Entering Research Learning Assessment (ERLA): Validity Evidence for an Instrument to Measure Undergraduate and Graduate Research Trainee Development

Amanda R. Butz[†*] and Janet L. Branchaw[†‡]
[†]Wisconsin Institute for Science Education and Community Engagement and [‡]Department of Kinesiology, University of Wisconsin–Madison, Madison, WI 53706

ABSTRACT

Expanding the scope of previous undergraduate research assessment tools, the *Entering Research* Learning Assessment (ERLA) measures undergraduate and graduate research trainee learning gains in the seven areas of trainee development in the evidence-based *Entering Research* conceptual framework: Research Comprehension and Communication Skills, Practical Research Skills, Research Ethics, Researcher Identity, Researcher Confidence and Independence, Equity and Inclusion Awareness and Skills, and Professional and Career Development Skills. In this paper, we present multiple sources of validity evidence for the ERLA trainee self-assessment and mentor assessment of trainee learning gains. Evidence of internal structure of the initial scales via exploratory factor analysis ($N_{trainees}$ = 193; $N_{mentors}$ = 130) revealed seven factors that align with the *Entering Research* conceptual framework. Validity evidence for internal structure using confirmatory factor analysis, convergent validity, and evidence of internal consistency for the revised scale were examined with a larger sample ($N_{trainees}$ = 489; $N_{mentors}$ = 256). Evidence of internal structure and alignment for a paired version of the ERLA was also examined with a subset of the original sample (N = 121 pairs). Each analysis revealed acceptable model–data fit. Guidance on using the ERLA instruments and interpreting their scores is presented.

INTRODUCTION

Mentored research experiences play an important role in the development of undergraduate and graduate students in science, technology, engineering, mathematics, and medicine (STEMM); positive research experiences and mentor–trainee relationships predict long-term trainee success (Eagan *et al.*, 2013; National Academies of Sciences, Engineering, and Medicine [NASEM], 2017, 2018). Notably, interventions that include mentored undergraduate research experiences have been shown to increase the persistence of students from diverse populations in STEMM (e.g., Eagan *et al.*, 2013). To understand how mentored research experiences produce positive outcomes, research training program directors and education researchers need assessment tools that can accurately measure the gains in research trainee learning and development that occur as a result of the research experience. An important component of assessing mentored research experiences has been trainee self-reported assessment of skills or learning gains.

Several tools exist to measure learning gains and outcomes of research experiences; yet few tools offer a multidimensional approach to assessing trainee learning. Of the multidimensional tools available, many omit key areas of trainee development, such as research ethics or promoting equity and inclusion in the research environment. A comprehensive assessment can help training program directors efficiently assess the

Ross Nehm. *Monitoring Editor*
Submitted Jul 31, 2019; Revised Jan 31, 2020; Accepted Feb 20, 2020
CBE Life Sci Educ June 1, 2020 19:ar18
DOI:10.1187/cbe.19-07-0146
*Address correspondence to: Amanda R. Butz (abutz2@wisc.edu).

progress of trainees while they simultaneously identify gaps in their training. To expand the scope of current self-assessment instruments, we developed and collected validity evidence for the *Entering Research* Learning Assessment (ERLA), which collects trainee assessment data from trainees and their mentors that can be aligned and compared. The ERLA expands the comprehensiveness of available assessments by incorporating two important but largely underassessed areas of the research training experience: research ethics and equity and inclusion. The ERLA aligns with learning objectives and outcomes that were derived from the literature and informed by STEMM practitioners and social science researchers. Finally, the ERLA includes an assessment of trainee learning gains from the perspective of the mentor that complements the trainee self-assessment and aligns with these same standards.

Evidence of Validity

Collecting validity evidence for an instrument is an ongoing process. The American Educational Research Association, in conjunction with the American Psychological Association, National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing have developed standards for educational and psychological testing and provide several recommendations related to validity evidence for instruments (AERA *et al.*, 2014). Included in their recommendations are descriptions of different types of validity evidence that can be used separately or together to provide evidence that the instrument can be used and its scores interpreted in the intended way. These include evidence based on test content, evidence of response processes, evidence based on internal structure, and evidence based on relations to other variables. Not all types of validity evidence are needed in order to consider whether an instrument is "valid"; instead, different types of validity evidence provide information on whether the instrument scores can be interpreted in the way in which they were intended for a specific population and context (AERA *et al.*, 2014). Reliability and internal consistency measurements (e.g., Cronbach's alpha) provide information about the expected consistency of scores across different administrations. (See Supplemental Material for brief summaries of validity evidence and reliability.)

Many metrics are used to evaluate evidence of internal structure (i.e., model–data fit) in new and revised instruments. These include the chi-square statistics, chi-square/degrees of freedom (χ^2/df) ratios, root-mean-square error of approximation (RMSEA), comparative fit index (CFI), and Tucker Lewis index (TLI). Lower χ^2 values and χ^2/df ratios of between 2 and 3 often indicate acceptable fit (Schreiber *et al.*, 2006). CFI values above 0.95, TLI values less than 0.95, and RMSEA values less than 0.06 are also used as criteria for acceptable model–data fit (Hu and Bentler, 1999; Schreiber *et al.*, 2006). We used these guidelines to develop the ERLA and to evaluate how other assessments of trainee learning gains have been developed. It is important to note, however, that different threshold values for each of these fit statistics have been recommended by different researchers (e.g., Hu and Bentler, 1999; Schreiber *et al.*, 2006; Knekta *et al.*, 2019).

Several instruments have been developed and evidence of validity and reliability collected to assess skills and learning gains in the context of mentored research experiences. Some instruments focus on only one or two aspects of the research experience and trainee development, such as research self-efficacy and outcome expectations (Byars-Winston *et al.*, 2016) or science identity and values (Chemers *et al.*, 2010; Estrada *et al.*, 2011). Here, we focus on instruments that were designed to assess trainee learning and skill gains across multiple dimensions. These instruments, the Undergraduate Research Student Self-Assessment tool (URSSA; Weston and Laursen, 2015), the Survey of Undergraduate Research Experiences (SURE; Lopatto, 2004), and the Mentor Competency Assessment (MCA; Fleming *et al.*, 2013) adapted for use with trainees, served as the starting point for understanding the current landscape of research trainee assessment and for development of the ERLA.

Existing Multidimensional Instruments for Assessing Research Trainee Learning Gains

A brief overview of the validity evidence for the URSSA, SURE, and MCA instruments is provided in Table 1. The URSSA (Weston and Laursen, 2015) is a 34-item survey designed to assess research trainees' gains. Respondents are asked to self-assess how much they gained as a result of their most recent research experience, with options ranging from "no gains" to "great gains." Evidence of content validity for this instrument was provided by developing and aligning items to the key benefits and gains of research experiences identified through qualitative interviews with students and faculty engaged in undergraduate research experiences: personal and professional development, knowledge in thinking and working related to research, gains in skills, gains in professional socialization, and preparation for graduate school and career (Hunter *et al.*, 2009). Evidence of validity based on internal structure was provided through a confirmatory factor analysis and internal consistency statistics.

The SURE (Lopatto, 2004, 2007) is another commonly used survey for assessing research trainee learning gains. Like the URSSA items, the SURE items were designed to align with key themes identified in previous research. To identify these themes, Lopatto (2003) invited science faculty to share what they thought were the essential features of a successful undergraduate research experience; students from four institutions were then asked to select the most important benefits of undergraduate research experiences that had been identified by faculty and by the literature. This resulted in the 20 learning gains items. The SURE is designed to assess the benefits that trainees gained from their research experience. Responses can range from "no gain" to "very large gain," with an option to indicate that a response is not applicable. Evidence of content validity for the SURE is provided through the alignment of items with the research experience outcomes identified by Lopatto (2003) in his prior work with science faculty and students. Internal consistency statistics were also calculated.

The MCA (Fleming *et al.*, 2013) was developed to assess mentors' skills across six mentoring competencies; items were aligned to a mentor training curriculum in the *Entering Mentoring* series (Pfund *et al.*, 2013, 2015). Items were developed to align with each of these competencies and were reviewed by survey experts and via cognitive interviews with mentors and trainees. Validity evidence based on internal structure was presented, as was evidence of internal consistency. An adapted version of the MCA has been developed for use with trainees.

TABLE 1. Existing multidimensional instruments for assessing skills and knowledge of research trainees and mentors

Instrument[a]	Number of items and constructs measured	Population from which validity evidence was collected	Model–data fit statistics	Internal consistency statistics (α)
URSSA (Weston and Laursen, 2015)	34 items evaluating undergraduate trainees' gains: Skills (12 items); Thinking and Working Like a Scientist (8 items); Personal Gains (6 items); Attitudes and Behaviors as a Researcher (8 items)	506 undergraduate trainees from United States and Canada	$\chi^2(458) = 1418$* RMSEA = 0.064 CFI = 0.76	0.83–0.92
SURE (Lopatto, 2004, 2007)	20 items evaluating undergraduate trainees' gains related to the research experience. Items are designed to be scored individually.	1135 undergraduates (59% women; 57% White; 9% African American; 16% Asian American; 5% Hispanic; 6% other or multiracial)	N/A	0.92–0.94
MCA (Fleming *et al.*, 2013)	26 items evaluating mentors' skill gains: Effective Communication (6 items); Aligning Expectations (5 items); Assessing Understanding (3 items); Fostering Independence (5 items): Addressing Diversity (2 items); Promoting Professional Development (5 items)	283 mentors (40% women; 90% White; 2% African American; 8%; Asian; 2% other; 7% Hispanic/Latino). Participants were mentors of faculty (46%) research scientists (5%), and students/fellows (49%)	χ^2 (284) = 663.20* RMSEA = 0.069 CFI = 0.85 TLI = 0.83	0.62–0.90
ERLA (present study)	53 items evaluating undergraduate and graduate trainee's gains: Research Comprehension and Communication Skills (15 items); Practical Research Skills (13 items); Research Ethics (3 items); Researcher Identity (6 items); Researcher Confidence and Independence (7 items); Equity and Inclusion Awareness and Skills (5 items); Professional and Career Development Skills (4 items). Parallel version for mentors to assess trainees' skills also available (47 items)	490 undergraduate and graduate trainees (see Table 5)	χ^2 (1304) = 3333.766* χ^2/df= 2.56 RMSEA = 0.056 CFI = 0.957	0.86–0.95

[a]The MCA assesses the skill gains of research mentors. An adapted version of this instrument used to assess trainee gains is also available, but no validity evidence on this version of the instrument is currently available. The validity evidence presented for the ERLA refers to the data collected for trainees and presented in stage 4; see Table 6 in this paper for additional information on the properties of this scale for trainees and mentors.
*$p < 0.001$.

Limitations of Existing Multidimensional Assessment Tools

The URSSA, SURE, and mentee version of the MCA are all tools widely used by training programs to assess the benefits of mentored research experiences and the skill levels of research trainees. However, these instruments have several limitations.

Content Limitations. Two recent reports, *Undergraduate Research Experiences for STEM Students* (NASEM, 2017) and *Graduate STEM Education for the 21st Century* (NASEM, 2018), recommend learning goals for undergraduate and graduate trainees. These goals include expected domains such as disciplinary knowledge and skills and professional skills that are addressed in the URSSA, SURE, and MCA, but also include learning goals focused on ethical development, cultural issues in research, and promoting inclusive learning environments, which are not addressed in the URSSA and SURE. The ERLA and the *Entering Research* conceptual framework with which it aligns were informed by the broad recommendations in these reports and therefore provide a more comprehensive training and assessment strategy for research training programs that aspire to respond to the recommendations.

The URSSA, SURE, and MCA each focus on the research experience and/or the research mentoring relationship, but these are not the only places where students may gain research skills and knowledge. Indeed, the communities and structured learning experiences (e.g., seminars and workshops) in formal research training programs have the potential to augment and expand upon what is learned in the research experience (Balster *et al.*, 2010). Thus, it is important to use assessments that can capture how training programs and research experiences work in concert to promote trainee development. In developing the ERLA, we sought to create an instrument that was comprehensive enough in scope to capture learning across the research experience, the mentoring relationship, and the training program activities. To that end, the ERLA question stem invites trainees to consider gains over the course of their research experiences, not just their experiences doing research, and assesses areas of trainee development not assessed in previous instruments.

The URSSA and SURE contain items that align with aspects of the research experience important to trainee development (i.e., building research skills, communicating about research, and the research mentoring relationship) but do not address equity and inclusion, a critical factor shown to impact research trainee development and retention in STEM. An unwelcoming climate in the research environment, stereotype threat, and bias

can negatively impact the participation of individuals from historically underrepresented groups in STEM (Steele, 1997; President's Council of Advisors on Science and Technology, 2012; Valantine and Collins, 2015). Incorporating discussions of these topics into the mentored research and training experiences acknowledges that cultural identities and the extent to which trainees feel welcome in the research environment can impact trainee perceptions of research and the research environment (Bumpus, 2015; Butz *et al.*, 2018). Currently, few tools exist to assess underrepresented trainees' growth in their capability to navigate these challenges or to assess well-represented trainees' abilities to recognize their biases and assumptions and act as advocates for those who may be marginalized or excluded from research. New tools are needed to empower research training programs to measure the development of their trainees' awareness and skills in equity and inclusion as they work to respond to calls to broaden participation in STEMM.

Notably, the adapted version of the MCA for trainees contains items designed to assess equity and inclusion; these items were adapted in the development of the equity and inclusion ERLA items. However, the trainee version of the MCA does not provide a comprehensive assessment of trainee research learning gains. It is based on the competencies of research mentor training and thus narrowly focuses on skills related to the mentoring relationship.

Limitations of Validity Evidence and Measurement. The URSSA, SURE, and trainee version of the MCA each provide validity evidence. However, this validity evidence comes with some important limitations. Weston and Laursen (2015) call for refinement of the URSSA instrument to further establish the factor structure of the scale. The SURE provides limited validity evidence. Although the internal consistency statistics for the SURE are high, these internal consistency statistics alone do not provide much information or guidance on the interpretation of scores for this instrument. Cronbach's alpha is sensitive to the number of items in a scale and is meant to be calculated for unidimensional scales; it does not provide evidence of the factor structure of a given scale (Cronbach, 1951). Although validity evidence for the mentor version of the MCA is provided, validity evidence for the trainee version of the MCA scales has not been systematically analyzed and reported. Finally, none of the three measures were designed to assess both undergraduate and graduate trainee development, nor were these instruments designed to be used as paired trainee and mentor assessment instruments.

The ERLA addresses these gaps. First, items for the ERLA align with the areas of trainee development and learning objectives outlined in the *Entering Research* conceptual framework (Branchaw *et al.*, 2020; Supplemental Figure S1), which apply to both undergraduate and graduate research trainees. Therefore, ERLA can assess undergraduate or graduate trainee development and provide scores across seven constructs associated with trainee development. Second, methodology used to develop ERLA included the collection of multiple sources of validity evidence (AERA *et al.*, 2014). The multiple sources address the call from previous researchers who have noted that measures in some domains of science education lack validity and reliability evidence (Campbell and Nehm, 2013). Finally, the ERLA begins to address the challenge of using only self-reported data to document research trainee gains. The paired assessment allows training program directors to obtain the mentor's assessment of trainee gains, which can be compared against the trainee's self-reported gains or used on its own. Together, the paired assessments provide two measures of trainee development and also provide information about the alignment between trainee and mentor assessments of trainee skills and knowledge. By collecting validity evidence for a paired scale and providing guidance for the scoring and use of these paired scores, ERLA can be used to assess trainee gains based not only on self-reported data from trainees, but also via comparison data collected from mentors based on their observations.

The Present Study

In response to the gaps identified previously and in conjunction with the development of a revised and expanded curriculum to support comprehensive research trainee development, we first developed a comprehensive set of meta–learning objectives (i.e., learning objectives that applied across multiple *Entering Research* curricular activities) for trainee development based on the literature, then adapted and created items that aligned with those meta–learning objectives to assess trainee development. These meta–learning objectives were eventually organized into what is now the *Entering Research* conceptual framework. Our goal in creating this conceptual framework and its accompanying instrument was to create an assessment tool that could measure several different dimensions of trainee development identified by prior research as important to research trainee success. In addition, we aimed to develop a tool that could capture trainees' learning and growth throughout the undergraduate and graduate trainee experiences. Finally, we sought to begin to address the limitation of existing instruments that rely solely on self-reported data by creating an instrument that could be used with mentor and trainee pairs to assess trainee skills and examine the effect of alignment, or similarity of the scores.

DEVELOPMENT OF THE *ENTERING RESEARCH* LEARNING ASSESSMENT

The creation of the ERLA occurred over a four-stage, iterative process of development, testing, and refinement (Figure 1). In stage 1, we defined meta–learning objectives and identified, adapted, and developed items to align with them, providing validity evidence for test content. In stage 2, we took the items developed in stage 1 and pilot tested them with research trainees and mentors, examining respondent feedback to identify problematic items and used exploratory factor analysis (EFA) to determine whether the items aligned with our intended organization of the meta–learning objectives into a conceptual framework; this provided initial validity evidence for internal structure. Items were then refined and the instrument expanded in stage 3 before we conducted a confirmatory factor analysis and alignment analyses with a new sample of trainees and their mentors in stage 4, which provided additional validity evidence for internal structure along with convergent validity evidence and evidence of internal consistency. We present the methods and results for each stage below.

Stage 1: Item Development

The ERLA was developed in conjunction with the *Entering Research* curriculum and conceptual framework (Branchaw *et al.*, 2020) using a backward design approach. Learning

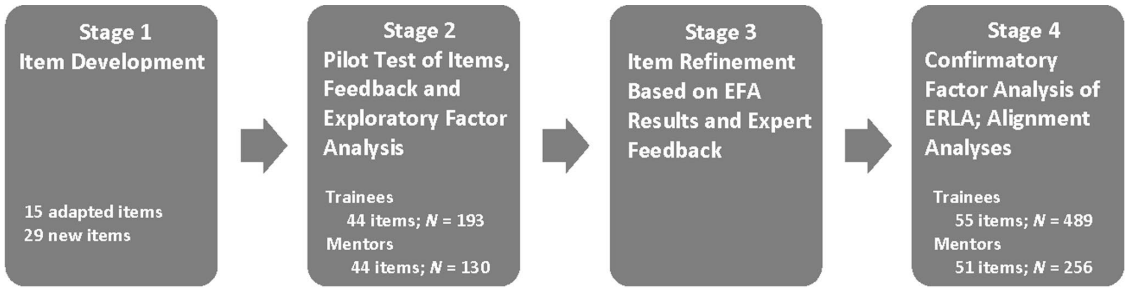

FIGURE 1. Development of *Entering Research* Learning Assessment (ERLA).

objectives were first defined, then assessments and activities were developed to align with the objectives (Wiggins *et al.*, 1998). In this section, we describe the process of defining the meta–learning objectives, including the literature that informs them; the development of the ERLA items; and the organization of the meta–learning objectives into areas of trainee development as a conceptual framework.

Stage 1A: Defining the Entering Research Meta–Learning Objectives. As we began to restructure the *Entering Research* curriculum from a linear, two-semester curriculum in the first edition to a more flexible collection of activities organized by meta–learning objectives in the second edition (Branchaw *et al.*, 2020), we looked to both practitioners and the literature to help us define the meta–learning objectives that should frame our curriculum. An initial list of meta–learning objectives was developed based on the content in the first edition of the curriculum, the literature on STEMM research trainee development and persistence, and the recommendations for research program design in two recent National Academies reports, *Undergraduate Research Experiences for STEM Students* and *Graduate STEM Education for the 21st Century* (NASEM, 2017, 2018; Table 2). These reports were written by committee members with expertise in education, STEMM research training, and institutional change. We asked a team of 17 STEMM and social science researchers and practitioners who were adapting the first edition activities and authoring new activities for the second edition of the curriculum to provide feedback on the initial list of objectives and to suggest additional objectives that they thought were addressed by the curriculum. Their feedback led to the generation of a final set of meta–learning objectives that are addressed in the *Entering Research* curriculum and are aligned with factors shown to be important in undergraduate and graduate research trainee persistence and success in STEMM (Table 2).

Stage 1B: Development of ERLA Items, Alignment with Meta–Learning Objectives, and Development of Response Scale. Once the meta–learning objectives were articulated, we developed new and adapted existing items (e.g., from URSSA, MCA). To provide evidence of validity based on test content, we aligned items with specific meta–learning objectives articulated in stage 1A. In all but one case, multiple items were developed for each meta–learning objective; for the meta–learning objective Develop Research Leadership and Mentoring Skills, only one item was developed ("Mentor others learning to do research"). One item from the URSSA (Weston and Laursen, 2015) Thinking and Working Like a Scientist subscale, "Understand the theory and concepts guiding their research project," was incorporated without adaptation, and three other items in this subscale were adapted. In addition, one item from the URSSA Skills and two items from the Personal Gains subscales were adapted. Six items were adapted from the trainee version of the MCA (Fleming *et al.*, 2013). Two items relating to researcher identity were adapted from the Scientific Identity scale developed by Estrada *et al.* (2011). Twenty-nine new items were developed by the three authors of the curriculum (Branchaw *et al.*, 2020). There were 44 items included in the first version of the ERLA (see Table 4 later in the article).

We chose to use a gains-based response scale similar to the one used on the URSSA rather than an assessment of absolute skill level. Therefore, ERLA responses represent trainee learning gains relative to the knowledge and experience they had when beginning the research experience. We used this approach because, as other researchers have noted (e.g., Weston and Laursen, 2015), trainees, especially novice trainees, have difficulty realistically assessing their absolute pre skill level on assessments that are administered as pre- and postsurveys.

Stage 1C. Organization of Meta–Learning Objectives into the Entering Research Conceptual Framework. With one exception, all meta–learning objectives developed during stage 1A were retained with minor revisions throughout instrument development. These meta–learning objectives were initially organized around five areas of trainee development (see Table 4 later in the article): Research Skills, Interpersonal Skills, Research Attitudes and Beliefs, Equity and Inclusion Awareness and Skills, and Professional and Career Development Skills. However, as data were collected during pilot testing (stage 2), this organization evolved. Three separate areas of trainee development emerged from the Research Skills and Interpersonal Skills areas: Research Comprehension and Communication Skills, Practical Research Skills, and Research Ethics. Likewise, two independent areas of trainee development emerged from the Research Attitudes and Beliefs area: Researcher Identify and Researcher Confidence and Independence (Supplemental Figure S1).

Stage 2: Pilot Test of Items, Feedback, and EFA

Participants. Participants were undergraduate trainees and their mentors (Table 3) who completed the ERLA as part of either a research symposium evaluation survey administered in Spring 2017 at a large research university in the Midwest or as part of the evaluation of activities pilot tested from the second

TABLE 2. *Entering Research* meta–learning objectives aligned to supporting research and National Academies recommendations

Entering Research meta–learning objective	Foundational research findings and alignment with recommendations from National Academies reports (NASEM 2017; 2018)
Develop Disciplinary Knowledge: Ability to understand the theory, content and concepts that inform research in a given discipline	• Develop disciplinary knowledge/ specialized expertise (Craney *et al*. 2011; Kardash, 2000; NASEM, 2017, 2018). • Understand the role of theory in research (Russell *et al.*, 2007). • "Develop STEM literacy" (NASEM, 2017, p. 71). • "Utilize disciplinary research practices" (NASEM, 2017, p. 71).
Develop Technical Research Skills: Ability to design and conduct research	• Develop data-collection, data analysis, interpretation, and problem-solving skills (Kardash, 2000; Bauer and Bennett, 2003; Lopatto, 2004; Russell *et al.*, 2007; Junge *et al.*, 2010; Gilmore *et al.*, 2015; NASEM, 2018). • Knowledge of experimental design and the research process (Russell *et al.* 2007; Thiry *et al.* 2011; NASEM, 2017, p. 109). • "Develop [research] skills/techniques" (NASEM, 2017, p. 71). • "Design a research strategy, including relevant quantitative, analytical, or theoretical approaches" (NASEM, 2018, p. 106).
Develop Research Communication Skills: Ability to communicate research to different audiences and in different formats (e.g., oral, written)	• Increase communication skills (Craney *et al.*, 2011; Junge *et al.*, 2010; Kardash, 2000). • "Obtain, evaluate, and communicate information" (NASEM, 2017, p. 71). • "Acquire the capacity to communicate, both orally and in written form, the significance and impact of a study or a body of work to all STEM professionals, other sectors that may utilize the results, and the public at large" (NASEM, 2018, p. 107).
Develop Logical/Critical Thinking Skills: Ability to think logically and critically about one's own research and the research of others	• Ability to contextualize research in the broader field (Kardash, 2000; Craney *et al.*, 2011; Gilmore *et al.*, 2015). • "Engage in argumentation from evidence" (NASEM, 2017, p. 71). • "Analyze and interpret data" (NASEM, 2017, p. 71). • Understand disciplinary research practices (NASEM, 2017). • "Evaluate outcomes of each experiment or study component and select which outcomes to pursue and how to do so through an iterative process" (NASEM, 2018, p. 106).
Develop Understanding of the Research Environment: Understand research culture and norms; develop skills to successfully navigate research environment	• Understand the nature of science and research-related work; engage in authentic research experiences (Harsh *et al.*, 2011). • "Know importance of iteration" (NASEM, 2017, p. 71). • "Learn and apply professional norms and practices of the scientific or engineering enterprise" (NASEM, 2018, p. 107).
Develop Effective Interpersonal Communication Skills: Skills to communicate clearly with research mentors and research team members and to ask clarifying questions to increase understanding of research	• Increase professional interpersonal communication skills (Kardash, 2000; Hunter *et al.*, 2007; Junge *et al.*, 2010; Laursen *et al.*, 2010; Craney *et al.*, 2011; Gilmore *et al.*, 2015; Carter *et al.*, 2016). • "Act professionally" (NASEM, 2017, p. 71). • "Develop professional competencies, such as interpersonal communication, budgeting, project management, or pedagogical skills that are needed to plan and implement research projects" (NASEM, 2018, p. 107).
Develop Responsible and Ethical Research Practices: Understand what constitutes ethical research practices	• Integrating ethics training that encourages active trainee participation leads to gains in ethics, knowledge, skills, and attitudes (Watts *et al.*, 2017). • Awareness of "the importance of ethics and responsible conduct" (NASEM, 2017, p. 40). • "Learn and apply … ethical responsibilities of scientists and engineers within the profession and in relationship to the rest of society, as well as ethical standards that will lead to principled character and conduct" (NASEM, 2018, p. 107).
Develop Research Leadership and Mentoring Skills: Prepare to take on increasing levels of leadership and mentorship	• Develop research leadership, collaboration, and management skills (Coker and Van Dyke, 2005; Mancha and Yoder, 2014; NASEM, 2018). • Leadership opportunities should be a component of undergraduate chemistry curriculum, including research experiences (Wenzel *et al.*, 2012). • Graduate students often serve as mentors for undergraduate trainees (NASEM, 2017).
Develop Identity as a Researcher: Think about oneself as a researcher	• Research experiences are a key contributor to science identity (Seymour *et al.*, 2004; Hunter *et al.*, 2007; Kardash and Edwards, 2012). • Science identity contributes to persistence in STEM (Barker, 2009; Estrada *et al.*, 2011; Villa *et al.*, 2013; Raelin *et al.*, 2014). • "Promote agency and develop STEM identity" (NASEM, 2017, p. 71).
Develop Independence as a Researcher: Ability to work independently and exhibit leadership at a level appropriate to one's training	• Self-efficacy and project ownership contribute to research independence (Adedokun *et al.*, 2013; Hanauer and Dolan, 2014). • "Increase ownership of project" (NASEM, 2017, p. 71). • "Develop professional competencies, such as interpersonal communication, budgeting, project management, or pedagogical skills that are needed to plan and implement research projects" (NASEM, 2018, p. 107).

(*Continues*)

TABLE 2. Continued

Entering Research meta–learning objective	Foundational research findings and alignment with recommendations from National Academies reports (NASEM 2017; 2018)
Develop Confidence as a Researcher: Confidence in one's ability to successfully conduct research	• Develop confidence and self-efficacy in research (Byars-Winston *et al.*, 2015; Chemers *et al.*, 2011; Estrada *et al.* 2011; Hunter *et al.*, 2007; Hurtado *et al.*, 2009 John and Creighton, 2012; Russell *et al.*, 2007; Seymour *et al.*, 2004). • Student confidence/self-efficacy (NASEM, 2017).
Advance Equity and Inclusion in the Research Environment: Recognize and mitigate stereotype threat, bias, and microaggressions; acknowledge intersection of personal and research identities	• Research programs that provide support for trainees from underrepresented groups increase persistence and success (Hathaway *et al.*, 2002; Mendoza-Denton *et al.*, 2002; Lee and Davis, 2000; Woodcock *et al.*, 2012). • "Develop a sense of belonging/inclusion" (NASEM, 2017, p. 71). • Implement "practices that create an equitable and inclusive institutional environment" (NASEM, 2018, p. 9).
Develop Skills to Deal with Personal Differences in the Research Environment: Build capacity to effectively engage with individuals from different backgrounds and identities to maximize the benefits of an increasingly diverse STEM workforce	• Heterogeneous research teams produce more innovative and effective solutions and products (Ferrini-Mundy, 2013; Page, 2008; Reagans and Zuckerman, 2001; Roberge and Van Dick, 2010; Saxena, 2014; Woolley *et al.*, 2010). • "Perform work as collaborative member of team" (NASEM, 2017, p. 71). • "Recognize and overcome stereotype threat" (NASEM, 2017, p. 71). • Work collaboratively with individuals from "diverse cultural and disciplinary backgrounds" (NASEM, 2018, p. 107; NASEM, 2017).
Explore and Pursue a Research Career: Explore career pathways and create professional development plans	• Awareness of what graduate school is like (Russell *et al.*, 2007). • Mentors who provide career and graduate school guidance (Carpi *et al.*, 2017). • Confirmation/clarification of career path (NASEM, 2017, p. 71; Russell *et al.*, 2007); exposure to research careers and career expectations (Hurtado *et al.*, 2009). • Retention in STEM major/commitment to the discipline (NASEM, 2017, p. 71). • Explore STEM career opportunities and pathways (NASEM, 2018).
Develop Confidence in Pursuing a Research Career: Confidence in one's ability to pursue a STEM research career and the next steps in one's training.	• Intention to pursue a PhD (Russell *et al.*, 2007). • Research experiences help trainees "solidify their career plans as research scientists" (Hurtado *et al.*, 2009, p. 211). • "Confirmation/clarification of career path" (NASEM, 2017, p. 71; Russell *et al.*, 2007). • Retention in STEM major/commitment to the discipline (NASEM, 2017, p. 71). • Explore STEM career opportunities and pathways (NASEM, 2018).

edition of *Entering Research* in Summer 2017 at four sites across the United States (2 West; 1 Midwest; 1 Northeast; Institutional Review Board [IRB] protocol 2017-0026).

Procedure. The original prompt for this instrument was "As a result of your research experience, indicate how much you gained in your ability to …" for trainees and "How much did your trainee GAIN in their ability to do the following as a result of their research experience?" for mentors. Response options were: "no gain," "a little gain," "moderate gain," "good gain," "great gain," and "not applicable." A subset of the trainee sample ($n = 142$) and all mentors in our sample ($n = 130$) also had the option to indicate that "this question was unclear." Individuals who indicated that a question was not applicable or was unclear were asked to provide information on why the statement was not applicable or was unclear. These open-ended responses were reviewed in combination with the results of the item analysis and EFA to identify potentially problematic items that needed modification.

Analyses. Both an EFA and item-level analysis were conducted as part of this pilot-testing stage. The EFA was conducted using SPSS v. 23 (IBM SPSS Statistics for Windows, 2015). Because we anticipated that our factors would correlate, we analyzed the data using principal axis factor with an oblique rotation. The data were treated as interval for this analysis, so any selections of "not applicable" or "this question was unclear" were excluded from analysis. To maximize the data available from individuals who completed most items but who may have chosen "not applicable" or "this question was unclear" as a response option, we chose pairwise deletion. After the factor structure was revealed, we examined internal consistency statistics (Cronbach's alpha) for each of the subscales.

The inflexion point of the scree plot and eigenvalues were examined to determine how many factors were appropriate to retain (Cattell, 1966). We did not specify an eigenvalue cutoff, but instead examined the data relative to the *Entering Research* conceptual framework to determine whether the number of factors extracted and the ways in which the items loaded on particular factors were consistent with the intended framework (Knekta *et al.*, 2019). Factor loadings were examined using the pattern matrix; any items with factor loadings lower than 0.4 were flagged for further review to determine whether items should be removed or revised. We expected to find a five-factor solution wherein all items aligned with the original *Entering Research* conceptual framework. However, any additional factors that emerged were further analyzed to determine whether they indeed represented a new factor or were simply an artificial factor that could not be reasonably represented in the framework. Individual items were assessed based on several criteria, including the extent to which participants noted items were not applicable or unclear and interitem correlations. The results of these analyses were examined holistically to determine whether individual items should be revised to better align with the intended framework or removed from the instrument.

TABLE 3. Demographic information for stage 2 sample

	Trainees (N = 193)	Mentors (N = 130)
Gender		
Female	48%	62%
Male	23%	33%
Other gender identities	<1%	—
Not reported	28%	5%
Race/ethnicity[a]		
White	52%	63%
Asian	11%	15%
African American	3%	5%
Native Hawaiian/Pacific Islander	—	<1%
Native American	<1%	—
Two or more races	2%	5%
Unknown	2%	—
Not reported	29%	11%
Hispanic	7%	9%

[a]Respondents could select Hispanic in addition to a race category. As a result, total percentages for the sample may add up to more than 100%.

Results. Preliminary validity evidence for internal structure and evidence of internal consistency were provided via responses from trainees and mentors. Feedback from respondents along with the results of the EFA led to the removal of three items and the revision of eight items before expert feedback on the scale was solicited as part of stage 3 (Supplemental Table S1). A brief overview of validity evidence reviewed during stage 2 is provided later.

Results for Trainee Sample. The Kaiser-Meyer-Olkin measure of sampling adequacy revealed a value of 0.932, indicating that our sample size was sufficient to reveal reliable factors (Hutcheson and Sofroniou, 1999). The initial EFA extracted seven factors that accounted for 67% of the variance (Table 4). Factor 1 (Research Comprehension and Communication Skills) accounted for 46% of the variance, factor 2 (Equity and Inclusion Awareness and Skills) for 7%, factor 3 (Professional and Career Development Skills) for 5%, factor 4 (Researcher Confidence and Independence) for 3%, factor 5 (Research Ethics) for 3%, and factors 6 (Practical Research Skills) and 7 (Researcher Identity) each accounted for 2%. The items retained from the

TABLE 4. *Entering Research* Learning Assessment (stages 1 and 2): Original items and results of Exploratory Factor Analysis

Item	RCC	PRS	RE	RID	RCI	EIA	PDS
Research Skills							
1. Understand the theory and concepts guiding your research project.[b]	**0.413**	0.181	0.088	0.027	0.210	0.011	0.088
2. *Connect your research experience to what you have learned in courses.*[b]	0.123	0.260	0.186	0.187	–0.214	0.114	0.025
3. Communicate the context, methods, and results of your research.	**0.567**	0.210	0.023	0.238	0.187	0.026	0.083
4. Tailor your research communications for different audiences (e.g., general public, disciplinary conference)	**0.373**	0.159	0.106	0.173	0.133	0.271	0.004
5. Identify forms of unethical practices or research misconduct.	0.075	0.071	**0.858**	0.082	0.089	0.068	0.057
6. Understand the consequences of unethical practices or research misconduct.	0.007	0.084	**0.913**	0.032	0.017	0.010	0.063
7. Take action to address unethical practices or research misconduct.	0.178	0.016	**0.525**	0.123	0.124	0.379	0.017
8. Use logic and evidence to interpret data.	**0.681**	0.156	0.212	0.111	0.116	0.002	0.062
9. Use logic and evidence to build arguments and draw conclusions from data.	**0.548**	0.082	0.161	0.144	0.156	0.056	0.060
10. Make connections between your research and societal issues.	**0.280**	0.125	0.117	0.000	0.090	0.211	0.004
11. Design and conduct a research project.	0.085	**0.500**	0.185	0.077	0.304	0.002	0.082
12. Keep detailed research records (e.g., a lab/field notebook).[c]	0.204	**0.494**	0.020	0.197	0.032	0.122	0.001
13. Analyze data.[b]	**0.610**	0.005	0.007	0.150	0.125	0.013	0.030
14. Work in the research environment comfortably.	**0.611**	0.246	0.135	0.126	0.004	0.054	0.162
15. *Be yourself when working in the research environment.*	**0.619**	0.232	0.063	0.123	0.106	0.164	0.051
16. Accept and use criticism of your research to improve your research.	**0.534**	0.227	0.064	0.007	0.073	0.099	0.136
17. Understand that the process of discovery is iterative and never-ending.	**0.758**	0.060	0.075	0.064	0.196	0.008	0.187

(*Continues*)

TABLE 4. Continued

Item	RCC	PRS	RE	RID	RCI	EIA	PDS
Interpersonal skills							
18. Listen for understanding and comprehension regarding your research project.	**0.476**	0.056	0.036	0.022	0.285	0.203	0.025
19. Ask questions to clarify your understanding of your research project.	**0.523**	0.014	0.106	0.136	0.229	0.092	0.190
20. Align your research experience goals and expectations with those of your research mentor.[d]	**0.545**	0.105	0.097	0.016	0.212	0.051	0.130
21. Practice regular and open communication with your mentor.	**0.661**	0.113	0.097	0.144	0.064	0.033	0.225
22. Practice regular and open communication with research team members.	**0.626**	0.019	0.065	0.024	0.068	0.154	0.027
23. *Mentor others learning to do research.*	0.163	0.105	0.019	0.025	0.319	0.427	0.072
Researcher Attitudes and Beliefs							
24. Think of yourself as a scientist/researcher.[e]	0.186	0.069	0.198	**0.603**	0.021	0.039	0.017
25. Feel like you belong in research.[e]	0.126	0.058	0.047	**0.705**	0.023	0.066	0.169
26. Call yourself a researcher when talking to others.	0.122	0.054	0.053	**0.717**	0.003	0.052	0.204
27. Work independently on your research project.[f]	0.048	0.215	0.015	0.151	**0.443**	0.077	0.187
28. Determine the next steps in your research project.[b]	0.040	0.336	0.022	0.001	**0.606**	0.006	0.196
29. Investigate and solve problems when they arise in your research (e.g., troubleshoot).	0.109	0.035	0.009	0.048	**0.742**	0.028	0.150
30. Be confident in conducting research.[f]	0.050	0.024	0.070	0.149	**0.589**	0.005	0.169
31. Be confident in coping with challenges when they arise in your research project.	0.125	0.024	0.153	0.107	**0.629**	0.048	0.062
32. Be confident in staying motivated and committed to your research project when things do not go as planned.	0.101	0.082	0.092	0.202	**0.493**	0.099	0.028
33.Be confident in completing your research training.	0.228	0.000	0.202	0.116	**0.427**	0.086	0.117
34. Be confident in pursuing a career in research.	0.068	0.004	0.035	0.481	0.091	0.039	**0.501**
Equity and Inclusion Awareness and Skills							
35. Identify the biases and prejudices that you have about others.[d]	0.038	0.123	0.014	0.029	0.069	**0.837**	0.085
36. Identify the biases and prejudices that others may have about you.[d]	0.014	0.093	0.083	0.072	0.105	**0.848**	0.120
37. Understand the impact of biases on your interactions with others in a research environment.	0.019	0.018	0.061	0.078	0.064	**0.918**	0.013
38. Work effectively with others in a research environment whose personal backgrounds are different from your own.[d]	0.204	0.091	0.187	0.053	0.019	**0.458**	0.036
39. Understand how others might experience research differently based on their identity (e.g., race, socioeconomic status, first-generation status)	0.007	0.053	0.140	0.076	0.127	**0.727**	0.013
40. Advocate for others who may be marginalized or excluded from the research environment.	0.055	0.008	0.056	0.078	0.133	**0.680**	0.031
Professional and Career Development Skills							
41. Explore possible research career pathways.	0.135	0.059	0.025	0.146	0.018	0.118	**0.659**
42. Set research career goals.[d]	0.071	0.075	0.100	0.082	0.098	0.103	**0.709**
43. Develop a plan to pursue a research career (determine the next step in their training).	0.005	0.008	0.077	0.094	0.072	0.030	**0.755**
44. Meet and establish relationships with research professionals in their field (network).[d]	0.066	0.052	0.074	0.033	0.183	0.083	**0.655**

[a]Question stem for each item was "As a result of your research experience, indicate how much you gained in your ability to..." for trainees and "How much did your mentee gain in their ability to do the following as a result of their research experience?" for mentors. The areas of trainee development from the Entering Research conceptual framework are listed in bold. Items appearing in italics were removed from the assessment at Stage 2. Adapted items are noted by superscripts: b, URSSA (Thinking and Working Like a Scientist); c, URSSA (Skills); d, MCA (Trainee adaptation); e, Estrada *et al.* (2011); f, URSSA (Personal Gains). RCC, Research Comprehension and Communication Skills; PRS, Practical Research Skills; RE, Research Ethics; RID, Researcher Identity; RCI, Researcher Confidence and Independence; EIA, Equity and Inclusion Awareness and Skills; PDS, Professional and Career Development Skills.

first iteration of the instrument provided preliminary evidence of internal structure and internal consistency with our trainee sample (factor loadings 0.28–92; $\alpha = 0.74 – 0.95$). An examination of inter-item correlations revealed that items correlated significantly and positively with one another, and that the items pertaining to each of the seven factors revealed in the EFA correlated with one another.

The frequency with which trainees selected "not applicable" or "this question was unclear" was also assessed, with most items having fewer than 10% of respondents indicating that an item was not applicable or unclear. The exception was one item from the research ethics subscale, "take action to address unethical practices or research misconduct," for which 15% of trainees indicated that the item was not applicable. When looking at respondent feedback to this item, many individuals noted that they had selected this option because they had not had to address unethical practices in the course of their research experience. We decided to retain this item, because it is applicable to the research experience and this topic may be addressed in training experiences or in training program seminars.

Revision of Entering Research Conceptual Framework and Refinement of Meta–Learning Objectives. Two factors emerged from items originally placed in the Research Skills and Interpersonal Skills subscales (Supplemental Figure S1). Many of the items in the Research Skills subscale and all of the items in the Interpersonal Skills subscale loaded onto one factor, which we renamed Research Comprehension and Communication Skills. A second factor, Practical Research Skills, also emerged from the Research Skills subscale. We retained this factor, which aligned with the learning objective Develop Technical Research Skills and subsequently split this meta–learning objective into two separate meta–learning objectives, Develop Ability to Design a Research Project and Develop Ability to Conduct a Research Project.

Two factors emerged from the items originally included under Research Attitudes and Beliefs: Researcher Identity and Researcher Confidence and Independence. One item from this subscale, "Be confident in pursuing a career in research," loaded onto the Professional and Career Development Skills factor. We decided that this loading was appropriate and retained that item in the Professional and Career Development subscale during the next stage of development. The original meta–learning objective, Develop Confidence as a Researcher and in Pursuing a Research Career, with which this item and other confidence items were aligned, was revised into two separate meta–learning objectives: Develop Confidence as a Researcher and Develop Confidence to Pursue a Research Career. These objectives were aligned with Researcher Confidence and Independence and Professional and Career Development Skills, respectively.

Results for Mentor Sample. Low sample size and high numbers of mentors indicating some items were "not applicable" or that items were unclear resulted in a mentor sample that was too small to perform an EFA. An examination of inter-item correlations for the mentor items revealed similar results to those found with the trainee data, though correlation coefficients were often smaller and, in one case, not statistically significant. The item "make connections between your research and societal issues," was not significantly correlated or had low correlation coefficients with several of the items in the Research Comprehension and Communication Skills scale, lending further support to our decision to revise that item. The frequency with which mentors selected "not applicable" or "this question was unclear" was higher compared with trainees, especially in items pertaining to research ethics (17–39% of respondents indicating items were not applicable) and equity and inclusion awareness (13–26% of respondents indicating items were not applicable). Mentors commented that items were confusing and/or difficult to answer (e.g., "I can't speak for my trainee") and consequently did not rate their trainees' gains. This was particularly true for items relating to equity and inclusion. In some cases, mentors expressed frustration with these items (e.g., "Doesn't matter, questions like this promote biases and prejudice"; "It's about the science, not about them"). Three items from this subscale were flagged for further review based on mentor feedback that these items were difficult to assess.

Stage 3: Item Refinement Based on EFA Results and Expert Feedback

Item Refinement. With the revised framework in place, we reexamined items with lower factor loadings to determine whether the items could be revised or clarified. All items were evaluated to ensure that the structure of each item following the question stem was consistent. Assessment of item factor loadings, participant feedback on items, and the revised conceptual framework led to the revision of some original items (Supplemental Table S1). The revised instrument included 13 additional items, 10 of which were created to assess the Practical Research Skills factor that emerged during the EFA, because only two original items loaded on this factor during stage 2. Two of these items were adapted from the URSSA; one from the Skills subscale and one from the Thinking and Working like a Scientist subscale. One of the research identity items ("Behave like a researcher"), which was originally intended to partner with the trainee item "Think like a scientist/researcher") was revised to "Behave like a researcher in your discipline," and an identical item was developed and added to the trainee survey. Two additional items were added to the Researcher Identity subscale: "Fit in with the research culture of your discipline" and "Fit in with the culture of your research group."

Item Stem and Response Scale Refinement. We also made some refinements to the item stem and response scale. In subsequent iterations of the ERLA, we changed the prompt for the item stem to "How much did you gain in your ability to do the following over the course of your research experience?," to acknowledge that research learning also occurs outside actually conducting research (e.g., participation in a course based on *Entering Research*) and can impact trainee learning gains. The response option "not applicable" was removed for both trainees and mentors to reflect our assumption that all of the items included in the final instrument would apply to research experiences across STEMM disciplines. For mentors, we added the response option "did not observe," to allow mentors to share if they did not have the opportunity to observe the trainee engaged in a particular skill. These responses were combined with "no gain" as part of the scoring process. We chose to combine these two options in the scoring process,

because all of the response options for mentors are based on their observations. If a mentor did not observe a particular behavior or skill for a trainee, then a response of "did not observe" is equivalent to no observable gain (i.e., "no gain"). Though these types of responses (e.g., "don't know," "unsure") are often treated as missing by researchers (Schafer and Graham, 2002), these responses have been treated in many different ways, including recoding a response into a different category (Denman *et al.*, 2018). Researchers have noted that decisions on the scoring of such responses must be done "in the context of a specific instrument" (DeMars and Erwin, 2004, p. 87). With this in mind, we chose to incorporate the "did not observe" responses into mentors' scores, because differences between mentor and trainee assessments of trainees' skills provide important information about the mentored research experience that program directors may find useful for program evaluation. We later explore how the factor structure of the ERLA is affected when mentors' responses of "did not observe" are treated as missing.

Expert Review. Once these initial revisions based on the EFA and mentor and trainee feedback were incorporated, we asked two individuals from different STEM disciplines (chemistry and engineering) with experience facilitating *Entering Research* and working with undergraduate and graduate research trainees to review the items and provide feedback. Feedback from these experts led to additional refinements. The three items under the Equity and Inclusion Awareness and Skills subscale that were initially flagged for review in stage 2 were removed in stage 3 due to expert feedback that aligned with the feedback that we had received previously from mentors indicating that they could not accurately assess their trainees' gains on these items. These experts provided additional suggestions for revised wording and suggested an additional item, "Make detailed observations," which was incorporated into the Practical Research Skills subscale. Final edits based on their feedback yielded a trainee survey with 55 items and a mentor survey with 51 items.

Stage 4: Confirmatory Factor Analysis of ERLA Scale and Alignment Analyses

Participants. Novel participants for stage 4 (Table 5) were recruited as part of their participation in a mentored research experience at a large research university in the midwestern United States (IRB protocol 2018-0312) and through snowball sampling via emails sent to undergraduate and graduate research programs and centers across the United States requesting that they forward the survey on to their trainees and their mentors for completion (IRB protocol 2017-0026). Twenty-four percent (24%) of trainees were first- or second-year undergraduates; 47% were undergraduates in their third year or beyond; 13% were postbaccalaureate students or graduate students in their first or second year; and 15% were graduate students in their third year of training or beyond; 2% did not report their training stages. One trainee completed the assessment more than once to assess multiple mentors. These two surveys were treated as individual observations to examine validity evidence for the instrument, as trainees were asked to complete each survey keeping in mind a specific mentor. This led to a total sample size of 490.

TABLE 5. Demographic information for stage 4 sample

	Trainees (N = 489)	Mentors (N = 256)
Gender		
Female	65%	52%
Male	31%	45%
Other gender identities	1%	<1%
Not reported	2%	4%
Race/ethnicity[a]		
White	66%	77%
Asian	19%	12%
African American	4%	<1%
Native Hawaiian/Pacific Islander	—	—
Native American	1%	<1%
More than one race	6%	3%
Unknown	1%	1%
Not reported	5%	6%
Hispanic	8%	5%

[a]Respondents could select Hispanic in addition to a race category. As a result, total percentages for the sample may add up to more than 100%.

The majority of mentors (96%) were mentors of undergraduate students. Some mentors completed the assessment more than once because they were assessing multiple trainees. Each survey response was treated as an observation to examine validity evidence for the instrument, as mentors were asked to complete each survey keeping in mind a specific trainee. This led to a total sample size of 309.

Measures. All questions were administered to participants via an online survey. The survey included the revised version of the ERLA (55 items for trainees and 51 items for mentors) and questions asking participants to rate the research experience and the overall quality of their mentoring relationship. Due to a survey error, the majority of mentors in our sample did not have the opportunity to rate the research experience, resulting in a lower sample size for this item (N = 182).

Analyses. We examined responses to the ERLA for missing data to determine whether our data met the criteria for ignorable missingness (i.e., data missing completely at random [MCAR] or missing at random) as this has implications for subsequent treatment and analysis of data (Myers *et al.*, 2013). Little's MCAR test (1988) revealed that data were missing completely at random in the mentor sample. The data were not found to be missing completely at random in the trainee sample; however, the percentage of missing data for all variables was less than 5%, indicating that the missing data could be ignored (Tabachnick and Fidell, 2007). An examination of the mentor and trainee data revealed that the data were not univariately normally distributed. Due to the Likert-type response scale, it was best to treat the data as ordinal. The Weighted Least Mean Square and Variance-Adjusted estimator available in MPLUS (Muthén and Muthén, 1998–2017) is appropriate for this type of data (Knekta *et al.*, 2019) and was used to conduct the confirmatory factor analysis.

When using the ERLA to conduct research, researchers who do not wish to factor "did not observe" mentor responses into scale scores may want to treat these responses as missing data.

To confirm that the model identified in our previous analysis would fit the data if the "did not observe" responses were treated as missing, we ran a separate confirmatory factor analysis. When treated as missing, the mentor sample data were not found to be missing completely at random based on Little's MCAR test (1988). Because the percentage of missing data was greater than 5% (13.04% for all variables), multiple imputation was conducted in MPLUS using maximum likelihood estimation to replace any missing values due to responses of "did not observe" or nonresponse. Fifty imputed data sets were created based on the recommendations by Graham *et al.* (2007) that more imputations will increase statistical power. These data sets were used to conduct a confirmatory factor analysis using the approach outlined above.

For examination of evidence of convergent validity, correlations between each of the ERLA subscales and ratings of the research experience and quality of the mentoring relationship were examined for trainees and mentors, respectively. Several fit statistics noted in the *Introduction* of this paper were used to examine model–data fit (i.e., lower χ^2 values; χ^2/df ratio between 2 and 3; CFI > 0.95; and RMSEA <0.06; Hu and Bentler, 1999; Schreiber *et al.*, 2006).

To examine whether the trainee and mentor versions of the instrument could be used together to assess the degree to which trainee self-assessments of skill gain and mentor assessments of trainee skill gain align, we first examined validity evidence for internal structure for a version of the trainee instrument that aligned with items presented on the mentor instrument. Internal structure validity evidence was assessed using the model–data fit statistics and criteria outlined earlier. We next examined the descriptive statistics for trainee and mentor scores by calculating the means and standard deviations for a paired subset of our larger sample ($n = 121$) and examined the extent to which trainees and mentors aligned on their assessments by calculating difference scores and Spearman correlations for each subscale.

Results

Validity Evidence Based on Internal Structure and Evidence of Internal Consistency. Initial fit statistics for the 55 trainee items fit to a seven-factor model were $\chi^2(1409) = 3920.362$, $p < 0.001$; $\chi^2/df = 2.78$; RMSEA = 0.060, 90% CI [0.058, 0.063]; CFI = 0.949. In addition to the confirmatory factor analysis model-fit statistics, model modification indices, internal consistency statistics (i.e., Cronbach's alpha), and item-total correlations were examined to identify any problematic items. Modification indices suggested that the fit statistics for the model would significantly improve with the removal of one item in the Equity and Inclusion Awareness and Skills subscale, "Work effectively in a research environment with individuals whose personal backgrounds are different from your own," and one item in the Professional and Career Development Skills subscale, "Meet and establish relationships with research professionals in your field (network)." In addition, the subscale internal consistency analyses revealed lower item-total correlations for the items and a lower overall alpha value when these items were retained in the subscale. For these reasons, the items were removed, resulting in final model-fit statistics of $\chi^2(1304) = 3333.766$, $p < 0.001$; $\chi^2/df = 2.56$; RMSEA = 0.056, 90% CI [0.054, 0.059]; CFI = 0.957.

Initial fit statistics for the 51 mentor items fit to a seven-factor model were $\chi^2(1203) = 2829.838$, $p < 0.001$; $\chi^2/df = 2.35$; RMSEA = 0.066, 90%CI [0.063, 0.069]; CFI = 0.938. As was done with the trainee data, model modification indices, internal consistency statistics (i.e., Cronbach's alpha), and item-total correlations were examined to identify any problematic items. Modification indices suggested that the fit statistics for the model would improve with the removal of the Equity and Inclusion Awareness and Skills item "Work effectively in a research environment with individuals whose personal backgrounds are different from your own," one item from the Researcher Identity subscale ("Call themselves a researcher when talking to others"), and one item from the Practical Research Skills subscale ("Do experiments"). Further examination of these items revealed that Cronbach's alpha would improve if the items were removed from the Equity and Inclusion Awareness and Skills and Researcher Identity subscales and that the interitem correlation for the practical research skills item was comparatively low. The Professional and Career Development Skills item removed from the trainee survey was removed here as well to improve the alpha for the subscale. The final fit statistics for the mentor version of the ERLA with these four items removed was $\chi^2(1013) = 2306.844$, $p < 0.001$; $\chi^2/df = 2.28$; RMSEA = 0.064. 90% CI [0.061; 0.068]; CFI = 0.949. Similar fit statistics were found when "did not observe" responses were treated as missing, $\chi^2(1013) = 2370.431$, $p < 0.001$; $\chi^2/df = 2.34$; RMSEA = 0.066; CFI = 0.961. Factor loadings and Cronbach's alpha values for the subscales of the final trainee and mentor surveys are presented in Table 6; item factor loadings of the mentor version of the ERLA with "did not observe" treated as missing are presented in Supplemental Table S2; intercorrelations between subscales of the ERLA are presented in Supplemental Tables S3.1 and S3.2. For trainees, correlations of subscales ranged from 0.643 to 0.970; for mentors, correlations of subscales ranged from 0.493 to 0.975.

Convergent Validity Evidence. Correlations between each of the subscales of the ERLA and trainees' and mentors' perceptions of the research experience and overall quality of the mentoring relationship are presented in Table 7. For trainees and mentors, all seven subscales of the ERLA were significantly and positively related to both overall ratings of the research experience and ratings of the quality of the mentoring relationship.

Alignment Analyses. Research training program directors may use the ERLA with pairs of mentors and trainees to examine trainees' perceived gains compared with their mentors' assessments of the trainees' gains. Therefore, we also examined the fit of a seven-factor model for trainees that included only the items that align with the 47-item mentor version of the instrument. The fit statistics for the aligned scale were $\chi^2(1013) = 2828.499$, $p < 0.001$; $\chi^2/df = 2.79$; RMSEA = 0.060, 90% CI [0.058, 0.063]; CFI = 0.958. The final factor loadings for the aligned version of the trainee scale are presented in Table 6.

To investigate whether alignment of trainee- and mentor-reported gains differed across the different areas of trainee development, we used a subset of paired mentor–trainee responses from our sample ($n = 121$). We examined the average subscale scores for mentors and their trainees, using only the items that were common across both surveys and calculated the extent to

TABLE 6. *Entering Research* Learning Assessment (stage 4): Final items and results of confirmatory factor analysis[a]

Subscale and item number		Item		Internal consistency and factor loadings		
T	M	Trainee (T)	Mentor (M)	T	$T_{aligned}$	M
Research Comprehension and Communication Skills				$\alpha = 0.95$	$\alpha = 0.95$	$\alpha = 0.94$
1	1	Understand the theory and concepts guiding your research project.[b]	Understand the theory and concepts guiding their research project.	0.748	0.747	0.806
25	22	Communicate the context, methods, and results of your research.	Communicate the context, methods, and results of their research.	0.848	0.850	0.851
14	12	Tailor your research communications for different audiences (e.g., general public, disciplinary conference)	Tailor their research communications for different audiences (e.g., general public, disciplinary conference).	0.723	0.720	0.649
24	21	Use logic and evidence to interpret data.	Use logic and evidence to interpret data.	0.852	0.851	0.845
44	39	Use logic and evidence to build arguments and draw conclusions from data.	Use logic and evidence to build arguments and draw conclusions from data.	0.876	0.875	0.904
19	16	Communicate the relevance of your research to others.	Communicate the relevance of their research to others.	0.780	0.782	0.776
6	5	Analyze data.[b]	Analyze data.	0.692	0.690	0.711
28	25	Work in the research environment comfortably.	Work in the research environment comfortably.	0.904	0.904	0.849
45	40	Accept and use criticism of your research to improve your research.	Accept and use criticism of their research to improve their research.	0.858	0.858	0.792
29	26	Understand that the process of discovery is iterative and never ending.	Demonstrate understanding that the process of discovery is iterative and never ending.	0.821	0.819	0.801
13	11	Demonstrate understanding and comprehension regarding your research project.	Demonstrate understanding and comprehension regarding their research project.	0.807	0.810	0.828
9	7	Ask questions to clarify your understanding of your research project.	Ask questions to clarify their understanding of their research project.	0.786	0.787	0.775
41	36	Align your research experience goals and expectations with your research mentor's.[d]	Align their research experience goals and expectations with your goals and expectations.	0.847	0.848	0.792
2	2	Practice regular and open communication with your research mentor.	Practice regular and open communication with you.	0.698	0.700	0.786
34	31	Practice regular and open communication with your research team members.	Practice regular and open communication with your research team members.	0.793	0.792	0.788
Practical Research Skills				$\alpha = 0.92$	$\alpha = 0.92$	$\alpha = 0.91$
10	8	Design a research project.	Design a research project.	0.685	0.683	0.818
18	15	Keep detailed research records (e.g., a lab/field notebook).[c]	Keep detailed research records (e.g., a lab/field notebook).	0.687	0.683	0.674
47	41	Conduct a research project.	Conduct a research project.	0.841	0.839	0.847
15	—	Do experiments.		0.627		
30	27	Collect data.	Collect data.	0.786	0.781	0.682
21	18	Use the tools, materials, and equipment needed to conduct research.	Use the tools, materials, and equipment needed to conduct research.	0.813	0.810	0.787
31	28	Understand the safety precautions relating to your research.	Demonstrate understanding of the safety precautions relating to their research.	0.756	0.748	0.679
37	33	Work effectively with the subject of study (e.g., mathematical models, mice, plans, rock formations).	Work effectively with the subject of study (e.g., chemicals, mathematical models, mice, plants, rock formations).	0.829	0.826	0.814
12	10	Formulate a research question/hypothesis.[b]	Formulate a research question/hypothesis.	0.795	0.794	0.798
33	30	Make a case for your research question based on the literature.	Make a case for their research question based on literature.	0.819	0.818	0.776
17	14	Determine the appropriate experimental approach to investigate your research question.	Determine the appropriate experimental approach to investigate their research question.	0.830	0.825	0.805
49	43	Determine an analysis plan/statistical methods to analyze your data.[c]	Determine an analysis plan/statistical methods to analyze their data.	0.725	0.720	0.796
53	47	Make detailed observations.	Make detailed observations.	0.872	0.871	0.860

(*Continues*)

TABLE 6. Continued

Subscale and item number		Item		Internal consistency and factor loadings		
T	M	Trainee (T)	Mentor (M)	T	$T_{aligned}$	M
Research Ethics				α = 0.86	α = 0.86	α = 0.85
4	3	Identify forms of unethical practices or research misconduct.	Identify forms of unethical practices or research misconduct.	0.814	0.809	0.845
20	17	Understand the consequences of unethical practices or research misconduct.	Demonstrate understanding of the consequences of unethical practices or research misconduct.	0.871	0.879	0.909
42	37	Take action to address unethical practices or research misconduct.	Take action to address unethical practices or research misconduct.	0.890	0.885	0.885
Researcher Identity				α = 0.91	α = 0.87	α = 0.86
3	—	Think of yourself as a scientist/researcher.[e]		0.789		
43	38	Feel like you belong in research.[e]	Act like they belong in research.	0.899	0.883	0.903
38	—	Call yourself a researcher when talking to others.		0.822		
40	35	Behave like a researcher in your discipline.	Behave like a researcher in your discipline.	0.917	0.895	0.898
11	9	Fit in with the research culture of your discipline.	Fit in with the research culture of your discipline.	0.819	0.805	0.875
32	29	Fit in with the culture of your research group.	Fit in with the culture of your research group.	0.813	0.798	0.758
Researcher Confidence and Independence				α = 0.91	α = 0.91	α = 0.92
22	19	Work independently on your research project.[f]	Work independently on their research project.	0.726	0.728	0.755
5	4	Determine the next steps in your research project.[b]	Determine the next steps in their research project.	0.754	0.749	0.797
50	44	Investigate problems when they arise in your research (e.g., troubleshoot).	Investigate problems when they arise in their research (e.g., troubleshoot).	0.841	0.842	0.876
36	32	Confidence in conducting research.[f]	Confidence in conducting research.	0.921	0.920	0.919
48	42	Confidence in coping with challenges when they arise in your research project.	Confidence in coping with challenges when they arise in their research project.	0.873	0.875	0.887
16	13	Confidence in staying motivated and committed to your research project when things do not go as planned.	Confidence in staying motivated and committed to their research project when things do not go as planned.	0.829	0.828	0.809
52	46	Confidence in completing your research training.	Confidence in completing their research training.	0.900	0.902	0.925
Equity and Inclusion Awareness and Skills				α = 0.92	α = 0.81	α = 0.84
7	—	Identify the biases and prejudices that you have about others.[d]		0.835		
35	—	Identify the biases and prejudices that others may have about you.[d]		0.861		
46	—	Understand the impact of biases on your interactions with others in a research environment.		0.972		
51	45	Understand how others might experience research differently based on their identity (e.g., race, socioeconomic status, first-generation status).	Demonstrate understanding of how others might experience research differently based on their identity (e.g., race, socioeconomic status, first-generation status).	0.847	0.875	0.938
26	23	Advocate for others who may be marginalized or excluded from the research environment.	Advocate for others who may be marginalized or excluded from the research environment.	0.825	0.852	0.878
Professional and Career Development Skills				α = 0.90	α = 0.90	α = 0.91
39	34	Explore possible research career pathways.	Demonstrate understanding of possible research career pathways.	0.910	0.908	0.892
8	6	Set research career goals.[d]	Set research career goals.	0.851	0.848	0.889
27	24	Develop a plan to pursue a research career (determine the next step in your training).	Develop a plan to pursue a research career (determine the next step in their training).	0.896	0.898	0.888
23	20	Confidence in pursuing a career in research.	Confidence in pursuing a career in research.	0.875	0.877	0.889

[a]T, trainee version of ERLA; M, mentor version of ERLA. Question stem for each item was "How much did you [your trainee] gain in your [their] ability to do the following over the course of your [their] research experience?" Adapted items are noted by superscripts: b, URSSA (Thinking and Working Like a Scientist); c, URSSA (Skills); d, MCA (Trainee adaptation); e, Estrada *et al.* 2011; f, URSSA (Personal Gains). The first column for trainees reports factor loadings for the full ERLA scale. The second column ($T_{aligned}$) provides factor loadings for the ERLA including only items that align directly with the mentor version of the scale.

TABLE 7. Means, standard deviations, and spearman correlations between ERLA and self-reported rating of the research experience and overall quality of the mentoring relationship[a]

	1	2	3	4	5	6	7	8	9	*M*	SD
1. Research Comprehension and Communication Skills	—	0.846	0.464	0.865	0.891	0.521	0.696	0.600	0.679	4.13	0.76
2. Practical Research Skills	0.873	—	0.510	0.811	0.816	0.545	0.674	0.549	0.624	3.85	0.90
3. Research Ethics	0.631	0.661	—	0.430	0.396	0.687	0.412	0.293	0.325	2.86	1.37
4. Research Identity	0.824	0.811	0.596	—	0.814	0.439	0.680	0.586	0.666	4.05	0.95
5. Researcher Confidence and Independence	0.888	0.873	0.609	0.848	—	0.476	0.673	0.582	0.664	4.10	0.87
6. Equity and Inclusion Awareness and Skills	0.615	0.603	0.775	0.563	0.600	—	0.408	0.258	0.315	2.94	1.50
7. Professional and Career Development Skills	0.746	0.718	0.593	0.811	0.760	0.573	—	0.434	0.575	3.65	1.19
8. Research Experience	0.563	0.488	0.335	0.505	0.525	0.280	0.446	—	0.778	4.35	0.87
9. Relationship Quality	0.511	0.465	0.278	0.444	0.481	0.255	0.401	0.744	—	4.34	0.86
M	4.19	3.99	3.50	3.93	4.09	3.41	3.77	4.50	4.44		
SD	0.69	0.76	1.08	0.87	0.78	1.10	1.01	0.90	0.89		

[a]Pairwise intercorrelations for trainees (N = 482–490) are presented below the diagonal, and intercorrelations for mentors (N = 182–309) are presented above the diagonal. Due to a survey error, several mentors were not presented with the research experience question. Responses for ERLA subscale could range from 1 (no gain) to 5 (great gain). Reponses to research experience and relationship quality could range from 1 (poor) to 5 (excellent). All correlations were statistically significant, $p < 0.001$.

which mentors and trainees scores aligned by subtracting the trainee scores from the mentor scores on each of the ERLA subscales and examining Spearman correlations between trainees' and mentors' scores. A positive difference score indicated that mentors rated trainee gains higher than the trainee, while a negative score indicated that the trainee rated the gains higher than the mentor rated the trainee's gains. Trainee–mentor pairs who were within an absolute value of 0.50 difference between scores were considered aligned. A positive, statistically significant correlation would indicate that trainees and mentors rate trainees' skill gains similarly.

Figure 2 shows that higher percentages of trainee–mentor pairs in our sample were more closely aligned on their assessment of trainee gains related to Research Comprehension and Communication Skills, Practical Research Skills, Researcher Identity, and Researcher Confidence and Independence, while misalignment was more frequent in assessing trainees' gains related to Research Ethics, Equity and Inclusion Awareness and Skills, and Professional and Career Development Skills. In cases in which trainees' assessment of their gains differed from their mentors' assessments, it was most often the case that a mentee rated his or her gains higher than his or her mentor's observation of gains.

Overall, Spearman correlations between trainees' and mentors' ratings of trainees' skills showed low correlation coefficients ($-0.009 \leq \rho \leq 0.216$), though two relationships were statistically significant: trainees' and mentors' ratings of trainee Research Comprehension and Communication Skills and Researcher Confidence and Independence ($\rho = 0.183$ and $\rho = 0.216$, respectively; $p < 0.05$; Supplemental Table S4). A more in-depth explanation of the implications of trainee–mentor alignment on the ERLA is presented in the Supplemental Material (see Analysis of Trainee/Mentor Alignment on ERLA as Predictive of Research Trainee Outcomes and Tables S5–S11).

DISCUSSION

The *Entering Research* Learning Assessment (ERLA), a new survey instrument for use with undergraduate and graduate research trainees and their mentors, was created using an iterative process of development, refinement, and collection of validity evidence. The ERLA scales are based on an evidence-based conceptual framework of seven areas of trainee development derived from the research literature (Branchaw *et al.*, 2020) and aligned with the recommendations of two National Academies reports on STEM undergraduate research experiences and graduate education (NASEM, 2017, 2018). Items and scales from existing instruments with validity evidence were leveraged to create ERLA, which provides a more comprehensive assessment of trainee learning gains compared with existing instruments and provides paired data from trainees and their mentors to allow alignment and comparison of reported gains. Unlike prior instruments, the ERLA addresses trainee growth in the areas of equity and inclusion and research ethics, both of which were identified as important to STEMM trainee development and success in the National Academies reports (NASEM, 2017, 2018) but have been infrequently measured in comprehensive assessments of research trainees. The ERLA can be used by any research training program interested in assessing trainee learning across the seven areas of trainee development, not only by those using the *Entering Research* curriculum in their program.

Evidence of Validity

Multiple types of validity evidence were collected, including evidence based on test content, internal structure, and convergent validity evidence, as well as evidence of internal consistency for the trainee and mentor versions of the ERLA.

Evidence based on test content assesses the extent to which an instrument relates to the construct(s) it intends to measure. The *Standards for Educational and Psychological Testing* note that evidence based on test content "can include logical or empirical analyses of the adequacy with which the test content represents the content domain and the relevance of the content domain to the proposed interpretation of test scores" (AERA *et al.*, 2014, p. 14). Validity evidence based on test content is provided through ERLA's alignment with the areas of trainee development and meta–learning objectives in the evidence-based *Entering Research* conceptual framework. Using a backward design approach, we first developed the meta–learning objectives used to frame the *Entering Research* curriculum based on feedback from a community of STEM practitioners, from key themes identified in the literature, and from the

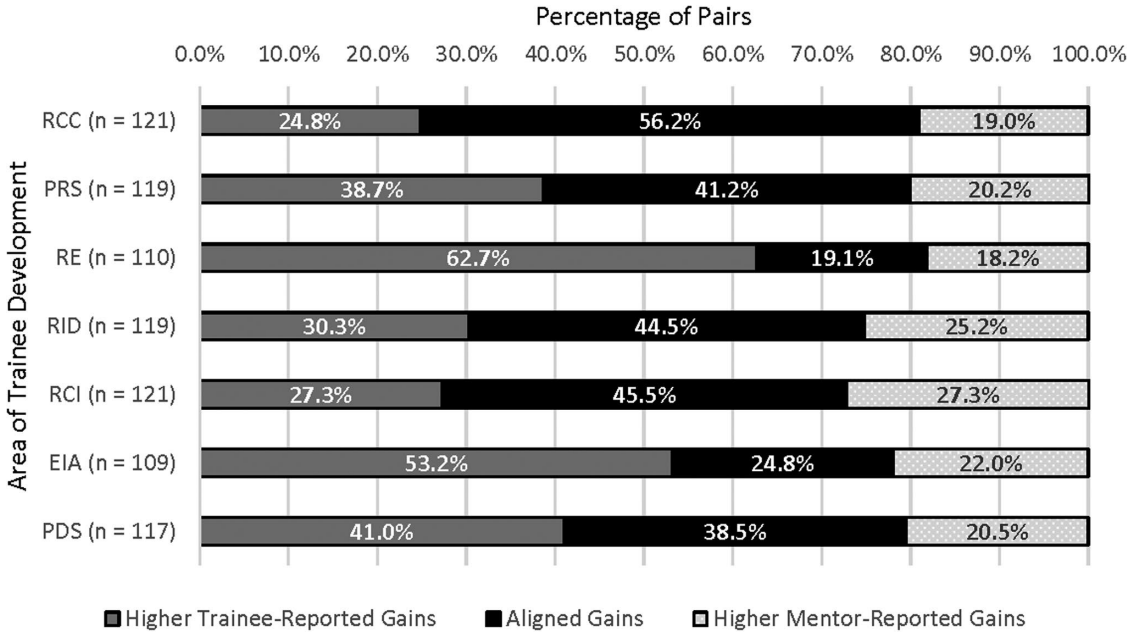

FIGURE 2. Degree of alignment between trainees' self-reported gains and mentors' assessment of trainee gains ($n = 121$). Alignment was calculated by subtracting the trainee's score for each subscale from the mentor's score for each subscale. RCC, Research Comprehension and Communication Skills; PRS, Practical Research Skills; RE, Research Ethics; RID, Researcher Identity; RCI, Researcher Confidence and Independence; EIA, Equity and Inclusion Awareness and Skills; PDS, Professional and Career Development Skills.

National Academies reports (NASEM, 2017, 2018). The final meta–learning objectives align with several goals, outcomes, and recommendations for optimizing undergraduate and graduate research training programs (NASEM, 2017, 2018), thus providing evidence that this instrument represents key aspects of undergraduate and graduate research experiences noted by practitioners and the literature on research training (Table 2).

The trainee and mentor versions of the ERLA each showed evidence of acceptable model–data fit (i.e., evidence of internal structure) based on the χ^2 and χ^2/df ratios, CFI, and RMSEA statistics reported for both versions of the instrument. This validity evidence suggests that the seven-factor structure of our instrument revealed in the EFA is a good fit to the data collected with a separate sample in stage 4 and that our items can be organized and scored according to the *Entering Research* conceptual framework. We were also able to collect validity evidence for internal structure for an aligned version of the ERLA for trainees that can be paired with the mentor scale to examine alignment between trainees' self-reported gains and mentors' perceptions of trainees' gains.

Intercorrelations between each of the ERLA subscales suggests that several of the ERLA subscales are highly related to one another (see Supplemental Tables S3.1 and S3.2). Research Comprehension and Communication Skills was highly correlated with the Practical Research Skills, Research Identity, and Researcher Confidence and Independence subscales for both trainees and mentors. This is not surprising, as mentored research experiences provide opportunities to increase proficiency in conducting, understanding, and communicating about research, which in turn can increase researcher identity, research self-efficacy, and independence (NASEM, 2017, 2018).

Evidence of convergent validity for both the trainee and mentor ERLA was collected by examining the relationship between each of the ERLA subscales and respondents' perceptions of the overall research experience and quality of the mentoring relationship. These relationships were all significant and positive, ranging between 0.278 and 0.563 for trainees and 0.258 and 0.679 for mentors. These relationships are similar in direction and magnitude to the relationships between the URSSA subscales and satisfaction reported by Weston and Laursen (2015), which ranged from 0.27 to 0.64. These findings suggest that the ERLA instrument correlates with measures of satisfaction with the research experience and overall quality of the mentoring relationship in a way that was expected and that aligns with previous findings.

Internal consistency statistics for each of the seven subscales were 0.81 or higher for both trainees and their mentors. These internal consistency statistics are similar to those reported for other instruments designed to assess research experiences or mentoring relationships (see Table 1); however, it is important to note that many factors, including the dimensionality of the instrument and number of items can contribute to these statistics; thus, these numbers cannot be used to make a direct comparison. Though we were only able to retain two of the three items originally included in the Equity and Inclusion Awareness and Skills subscale on the mentor version, we were encouraged by the high level of internal consistency of these items ($\alpha = 0.84$) and believe that this subscale shows sufficient evidence of validity to be incorporated into the instrument.

Multiple Measures of Research Trainee Learning Gains

Many assessments of trainee learning (e.g., URSSA, SURE) provide only trainee self-assessments of learning. The mentor version of the ERLA instrument offers research training program directors a second measure, beyond research trainee

self-reported data, of their research trainees' learning gains. Ideally, research training program directors should use multiple measures of trainee learning gains as evidence of trainee learning and, therefore, training program effectiveness (Gonyea, 2005; Kirkpatrick and Kirkpatrick, 2016). These include trainee self-assessments; mentor assessments of trainees; measures of the quality and quantity of the trainees' research products (e.g., papers and presentations); direct measures of trainee learning, such as written exams to assess content knowledge or practical exams to assess technical skills; and data that reflect satisfactory progress in the research program or toward research degree completion (e.g., passing a preliminary exam, successfully defending a thesis proposal). Along with the trainee and mentor ERLA surveys, we are exploring the development of a third, aligned ERLA survey for individuals in a trainee's mentoring network. This survey could be used by individuals who interact regularly with the trainee and are responsible for tracking progress, such as research training program directors, thesis committee members, and near peers. Learning assessment data from these individuals would provide additional and unique evidence of trainee learning.

Evidence from multiple measures of trainee learning gains will not only provide strong support for effectiveness claims about research training programs, but may also be used to evaluate whether research trainees are able to accurately self-assess their own learning gains. We hypothesize that accurate assessment of skills is a contributor to persistence and success in STEMM research. In general, research trainees' ability to accurately assess their learning gains reflects the ability to be metacognitive in planning, monitoring, and assessing his or her understanding and performance (Tanner, 2017; https://cft.vanderbilt.edu/guides-sub-pages/metacognition). Metacognition and accurate assessment of one's capabilities has been shown to be an important factor in students' academic success and performance (Countinho *et al.*, 2005; Sitzmann and Johnson, 2012). There is also evidence suggesting that, as trainees gain additional experience in research, their accuracy in self-assessment will also improve (Panadero *et al.*, 2016). Additional research on how the relationship between the ERLA and key trainee outcomes changes as a function of training stage will yield further insight into the extent to which trainee and mentor assessments of trainee learning gains relate to trainee outcomes across different stages of trainee development. The extent to which trainee and mentor assessments of trainees' skill gains align can serve as a starting point for trainees to reflect on their progress and can also serve as the basis for a conversation on trainee progress between trainees, their mentors, and research training program directors.

Paired Trainee–Mentor Data: What Does Alignment or Misalignment Mean?

Beyond providing multiple measures of research trainee learning gains, the paired ERLA trainee and mentor surveys provide the opportunity to examine whether and to what extent the trainee and mentor assessments of the trainee's learning gains align or are similar. Our initial findings with a subset of paired data from the larger sample suggest that trainees' and mentors' assessments of trainees' learning gains in the area of Research Comprehension and Communication Skills are more closely aligned (56% of pairs) compared with other areas of trainee development assessed with the ERLA. The largest differences between trainee and mentor assessments of trainee learning gains appear in the Research Ethics and Equity and Inclusion Awareness and Skills subscales of the ERLA. Further examination of the data revealed that many mentors selected "did not observe" for items on both of these subscales, which contributed to comparatively lower means for mentors, because "did not observe" is scored as "1."

The higher instance of "did not observe" responses for the equity and inclusion and research ethics items may reflect the extent to which these topics are addressed in the research experience itself, where the trainee and mentor interact. Other researchers have noted that culture and/or racial/ethnic identity are often not acknowledged or considered to be relevant to the research experience (e.g., Davidson and Foster-Johnson, 2001; Prunuske *et al.*, 2013). Comments from mentors during stage 1 pilot testing of the original version of ERLA stating that culture and/or racial/ethnic identity are not relevant to the research experience support these findings. Similarly, the discussion of research ethics is often left to leaders of professional development seminars, not research mentors, and ethics may not be explicitly discussed as part of the mentored research experience (DuBois *et al.*, 2008; Gasparich and Wimmers, 2014). Therefore, it is not surprising that mentors frequently reported "did not observe" for these items.

Another possible interpretation of the alignment or misalignment of trainee–mentor scores may be that it reflects the quality of the communication or alignment of expectations between mentors and trainees. However, the context of the research experience should be taken into account when considering this possible interpretation. Variations in trainee stage, expectations for independence, and research group climate could all influence the degree to which alignment may occur. For example, more advanced graduate trainees may not need a high level of oversight to be successful and may be able to more accurately assess their own skills, while earlier-stage trainees may overestimate their gains, as they are still coming to understand what is needed to be successful in research and are unable to accurately assess their skill levels (Kardash, 2000; Dunning *et al.*, 2003). Mentors who expect high levels of independence from their trainees may check in with their trainees less frequently, thus decreasing their ability to easily assess their trainees' learning gains over a given period of time. Conversely a very hands-on mentor can compare trainees' gains to those of previous trainees, resulting in a more informed assessment of trainee learning.

Though more research is needed to understand the implications of trainee overassessment, existing research on overconfidence yields several important considerations. Higher levels of confidence (and therefore overassessment of skill) may also be present among individuals who possess an entity, or fixed, mindset about intelligence (i.e., your intelligence cannot be changed; Ehrlinger *et al.*, 2016). The research by Ehrlinger *et al.* (2016) suggests that individuals who believe that you either have what it takes to be a scientist or you do not (i.e., having a fixed mindset) may be overconfident in their capabilities and as a result may avoid more challenging tasks in research. Moreover, a fixed mindset or a decrease in growth mindset is associated with higher levels of dropout from STEM majors (Dai and Cromley, 2014). It could be that overestimation of skills by trainees is an indicator of fixed mindset in trainees,

that, if caught early in the training experience, could be targeted through interventions. Conversely, other researchers have noted the benefits of overconfidence. For example, Bandura (1997) notes moderate amounts of overconfidence in one's capabilities can be adaptive when individuals are faced with new challenges. Therefore, overestimation of learning gains by novice trainees relative to a mentor's assessment of their learning gains may actually reflect a positive outcome. Future research to explore these possible interpretations of the trainee–mentor alignment (or misalignment) is planned.

Future research on ERLA will also focus on the extent to which trainee–mentor alignment on the ERLA changes over the course of the mentoring relationship and the predictive capacity of alignment for longer-term trainee outcomes, such as enrollment in graduate programs and scholarly productivity. In the meantime, this aligned instrument with evidence of internal structure and internal consistency provides a way to examine trainees' self-reported gains relative to their mentors' observations of their gains in the context of a research experience.

Recommendations for Use

The ERLA surveys measure gains and are therefore designed to be administered at the end of a research experience to collect summative data. However, they may also be administered midway through a research experience to collect formative data to guide program adjustments in real time. Midpoint administration may be particularly useful in undergraduate research programs lasting longer than one semester and in graduate training programs, where annual collection of data would allow program directors to track graduate student learning gains as they progress in the program.

Though the ERLA surveys were developed and align with the *Entering Research* curriculum and conceptual framework (Branchaw *et al.*, 2020), they can be used by any mentor, program director, or instructor interested in assessing research trainee learning gains in the seven areas of trainee development. Users may use the entire instrument or opt to incorporate individual subscales of the ERLA to assess one or more areas of trainee development.

When interpreting survey results, program directors using the ERLA should be aware of the extent to which each of the areas of trainee development is addressed in the research experiences of their trainees, either as part of the enrichment activities of a formal program or as part of the mentored research experience itself. Gains scores and differences between trainee and mentor assessments of trainee gains should be interpreted based on whether and to what extent the various areas of trainee development are explicitly addressed in their programs. Although we acknowledge that the exact skills and skill gains may look different at different training stages, the basic skill areas in which trainees are expected to grow are common. In future research, we plan to examine the extent to which the areas of trainee development are addressed in research experiences and within the context of trainee seminars to determine how each of these components of the overall experience contributes to trainee gains.

Based on the results of the confirmatory factor analyses presented in this paper, we are confident that the measurement model that treated mentors' responses of "did not observe" the same as "no gain" and the measurement model that treated mentors' responses of "did not observe" as missing both demonstrate acceptable model–data fit. Therefore, either approach can be used for scoring the mentor scale. Though it is common practice is to treat "did not observe" answers as missing, we believe the "did not observe" responses provide important information to practitioners interested in understanding the quality of mentor–trainee relationships. Therefore, we recommend that practitioners use the mentor version of the ERLA in conjunction with the trainee version of the ERLA to examine the extent to which trainees' self-reported gains align with the gains observed by their mentors and that they score the mentors' responses of "did not observe" the same as "no gain" to easily identify differences in trainee and mentor perceived gains. These alignment results can be used in conjunction with other program information (e.g., course syllabi, seminar agendas, other evaluation data) to ascertain whether or not research training programs are meeting their intended goals for trainee development. By contrast, practitioners and researchers who are not interested in using the trainee and mentor scales together to assess alignment or for program evaluation may elect to treat "did not observe" responses from mentors as missing. If researchers and practitioners choose to do this, we suggest they follow recommendations for treatment of missing data (e.g., multiple imputation techniques) based on the extent and nature of missing data in their sample before analyzing their data.

A copy of the scale manual for the ERLA, with example surveys and scoring instructions, is included in the Supplemental Material for this paper and with the *Entering Research* curriculum (Branchaw *et al.*, 2020). Additionally, the ERLA instrument, along with tools to evaluate research training programs, are preloaded into surveys hosted by the Wisconsin Institute for Science Education and Community Engagement (WISCIENCE; (https://wiscience.wisc.edu/program/evaluating-entering-research). These surveys are administered by WISCIENCE on behalf of research training programs and aggregated reports of data, including comparisons of mentor and trainee ERLA data, are provided.

Limitations

The ERLA provides two sources of data to assess trainee learning gains: self-assessment data from the trainee and mentor assessment data. Ideally, as outlined in the *Discussion*, other objective forms of data should be collected to contextualize and inform the interpretation of the ERLA data. The ERLA instrument, like other commonly used assessments of research experiences (e.g., URSSA), measures trainee gains at one point in time. Therefore, the data collected with ERLA provide a snapshot and should be interpreted within the context in which they were collected. For example, a trainee who reports "no gain" on a particular item or a comparatively low mean on a subscale of the ERLA may have chosen that option because the topic was not addressed or because he or she was already proficient in the area. ERLA users need to consider their trainees' career stages and prior research experiences and the topics addressed in their training seminars when interpreting results.

We were not able to ascertain the extent to which trainees who completed the ERLA in stages 2 and 4 were engaged in training programs compared with those engaged in mentored research without the support of a structured training program. Future research examining gains of trainees in formal research

training programs compared with those not engaged in formal programs could shed light on the importance of formal programs in supporting trainee development, particularly in the areas of equity and inclusion and research ethics.

Though we have presented robust validity evidence for ERLA, further evidence should be collected with more diverse populations of trainees and at various undergraduate and graduate career stages. Specifically, although initial evidence for content validity was provided through the alignment of items to the *Entering Research* conceptual framework, review by a panel of experts would provide additional evidence of validity based on test content. Also, due to the low number of mentors of graduate trainees who completed our survey at stage 4, additional validity evidence for the mentor version of the ERLA should be collected with mentors of graduate trainees. Additional responses from mentors of graduate trainees could also provide additional insight into how the length of the mentoring relationship impacts the alignment of trainee and mentor scores on the ERLA, as graduate trainees often work with their research mentors for longer periods of time. We were not able to present evidence of criterion validity due to the cross-sectional nature of the data collected thus far, and evidence of the predictive validity of ERLA has yet to be collected, as trainees must be tracked to document their research career trajectories and outcomes to do this.

Collecting additional data from larger and more diverse samples will permit us to conduct additional analyses on the ERLA using more advanced psychometric evaluation techniques, such as those offered by item response theory (deAyala, 2009). These techniques can provide additional insight on the relationship between observed responses and the ERLA subscales identified in this paper. Option response function plots could provide evidence of whether response categories are being used as expected and in the expected order. This approach will help us further understand how the "did not observe" category is interpreted by mentors and will yield additional insight on what number or type of response categories are best used with research mentors. Differential item functioning analyses could reveal whether the instrument is operating similarly for trainees across different racial, ethnic, and gender identities and training stages. Efforts to collect more sophisticated evidence of validity are underway.

CONCLUSIONS

Gathering validity evidence is an ongoing, iterative process (Campbell and Nehm, 2013; Reeves and Marbach-Ad, 2016). In this paper, we present initial validity evidence for internal content, convergent validity, and evidence of internal consistency for trainee and mentor versions of the ERLA and for an aligned ERLA scale. Although more research is needed to gather evidence of criterion validity and to investigate the role that trainee–mentor alignment plays in predicting trainee outcomes, the evidence provided in this paper is sufficient for program directors to use the ERLA and interpret scores with confidence, both independent of and in conjunction with training implementations that use the *Entering Research* curriculum (Branchaw *et al.*, 2020).

The ERLA and the validity evidence presented in this paper move the field of research training forward for both practitioners and researchers. Practitioners can use the ERLA and conceptual framework as a guide for choosing program activities and training that complement the research experience and provide a holistic, trainee-centered program. This more comprehensive assessment of trainee learning will also be of use to researchers who wish to examine the short- and long-term impacts of research training experiences in relation to trainee learning and skill gains. The ERLA extends the work of previous researchers by providing a more comprehensive tool with which research training program directors can assess trainee learning and outcomes. In addition, ERLA aligns with two national, comprehensive reports on the common elements of undergraduate and graduate research training programs and training experiences (NASEM, 2017, 2018) and explicitly assesses trainee development in the areas of ethics and equity and inclusion awareness, which were overlooked in prior comprehensive assessments of trainee learning. The ERLA provides a tool that is needed to assess efforts underway across the nation to diversify the STEMM research workforce.

ACKNOWLEDGMENTS

Work reported in this publication was supported by the National Institutes of Health (NIH) Common Fund and Office of Scientific Workforce Diversity under award U54 GM119023 (NRMN), administered by the National Institute of General Medical Sciences and by the Wisconsin Institute for Science Education and Community Engagement (WISCIENCE) and the Department of Kinesiology at the University of Wisconsin–Madison. The work is solely the responsibility of the authors and does not necessarily represent the official view of the NIH or the University of Wisconsin–Madison. A special thanks to the trainees and mentors who participated in this NIH Diversity Program Consortium study.

REFERENCES

Adedokun, O. A., Bessenbacher, A. B., Parker, L. C., Kirkham, L. L., & Burgess, W. D. (2013). Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research on self-efficacy. *Journal of Research in Science Teaching*, *50*(8), 940–951.

American Academy of Arts and Sciences. (2013). *Arise 2: Unleashing America's research and innovation enterprise*. Cambridge, MA: American Academy of Arts and Sciences.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Balster, N., Pfund, C., Rediske, R., & Branchaw, J. (2010). *Entering Research*: A course that creates community and structure for beginning undergraduate researchers in STEM disciplines. *CBE—Life Sciences Education*, *9*, 108–118.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barker, L. (2009). Student and faculty of undergraduate research experiences in computing. *ACM Transactions on Computing Education*, *9*(1), 1–28.

Bauer, K. W., & Bennett, J. S. (2003). Alumni perceptions used to assess undergraduate research experience. *Journal of Higher Education*, *74*(2), 210–230.

Branchaw, J. L., Butz, A. R., & Smith, A. R. (2020). *Entering research: A curriculum to support undergraduate and graduate research trainees* (2nd ed.). New York: Macmillan.

Branchaw, J. L., Pfund, C., & Rediske, R. (2010). *Entering research: Workshops for students beginning research in science*. New York: Freeman.

Bumpus, N. (2015, December 7). Moving toward inclusion. *Science*. https://doi.org/10.1126/science.caredit.a1500273

Butz, A. R., Spencer, K., Thayer-Hart, N., Cabrera, I. E., & Byars-Winston, A. (2018). Mentors' motivation to address race/ethnicity in research mentoring relationships. *Journal of Diversity in Higher Education*, *12*(3), 242–254. https://doi.org/10.1037/dhe0000096

Byars-Winston, A. M., Branchaw, J., Pfund, C., Leverett, P., & Newton, J. (2015). Culturally diverse undergraduate researchers' academic outcomes and perceptions of their research mentoring relationships. *International Journal of Science Education*, *37*, 2533–2554.

Byars-Winston, A., Rogers, J., Branchaw, J., Pribbenow, C., Hanke, R., & Pfund, C. (2016). New measures assessing predictors of academic persistence for historically underrepresented racial/ethnic undergraduates in science. *CBE—Life Sciences Education*, *15*, 1–11. doi: 10.1187/cbe.16-01-0030

Campbell, C. E., & Nehm, R. H. (2013). A critical analysis of assessment quality in genomic and bioinformatics education research. *CBE—Life Sciences Education*, *12*, 530–541. doi: 10.1187/cbe.12-06-0073

Carpi, A., Ronan, D. M., Falconer, H. M., & Lents, N. H. (2017). Cultivating minority scientists: Undergraduate research increases self-efficacy and career ambitions for underrepresented students in STEM. *Journal of Research in Science Teaching*, *54*, 169–174.

Carter, D. F., Ro, H. K., Alcott, B., & Lattuca, L. (2016). Curricular connections: The role of undergraduate research experiences in promoting engineering students' communication, teamwork, and leadership skills. *Research in Higher Education*, *57*(3), 363–393.

Chemers, M. M., Syed, M., Goze, B. K., Zurbriggen, E. L., Bearman, S., Crosby, F. J., … & Morgan, E. M. (2010). *The role of self-efficacy in mediating the effects of science support programs (Technical Report No. 5)*. Santa Cruz: University of California.

Chemers, M. M., Zurbriggen, E. L., Syed, M., Goza, B. K., & Bearman, S. (2011). The role of efficacy and identity in science career commitment among underrepresented minority students. *Journal of Social Issues*, *67*(3), 469–491.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Coker, J. S., & Van Dyke, C. G. (2005). Evaluation of teaching and research experiences undertaken by botany majors at N.C. State University. *North American Colleges and Teachers of Agriculture Journal*, *49*, 14–19. Retrieved November 1, 2019, from www.jstor.org/stable/43765919

Coutinho, S., Wiemer-Hastings, K., Skowronski, J. J., & Britt, M. A. (2005). Metacognition, need for cognition and use of explanations during ongoing learning and problem solving. *Learning and Individual Differences*, *15*, 321–337. https://doi.org/10.1016/j.lindif.2005.06.001

Craney, C., McKay, T., Mazzeo, A., Morris, J., Prigodich, C., & de Groot, R. (2011). Cross-discipline perceptions of the undergraduate research experience. *Journal of Higher Education*, *82*(1), 92–113.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Dai, T., & Cromley, J. G. (2014). Changes in implicit theories of ability in biology and dropout from STEM majors: A latent growth curve approach. *Contemporary Educational Psychology*, *39*, 233–247. https://doi.org/10.1016/j.cedpsych.2014.06.003

Davidson, M. N., & Foster-Johnson, L. (2001). Mentoring in the preparation of graduate researchers of color. *Review of Educational Research*, *71*, 549–574.

deAyala, R. J. (2009). *The theory and practice of Item response theory*. New York: Guilford.

DeMars, C. E., & Erwin, T. D. (2004). Scoring *neutral* or *unsure* on an identify development instrument for higher education. *Research in Higher Education*, *45*, 83–95. doi: https://doi.org/10.1023/B:RIHE.0000010048.18517.b4

Denman, D. C., Baldwin, A. S., Betts, A. C., McQueen, A., & Tiro, J. A. (2018). Reducing "I don't know" responses and missing survey data: Implications for measurement. *Medical Decision Making*, *38*, 673–682. https://doi.org/10.1177/0272989x18785159

DuBois, J. M., Dueker, M. J. M., Anderson, E. E., & Campbell, J. (2008). The development and assessment of an NIH-funded research ethics training program. *Academic Medicine: Journal of the Association of American Medical Colleges*, *83*(6), 596.

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

Eagan, M. K., Hurtado, S., Chang, M. J., Garcia, G. A., Herrera, F. A., & Garibay, J. C. (2013). Making a difference in science education: The impact of undergraduate research programs. *American Educational Research Journal*, *50*, 683–713. doi: 10.3102/0002831213482038

Ehrlinger, J., Mitchum, A. L., & Dweck, C. S. (2016). Understanding overconfidence. Theories of intelligence, preferential attention, and distorted self-assessment. *Journal of Experimental Social Psychology, 63*, 94–100. https://doi.org/10.1016/j.jesp.2015.11.001

Estrada, M., Woodcock, A., Hernandez, P., & Schultz, P. W. (2011). Toward a model of social influence that explains minority student integration into the scientific community. *Journal of Educational Psychology*, *103*, 206–222. doi: 10.1037/a0020743

Ferrini-Mundy, J. (2013). Driven by diversity. *Science*, *340*(6130), 278.

Field, S., Kuczera, M., & Pont, B. (2007). *No more failures: Ten steps to equity in education*. Paris: Organisation for Economic Co-operation and Development. Retrieved November 1, 2019, from www.oecd.org/education/school/45179151.pdf

Fleming, M., House, S., Hanson, V. S., Yu, L., Garbutt, J., McGee, R., … Rubio, D. (2013). The mentoring competency assessment: Validation of a new instrument to evaluate skills of research mentors. *Academic Medicine*, *88*, 1002–1008.

Gasparich, G. E., & Wimmers, L. (2014). Integration of ethics across the curriculum: From first year through senior seminar. *Journal of Microbiology Education*, *15*(2), 218–223.

Gilmore, J., Vieyra, M., Timmerman, B., Feldon, D., & Maher, M. (2015). The relationship between undergraduate research participation and subsequent research performance of early career STEM graduate students. *Journal of Higher Education*, *86*(6), 834–863.

Gonyea, R. M. (2005). Self-reported data in institutional research: Review and recommendations. *New Directions for Institutional Research*, *127*, 73–89. https://doi.org/10.1002/ir.156

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*, 206–213. doi: 10.1007/s11121-007-0070-9

Hanauer, D. I., & Dolan, E. L. (2014). The Project Ownership Survey: Measuring differences in scientific inquiry experiences. *CBE—Life Sciences Education*, *13*(1), 149–158.

Harsh, J. A., Maltese, A. V., & Tai, R. H. (2011). Undergraduate research experiences from a longitudinal perspective. *Journal of College Science Teaching*, *41*, 84–91.

Hathaway, R. S., Nagda, B. A., & Gregerman, S. R. (2002). The relationship of undergraduate research participation to graduate and professional education pursuit: An empirical study. *Journal of College Student Development*, *43*(5), 1–18.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Hunter, A. B., Laursen, S. L., & Seymour, E. (2007). Becoming a scientist: The role of undergraduate research in cognitive, personal and professional development. *Science Education*, *91*(1), 36–74.

Hunter, A-B., Weston, T. J., Laursen, S. L., & Thiry, H. (2009). URSSA: Evaluating student gains from undergraduate research in science education. *Council on Undergraduate Research Quarterly*, *29*(3), 15–19.

Hurtado, S., Cabrera, N. L., Lin, M. H., Arellano, L., & Espinosa, L. L. (2009). Diversifying science: Underrepresented student experiences in structured research programs. *Research in Higher Education*, *50*, 189–214. doi: 10.1007/s11162-008-9114-7

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist*. London: Sage.

IBM SPSS Statistics for Windows. (2015). *[computer software]*. Armonk, NY: IBM Corporation.

John, J., & Creighton, J. (2012). "In practice, it doesn't always work out like that." Undergraduate experiences in a research community of practices. *Journal of Further and Higher Education*, *37*(6), 1–19.

Junge, B., Quiñones, C., Kakietek, J., Teodorescu, D., & Marsteller, P. (2010). Promoting undergraduate interest, preparedness, and professional pursuit in the sciences: An outcomes evaluation of the SURE program at Emory University. *CBE—Life Sciences Education*, *9*(2), 119–132.

Kardash, C. M. (2000). Evaluation of an undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of Educational Psychology*, *92*(1), 191–201.

Kardash, C. M., & Edwards, O. V. (2012). Thinking and behaving like scientists: Perceptions of undergraduate science interns and their faculty mentors. *Instructional Science*, *40*, 875–899.

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of evaluation*. Alexandria, VA: ATD Press.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, *18*(1): rm1. doi: 10.1187/cbe.18-04-0064

Laursen, S., Hunter, A.-B., Seymour, E., Thiry, H., & Melton, G. (2010). *Undergraduate research in the sciences: Engaging students in real science*. San Francisco, CA: Jossey-Bass.

Lee, R. M., & Davis, C. III. (2000). Cultural orientation, past multicultural experience, and a sense of belonging on campus for Asian American college students. *Journal of College Student Development*, *41*(1), 110–115.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *82*, 1198–1202.

Lopatto, D. (2003). The essential features of undergraduate research. *Council on Undergraduate Research Quarterly*, *24*, 139–142.

Lopatto, D. (2004). Survey of Undergraduate Research Experiences (SURE): First findings. *Cell Biology Education*, *3*(4). https://doi.org/10.1187/cbe.04-07-0045

Lopatto, D. (2007). Undergraduate research experiences support science career decisions and active learning. *CBE—Life Sciences Education*, *6*, 297–306. doi: 10.1187/cbe.07-06-0039

Mancha, R., & Yoder, C. Y. (2014). Factors critical to successful undergraduate research. *Council on Undergraduate Research Quarterly*, *34*, 38–46.

Mendoza-Denton, R., Downey, G., Purdie, V. J., Davis, A., & Pietrzak, J. (2002). Sensitivity to status-based rejection: Implications for African American students' college experience. *Journal of Personality and Social Psychology*, *83*(4), 896–918.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Myers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Thousand Oaks, CA: Sage.

National Academies of Sciences, Engineering, and Medicine (NASEM). (2017). *Undergraduate research experiences for STEM students: Successes, challenges, and opportunities*. Washington, DC: National Academies Press. https://doi.org/10.17226/24622

NASEM. (2018). *Graduate STEM education for the 21st century*. Washington, DC: National Academies Press. https://doi.org/10.17226/24622

Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Panadero, E., Brown, G. T. L., & Strijbos, J-W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review*, *28*, 803–830. https://doi.org/10.1007/s10648-015-9350-2

Pfund, C., Branchaw, J. L., & Handelsman, J. (2015). *Entering mentoring: A seminar to train a new generation of scientists* (2nd ed.). New York: Macmillan.

Pfund, C., House, S., Asquith, P., Spencer, K., Silet, K., & Sorkness, C. (2013). *Mentor training for clinical and translational researchers*. New York: Macmillan.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC: U.S. Government Office of Science and Technology. Retrieved June 15, 2019, from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/pcast-engage-to-excel-final_2-25-12.pdf

Prunuske, A. J., Wilson, J., Walls, M., & Clarke, B. (2013). Experiences of mentors training underrepresented undergraduates in the research laboratory. *CBE—Life Sciences Education*, *12*, 403–409. doi: 10.1187/cbe.13-02-0043

Raelin, J. A., Bailey, M. B., Hamann, J., Pendleton, L. K., Reisberg, R., & Whitman, D. L. (2014). The gendered effect of cooperative education, contextual support, and self-efficacy on undergraduate retention. *Journal of Engineering Education*, *103*(4), 599–624.

Reagans, R., & Zuckerman, E. W. (2001). Networks, diversity, and productivity: The social capital of corporate R&D teams. *Organization Science*, *12*(4), 502–517.

Reeves, T. D., & Marbach-Ad, G. (2016). Contemporary test validity in theory and practice: A primer for discipline-based education researchers. *CBE—Life Sciences Education*, *15*, 1–9. doi: 10.1187/cbe.15-08-0183

Roberge, M. É., & Van Dick, R. (2010). Recognizing the benefits of diversity: When and how does diversity increase group performance? *Human Resource Management Review*, *20*(4), 295–308.

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). Benefits of undergraduate research experiences. *Science*, *316*, 548–549.

Saxena, A. (2014). Workforce diversity: A key to improve productivity. *Procedia Economics and Finance*, *11*(Suppl. C), 76–85.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177. doi: 10.1037//1082-989X.7.2.147

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis: A review. *Journal of Education Research*, *99*, 323–337.

Seymour, E., Hunter, A. B., Laursen, S. L., & DeAntoni, T. (2004). Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. *Science Education*, *88*, 493–534.

Sitzmann, T., & Johnson, S. K. (2012). When is ignorance bliss? The effects of inaccurate self-assessments of knowledge on learning and attrition. *Organizational Behavior and Human Decision Processes*, *117*, 192–207. doi: 10.1016/j.obhdp.2011.11.004

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629. http://dx.doi.org/10.1037/0003-066X.52.6.613

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Tanner, K. D. (2017). Promoting student metacognition. *CBE—Life Sciences Education*, *11*(2), 113–120. https://doi.org/10.1187/cbe.12-03-0033

Thiry, H., Laursen, S. L., & Hunter, A. B. (2011). What experiences help students become scientists? A comparative study of research and other sources of personal and professional gains for STEM undergraduates. *Journal of Higher Education*, *82*(4), 358–389.

Valantine, H. A., & Collins, F. S. (2015). National Institutes of Health addresses the science of diversity. *Proceedings of the National Academy of Sciences USA*, *112*, 12240–12242. https://doi.org/10.1073/pnas.1515612112

Villa, E. Q., Kephart, K., Gates, A. Q., Thiry, H., & Hug, S. (2013). Affinity research groups in practice: Apprenticing students in research. *Journal of Engineering Education*, *102*(3), 444–466.

Watts, L. L., Medeiros, K. E., Mulhearn, T. J., Steele, L. M., Connelly, S., & Mumford, M. D. (2017). Are ethics training programs improving? A meta-analytic review of past and present ethics instruction in the sciences. *Ethics & Behavior*, *27*(5), 351–384.

Wenzel, T. J., Larive, C. K., & Frederick, K. A. (2012). Role of undergraduate research in an excellent and rigorous undergraduate chemistry curriculum. *Journal of Chemical Education*, *89*, 7–9. https://doi.org/10.1021/ed200396y

Weston, T. J., & Laursen, S. L. (2015). The Undergraduate Research Student Self-Assessment (URSSA): Validation for use in program evaluation. *CBE—Life Sciences Education*, *14*, 1–10.

Wiggins, G. P., McTighe, J., Kiernan, L. J., & Frost, F. & Association for Supervision and Curriculum Development. (1998) *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Woodcock, A., Graziano, W. G., Branch, S. E., Ngambeki, I., & Evangelou, D. (2012). Engineering students' beliefs about research: Sex differences, personality, and career plans. *Journal of Engineering Education*, *101*(3), 495–511.

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. M. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, *330*, 686–688. doi: 10.1126/science.1193147