975	0.961	0.989	0.954	0.960
	[4]T^2		0.950	0.691	0.079	0.030	− 0.111	0.419	− 0.216	− 0.166	− 0.183	0.933	0.335	− 0.180
1	S-P	0.296		0.292	0.281	0.290	0.288	0.295	0.288	0.293	0.292	0.297	0.288	0.290
	S-NP	0.302		0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302
	P-NP	0.977		0.963	0.965	0.965	0.966	0.957	0.949	0.953	0.963	0.976	0.952	0.963
	T^2	-0.566		− 0.673	− 1.517	− 0.895	− 0.992	− 0.448	− 0.808	− 0.567	− 0.674	− 0.427	− 0.863	− 0.817
2	S-P	0.162	0.162		0.162	0.162	0.156	0.156	0.156	0.156	0.155	0.162	0.155	0.156
	S-NP	0.160	0.160		0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160	0.160
	P-NP	0.961	0.964		0.964	0.961	0.957	0.957	0.953	0.957	0.956	0.961	0.953	0.957
	T^2	0.110	0.132		0.141	0.123	− 0.280	− 0.248	− 0.266	− 0.280	− 0.319	0.118	− 0.286	− 0.290
3	S-P	0.133	0.124	0.115		0.111	0.125	0.120	0.122	0.140	0.127	0.120	0.121	0.117

Destination	Source	0	1	2	3	4	5	6	7	8	9	10	11	12
	S-NP	0.099	0.099	0.099		0.099	0.099	0.099	0.099	0.099	0.099	0.099	0.099	0.099
	P-NP	0.911	0.918	0.910		0.902	0.880	0.854	0.854	0.823	0.823	0.922	0.859	0.872
	T^2	1.498	1.123	0.705		0.494	0.990	0.720	0.785	1.273	0.877	0.961	0.760	0.670
4	S-P	0.122	0.119	0.109	0.113		0.121	0.118	0.118	0.116	0.119	0.119	0.119	0.116
	S-NP	0.107	0.107	0.107	0.107		0.107	0.107	0.107	0.107	0.107	0.107	0.107	0.107
	P-NP	0.898	0.911	0.902	0.916		0.923	0.912	0.917	0.920	0.908	0.923	0.931	0.915
	T^2	0.620	0.546	0.097	0.295		0.654	0.481	0.512	0.441	0.547	0.584	0.621	0.426
5	S-P	0.013	0.012	0.010	0.012	0.014		0.011	0.012	0.013	0.013	0.013	0.016	0.011
	S-NP	0.007	0.007	0.007	0.007	0.007		0.007	0.007	0.007	0.007	0.007	0.007	0.007
	P-NP	0.986	0.986	0.986	0.986	0.979		0.984	0.982	0.985	0.976	0.986	0.978	0.979
	T^2	0.717	0.633	0.372	0.539	0.649		0.434	0.466	0.625	0.554	0.657	0.818	0.425
6	S-P	0.081	0.080	0.077	0.075	0.079	0.080		0.082	0.078	0.084	0.081	0.080	0.085
	S-NP	0.077	0.077	0.077	0.077	0.077	0.077		0.077	0.077	0.077	0.077	0.077	0.077
	P-NP	0.966	0.966	0.954	0.949	0.950	0.942		0.936	0.945	0.936	0.966	0.944	0.934
	T^2	0.250	0.218	0.019	− 0.111	0.095	0.164		0.228	0.047	0.373	0.274	0.141	0.400
7	S-P	0.047	0.047	0.047	0.047	0.047	0.044	0.044		0.044	0.044	0.047	0.044	0.043
	S-NP	0.047	0.047	0.047	0.047	0.047	0.047	0.047		0.047	0.047	0.047	0.047	0.047
	P-NP	0.944	0.944	0.944	0.945	0.944	0.942	0.941		0.941	0.941	0.944	0.941	0.940
	T^2	0.015	− 0.012	− 0.005	− 0.016	0.011	− 0.155	− 0.159		− 0.142	− 0.173	− 0.007	− 0.182	− 0.207
8	S-P	0.035	0.038	0.057	0.069	0.066	0.072	0.069	0.082		0.084	0.032	0.068	0.080
	S-NP	0.014	0.014	0.014	0.014	0.014	0.014	0.014	0.014		0.014	0.014	0.014	0.014
	P-NP	0.979	0.976	0.962	0.950	0.952	0.951	0.947	0.929		0.926	0.980	0.949	0.942
	T^2	1.904	2.015*	2.862**	3.192**	3.077**	3.399**	3.109**	3.32**		3.373**	1.623	3.114**	3.617**
9	S-P	0.055	0.055	0.050	0.050	0.052	0.052	0.054	0.055	0.055		0.055	0.054	0.053
	S-NP	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067		0.067	0.067	0.067
	P-NP	0.931	0.931	0.933	0.933	0.932	0.932	0.918	0.917	0.918		0.931	0.918	0.915
	T^2	− 0.585	− 0.586	− 0.878	− 0.857	− 0.737	− 0.768	− 0.569	− 0.563	− 0.534		− 0.582	− 0.585	− 0.626
10	S-P	0.253	0.247	0.247	0.229	0.244	0.230	0.245	0.235	0.234	0.232		0.241	0.234
	S-NP	0.251	0.251	0.251	0.251	0.251	0.251	0.251	0.251	0.251	0.251		0.251	0.251
	P-NP	0.979	0.987	0.963	0.946	0.971	0.957	0.951	0.961	0.944	0.962		0.948	0.963
	T^2	0.140	− 0.530	− 0.310	− 1.295	− 0.556	− 1.339	− 0.370	− 1.069	− 0.950	− 1.307		− 0.582	− 1.214
11	S-P	0.077	0.080	0.079	0.080	0.078	0.079	0.078	0.079	0.075	0.076	0.082		0.079
	S-NP	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077	0.077		0.077
	P-NP	0.931	0.934	0.931	0.936	0.932	0.931	0.932	0.932	0.940	0.940	0.938		0.932
	T^2	0.029	0.174	0.091	0.177	0.067	0.093	0.051	0.108	-0.084	-0.044	0.255		0.096
12	S-P	0.082	0.084	0.085	0.085	0.085	0.075	0.076	0.076	0.075	0.076	0.082	0.076	
	S-NP	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052	0.052	
	P-NP	0.847	0.850	0.849	0.850	0.850	0.848	0.848	0.848	0.848	0.848	0.847	0.848	
	T^2	0.980	1.065	1.099	1.079	1.094	0.756	0.788	0.776	0.759	0.777	0.964	0.768	

[1]Spearman's correlations between student scores & BKT-PSTC predictions

[2]Spearman's correlations between student scores & non-PSTC BKT predictions

[3]Spearman's correlations between BKT-PSTC predictions & non-PSTC BKT predictions

[4]Hotelling's two-tailed t-test

*Significant relationship at $p < 0.05$

**Significant relationship at $p < 0.01$

Funding This material is based upon work supported by the United States National Science Foundation under Awards 1431234, 1431290, and 1760894. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Data availability The data generated and analyzed during the current study are available in Utah State University's Open Access Institutional Repository at https://doi.org/10.26078/X535-HW49.

Code availability The statistical code used to analyze the data during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Feldon, D. F., Litson, K., Jeong, S., Blaney, J. M., Kang, J., Miller, C., ... Roksa, J. (2019). Postdocs' lab engagement predicts trajectories of PhD students' skill development. In *Proceedings of the National Academy of Sciences, 116*(42), 20910–20916. https://doi.org/10.1073/pnas.1912488116

Ackerman, P. L., & Beier, M. E. (2018). Methods for studying the structure of expertise: Psychometric approaches. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (pp. 213–232). New York: Cambridge University Press

Adjei, S., Selent, D., Heffernan, N., Pardos, Z., Broaddus, A., & Kingston, N. (2014). Refining learning maps with data fitting techniques: Searching for better fitting learning maps. In J. Stamper, Z. Pardos, M. Mavrikis & B. M. McLaren (Eds.) *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 413–414).

Baker, R. S. J .d., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probability in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring System* (pp. 406–415). https://doi.org/10.1007/978-3-540-69132-7_44

Baker, R., Corbett, A., Gowda, S., Wagner, A., MacLaren, B., Kauffman, L., ... Giguereet S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization* (pp. 52–63).

Baker, R.S., Gowda, S.M., & Salamin, E. (2018) Modeling the learning that takes place between online assessments. In *Proceedings of the 26th International Conference on Computers in Education* (pp. 21–28).

Beck, J. E., & Chang, K. (2007). Identifiability: A fundamental problem of student modeling. In *Proceedings of the International Conference on User Modeling*, pp. 137–146. Springer. https://doi.org/10.1007/978-3-540-73078-1_17

Beck, J. E., Chang, K. M., Mostow, J., & Corbett, A. (2008). Does help help? Introducing the Bayesian Evaluation and Assessment methodology. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 383–394). Springer.

Botelho, A., Wan, H., & Heffernan, N. (2015, March). The prediction of student first response using prerequisite skills. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 39–45).

Boote, D. N., & Beile, P. (2005). Scholars before researchers: On the centrality of the dissertation literature review in research preparation. *Educational Researcher, 34*(6), 3–15

Chatterjee-Padmanabhan, M., Nielsen, W., & Sanders, S. (2018). Joining the research conversation: Threshold concepts embedded in the literature review. *Higher Education Research and Development*, 38, 494–507

Chen, Y., González-Brenes, J. P., & Tian, J. (2016). Joint discovery of skill prerequisite graphs and student models. In: *Proceedings of the 9th International Conference on Educational Data Mining* (pp. 46–53)

Chen, P., Lu, Y., Zheng, V. W., & Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)* (pp. 39–48). IEEE

Christensen, B. T., & Schunn, C. D. (2009). The role and impact of mental simulation in design. *Applied Cognitive Psychology, 23*(3), 327–344. https://doi.org/10.1002/acp.1464.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/ correlation analysis for the behavioral sciences*. Routledge.

Corbett, A., T., & Anderson, J., R (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278

Cox, M. F., & Andriot, A. (2009). Mentor and undergraduate student comparisons of students' research skills. *Journal of STEM Education: Innovations and Research*, 10(1/2), 31

Delamont, S., & Atkinson, P. (2001). Doctoring uncertainty: Mastering craft knowledge. *Social Studies of Science*, 31(1), 87–107. doi:https://doi.org/10.1177/030631201031001005

Doignon, J. P., & Falmagne, J. C. (2012). *Knowledge spaces*. Springer.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3), 363

Feldon, D. F. (2010). Do psychology researchers tell it like it is? A microgenetic analysis of research strategies and self-report accuracy. *Instructional Science, 38*, 395–415.

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

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., Sun, V., & Rates, C. (2017). Doctoral threshold concepts in the biological sciences. *International Journal of Science Education, 18*, 2574–2593.

Florence, M. K., & Yore, L. D. (2004). Learning to write like a scientist: Coauthoring as an enculturation task. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 41(6), 637–668

Gappa, J. M., Austin, A. E., & Trice, A. G. (2007). *Rethinking faculty work: Higher education's strategic imperative*. New York: Jossey-Bass

Gardner, S. K. (2009). Conceptualizing success in doctoral education: Perspectives of faculty in seven disciplines. *The Review of Higher Education*, 32(3), 383–406

Goldstein, S. D., Lindeman, B., Colbert-Getz, J., Arbella, T., Dudas, R., Lidor, A., & Sacks, B. (2014). Faculty and resident evaluations of medical students on a surgery clerkship correlate poorly with standardized exam scores. *The American Journal of Surgery*, 207, 231–235

Gross, A. J., & Mantel, N. (1967). The effective use of both positive and negative controls in screening experiments. *Biometrics*, 23, 285–295. doi:https://doi.org/10.2307/2528162

Hackett, E. J., & Rhoten, D. R. (2009). The snowbird charrette: Integrative interdisciplinary collaboration in environmental research design. *Minerva, 47*(4), 407–440. https://doi.org/10.1007/s11024-009-9136-0.

Hmelo-Silver, C. E., Nagarajan, A., & Day, R. S. (2002). It's harder than we thought it would be": A comparative case study of expert–novice experimentation strategies. *Science education*, 86(2), 219–243

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322, 1259–1262

Kardash, C. M. (2000). Evaluation of undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of educational psychology*, 92(1), 191

Keefer, J. M. (2015). Experiencing doctoral liminality as a conceptual threshold and how supervisors can use it. *Innovations in Education and Teaching International*, 52, 17–28. doi:https://doi.org/10.1080/14703297.2014.981839

Kiley, M. (2009). Identifying threshold concepts and proposing strategies to support doctoral candidates. *Innovations in Education and Teaching International*, 46(3), 293–304. doi:https://doi.org/10.1080/14703290903069001

Kiley, M. (2015). 'I didn't have a clue what they were talking about': PhD candidates and theory. *Innovations in Education and Teaching International*, 52, 52–63. doi:https://doi.org/10.1080/14703297.2014.981835

Kiley, M., & Wisker, G. (2009). Threshold concepts in research education and evidence of threshold crossing. *Higher Education Research & Development*, 28(4), 431–441. doi:https://doi.org/10.1080/07294360903067930

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA: Harvard University Press

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (pp. 61–77). New York: Cambridge University Press

Kuhn, T. S. (1962). *The structure of scientific revolutions*. The University of Chicago Press.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. Arguing about science, 74–86

Land, R., Cousin, G., Meyer, J. H., & Davies, P. (2005). Threshold concepts and troublesome knowledge (3): implications for course design and evaluation. In C. Rust (Ed.), *Improving student learning diversity and inclusivity* (pp. 53–64). Oxford: Oxford Centre for Staff and Learning Development

Leshem, S. (2020). Identity formations of doctoral students on the route to achieving their doctorate. *Issues in Educational Research*, 30, 169–186

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198–1202

Lovitts, B. (2005). Being a good course taker is not enough: A theoretical perspective on the transition to independent research. *Studies in Higher Education*, 30(2), 137–154

Lovitts, B. (2007). *Making the implicit explicit: Creating performance expectations for the dissertation.* Stylus

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

McAlpine, L., & McKinnon, M. (2013). Supervision—The most variable of variables: Student perspectives. *Studies in Continuing Education, 35,* 265–280.

Meyer, J. H. F., & Land, R. (2003). Threshold concepts and troublesome knowledge (1): Linkages to ways of thinking and practising with the disciplines. In C. Rust (Ed.), *Improving student learning – Ten years on* (pp. 1–16). Oxford: OCSLD

Meyer, J., & Land, R. (Eds.). (2006). *Overcoming barriers to student understanding: Threshold concepts and troublesome knowledge*. Abingdon, UK: Routledge

Meyer, J., Land, R., & Baillie, C. (Eds.). (2010). *Threshold concepts and transformational learning*. Rotterdam, Netherlands: Sense Publishers

Milliron, M. D., Malcolm, L., & Kil, D. (2014). Insight and Action Analytics: Three Case Studies to Consider. *Research & Practice in Assessment*, 9, 70–89

Murre, J. M. J. (2014). S-shaped learning curves. *Psychonomic Bulletin & Review*, 21, 344–356

Nerad, M., & Heggelund, M. (Eds.). (2008). *Towards a Global PhD? Forces and Forms in Doctoral Education Worldwide*. Seattle: University of Washington Press

Nersessian, N., & Chandrasekharan, S. (2009). Hybrid analogies in conceptual innovation in science. *Cognitive Systems Research*, 10(3), 178–188

Nicola-Richmond, K., Pépin, G., Larking, H., & Taylor, C. (2018). Threshold concepts in higher education: A synthesis of the literature relating to measurement of threshold crossing. *Higher Education Research and Development*, 37, 101–114

Pardos, Z. A., & Dadu, A. (2018). dAFM: Fusing Psychometric and Connectionist Modeling for Q-matrix Refinement. *Journal of Educational Data Mining*, 10(2), 1–27. https://doi.org/10.5281/zenodo.3554689

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525–556

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in neural information processing systems* (pp. 505-513)

Roberts, R. (2016). Understanding the validity of data: A knowledge-based network underlying research expertise in scientific disciplines. *Higher Education*, 72, 651–668

Salwën, H. (2019). Threshold concepts, obstacles or scientific dead ends? *Teaching in Higher Education*. doi:https://doi.org/10.1080/13562517.2019.1632828

Sao Pedro, M., Baker, R., & Gobert, J. (2013). Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 185–192).

Sao Pedro, M. A., Jiang, Y., Paquette, L., Baker, R. S., & Gobert, J. D. (2014). Identifying transfer of inquiry skills across physical science simulations using educational data mining. In *Proceedings of the 11th International Conference of the Learning Sciences* (pp. 222–229).

Scheines, R., Silver, E., & Goldin, I. M. (2014, May). Discovering Prerequisite Relationships Among Knowledge Components. In J. Stamper, Z. Pardos, M. Mavrikis & B. M. McLaren (Eds.) *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 355-356). London

Schraagen, J., & Maarten (1993). How experts solve a novel problem in experimental design. *Cognitive Science*, 17(2), 285–309. doi:https://doi.org/10.1016/0364-0213(93)90013-X

Schunn, C. D., & Anderson, J. R. (1999). The Generality/Specificity of Expertise in Scientific Reasoning. *Cognitive Science*, 23(3), 337–370. doi:https://doi.org/10.1207/s15516709cog2303_3

Seashore Louis, K., Holdsworth, J. M., Anderson, M. S., & Campbell, E. G. (2007). Becoming a scientist: The effects of work-group size and organizational climate. *The Journal of Higher Education*, 78(3), 311–336

Shanahan, M., Foster, G., & Meyer, J. (2006). Operationalising a threshold concept in economics: A pilot study using multiple choice questions on opportunity cost. *International Review of Economics Education*, 29 –57

Singley, M., & Anderson, J., R (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press

Speelman, C. P., & Kirsner, K. (1997). The specificity of skill acquisition and transfer. *Australian Journal of Psychology*, 49(2), 91–100

Thagard, P. (2003). Pathways to biomedical discovery. *Philosophy of Science*, 70, 235–254

Thai-Nghe, N., Drumond, L., Horváth, T., & Schmidt-Thieme, L. (2012, July). Using factorization machines for student modeling. In *Proceedings of the Workshops of the International Conference on User Modeling, Adaptation, and Personalization.*

Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment & Evaluation in Higher Education*, 36(5), 509–547

Timmerman, B., Feldon, D., Maher, M., Strickland, D., & Gilmore, J. (2013). Performance-based assessment of graduate student research skills: Timing trajectory, and potential thresholds. *Studies in Higher Education, 38*, 693–710.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51, 1–10

Urquhart, S. M., Maher, M. A., Feldon, D. F., & Gilmore, J. (2016). Factors associated with novice graduate student researchers' engagement with primary literature. *International Journal for Researcher Development, 7*(2), 141–158. https://doi.org/10.1108/IJRD-11-2015-0029

Vidal, N., Smith, R., & Spetic, W. (2015). Designing and teaching business and society courses from a threshold concept approach. *Journal of Management Education*, 39, 497–530

Williams, L., & Kemp, S. (2019). Independent markers of master's theses show low levels of agreement. *Assessment & Evaluation in Higher Education*, 44, 764–771

Wisker, G. (2015). Developing doctoral authors: Engaging with theoretical perspectives through the literature review. *Innovations in Education and Teaching International*, 52, 64–74

Yudelson, M. V., Medvedeva, O. P., & Crowley, R. S. (2008). A multifactor approach to student model evaluation. *User Modeling and User-Adapted Interaction*, 18(4), 349–382

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.