COMMENTARY

So, You Want to Measure Something? An Introduction to Measurement Validity in Educational Research

Trevor T. Tuma, Erin L. Dolan, *University of Georgia*

Abstract

Educational researchers use a variety of data collection tools with the goal of measuring students' educational experiences, learning, and outcomes. Measurement validity is essential for ensuring that research findings lead to trustworthy conclusions. Yet it can be challenging to understand what measurement validity is, how it is established, and why it matters when drawing conclusions from educational data. In this commentary, we introduce educational measurement with a focus on validity. We explain the process of assessing the validity of an educational measurement tool and how conclusions can be undermined when we ignore validity. We offer recommendations for identifying measurement tools with strong validity evidence and point readers to resources that have informed our understanding and application of measurement validity in our own educational research.

Keywords: *assessment, measurement, validity, standards, faculty development*

doi: 10.18833/spur/9/1/8

As scholars of undergraduate research, scholarship, and creative inquiry (URSCI), we are interested in understanding what makes these experiences effective and impactful. To do this successfully, we must first articulate what we intend to study, and then identify tools that are capable of measuring it accurately. For instance, if we are interested in whether students believe they can do research—or the extent of their research self-efficacy—we must first define what research self-efficacy is, and then use a tool that can accurately and consistently measure it. This task is challenging because we can not just pull out a research self-efficacy "gauge" and measure levels of research self-efficacy. Research self-efficacy is not a physically observable variable like someone's height or blood pressure. Rather, we need to find a trustworthy way to gauge their research self-efficacy. This is the idea of validity in educational measurement.

Finding ways to make measurements that support valid inferences, meaning conclusions that are true or trustworthy, is critical for several reasons. Let us continue our research self-efficacy example to illustrate. If we use a survey of research self-efficacy that does not truly measure research self-efficacy, we may miss detecting a real and meaningful effect of an undergraduate research experience simply because we were not looking with the right tool (i.e., a false negative). Inaccurate measurements also can lead to false positives. For example, if we want to measure the gains students make in their research skills as a result of a research experience, but we use a survey to ask them about their skill gains, we may actually be measuring their research self-efficacy rather than their actual skills. In this case, the effect is real, but the conclusion is misleading because the variable being measured does not match the variable of interest. No study design or analytic approaches can compensate for erroneous measurement. It takes time, effort, money, and resources to collect educational data—both from the scholars doing the work and from the participants in the research. Accurate measurements help to ensure that this investment generates useful data and leads to trustworthy conclusions that advance the field by contributing knowledge that others can depend on.

In this commentary, we aim to help readers understand what validity means in the context of measurement, why measurement validity matters, and how to make measurement

decisions in educational settings. We hope to accomplish this by clearly defining measurement validity in the context of educational research and sharing some of the lessons we have learned based on our own experiences with measurement.

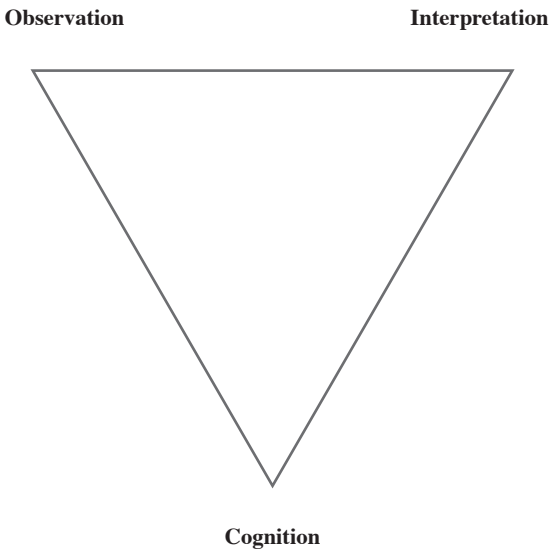
What Is Measurement Validity?

Confusion about measurement validity persists in educational scholarship, often because of the ways validity is discussed. For instance, scholars talk about "validated instruments," "valid surveys," and "valid data collection methods." This phrasing implies that validity is an inherent property of a data collection tool such as an exam or questionnaire. Rather, validity is not a static property, but an argument based on evidence regarding what is being measured, in what context, with what people, and for what purpose. Validity is about whether and to what extent trustworthy inferences can be made from data. To do this, we have to be able to trust that our measures are the right tools for the job.

This process starts by clearly defining what we want to measure. To illustrate this, we will describe our experiences developing a measure of mentoring in undergraduate research. We started this journey because we noticed that students who had negative experiences with their research mentors were choosing not to participate in our research studies. This made us curious about the idea of negative mentoring experiences: experiences that undergraduate researchers have with their research mentors that they view as negative for whatever reason. "Negative mentoring experiences" are a latent variable or "construct," an unobservable idea that can not be measured directly. We wanted to find a way to measure this construct, so we needed to find a way to make it observable and quantifiable. The idea of an assessment triangle is a useful framework for accomplishing this (Fig. 1; Glaser et al. 2001). We will illustrate how to use the assessment triangle by walking through our own measurement work (Limeri et al. 2024).

The assessment triangle visually depicts the three essential components of any assessment (Fig. 1). The first component is cognition, or the thoughts, feelings, or beliefs a person has that are of interest to assess. We have found it helpful to think of cognition as the construct of interest, including how we would define it conceptually. For our negative mentoring work, we started by reviewing literature to see if anyone had defined negative mentoring in undergraduate research. Although we found descriptions of negative mentoring of employees by supervisors in corporate settings, we found no description of negative mentoring that undergraduate researchers might experience, and we thought this would and should be distinct from negative mentoring that employees experience. Therefore we conducted a qualitative study of negative mentoring in undergraduate research to generate its conceptual

FIGURE 1. Assessment Triangle



Note: Adapted from Glaser et al. (2001).

definition (Limeri et al. 2019), which included absenteeism, abuse of power, interpersonal mismatch, lack of technical support, lack of psychosocial support, misaligned expectations, and unequal treatment. Measurement experts call this the "content" of the construct, that is, what the construct represents (American Educational Research Association et al. 2014).

With a conceptual definition of the construct in hand, we need to think about the second component of the assessment triangle, which is observation, or what a person says or does that demonstrates the cognition. We have found it helpful to think of observation as an operational definition: what we can observe that indicates the presence, absence, or extent of the construct. In the case of negative mentoring, we spent a year designing, pilot testing, and refining survey items that represented the range and types of negative mentoring observed in our qualitative study (Limeri et al. 2024).

The final component of the assessment triangle is interpretation, or the reasoning of how the observation logically follows from the cognition. To have confidence that our instrument, the Mentoring in Undergraduate Research Survey (MURS), was accurately measuring the range of mentoring that undergraduate researchers experienced, we collected multiple sources of evidence and made an argument that the evidence indicated the observations (i.e., the responses on the survey) reflected the cognition (i.e., how the undergraduate researchers experienced mentoring). Specifically, we interviewed undergraduate researchers who had a range of mentoring experiences, which yielded

evidence that students were understanding and interpreting the survey items (i.e., their response processes) as we intended. We conducted a sorting activity with experts in mentoring and undergraduate research in which we asked them to group the survey items into the various types of negative mentoring we had observed previously. This produced evidence that our survey items represented the construct clearly and evenly across a range of key ideas and experiences, and that we hadn't missed any key ideas or experiences. We collected data from a national sample of more than 500 undergraduate researchers and conducted statistical tests to assess whether the items represented the construct as we hypothesized. At the same time, we collected data on students' personality traits and ways of relating to other people to make sure we were not accidentally measuring other constructs, such as students' tendencies to view things negatively or develop insecure relationships. Finally, we collected data longitudinally from a second national sample of undergraduates who were starting research for the first time. This allowed us to determine the prevalence (or base rate) of negative mentoring (good news! it occurs infrequently) and assess the extent to which experiencing negative mentoring is detrimental (bad news—it is). This evidence showed that the MURS was useful for making interpretations about undergraduate research mentoring experiences.

Collectively, we argued that the evidence we collected supported the inference that the MURS was measuring the mentoring that undergraduate researchers experience. This is the foundation of validity, or an argument that a measure is being used in a reasonable way to make inferences about a construct of interest. This does not mean that the MURS is "validated," because it depends on how it is being used and what inferences are being made based on the results. In the next section, we explain how to make decisions about measurement that equip you to make valid inferences about your constructs of interest.

Are You Ready to Measure?

Now that you have more insights into all that goes into assessing the validity of a measure, you know that choosing the right one requires carefully considering whether the interpretation and use of a measure are appropriate for your specific context, population, and purpose. For reference, we will continue to use the term "measure" as a catch-all word for surveys, tests, and other data collection tools that scholars use to solicit written responses from participants, either in person or online. There are other approaches to measuring in educational research, such as observing student behaviors, calculating grades, or interviewing students, but those are beyond the scope of what we address here. We will use the term "item" to refer to questions, statements, or prompts to which students respond; responses to items are considered observations in our assessment triangle.

Choosing the right measure involves checking for:

- A clear conceptual definition of the construct being measured (i.e., cognition);
- A well-aligned operational definition of the construct, such that the items reflect what people actually think, feel, or experience related to the construct (i.e., observations); and
- A logical argument that responses to the items are a trustworthy reflection of the construct (i.e., interpretation).

We will illustrate the importance of these considerations by discussing how a widely used college assessment tool, the Test of Scientific Literacy Skills (TOSLS), can be misused, leading to invalid conclusions.

The TOSLS is a multiple-choice test designed to measure undergraduates' understanding of scientific concepts in college science courses (Gormally, Brickman, and Lutz 2012). Initial validation efforts showed that the TOSLS could provide meaningful insights on the gains that introductory undergraduate students made in their scientific literacy based on how their course was taught (Gormally et al. 2012). For example, students in nonmajor project-based courses showed greater learning gains than those in traditional lecture-based courses (Gormally et al. 2012). Given these encouraging results, it may be tempting to assume that the TOSLS is a "valid measure" of scientific literacy. Yet this is when it is essential to consider the nuances of validity.

What if we tried to use the TOSLS to measure the scientific literacy of science PhD students? The TOSLS almost certainly would not generate useful information, not because the tool was flawed but because the population differed from the population that the measure was designed for. The TOSLS is likely too easy for science PhD students, who would all (hopefully!) score 100 percent. Although it might be encouraging to see uniformly high performance, the scores would tell us very little about the students' actual scientific literacy. What if we tried to use the TOSLS to measure the scientific literacy of middle school students? The TOSLS might be too difficult for reasons other than their developmentally appropriate level of scientific literacy, such as their reading level. What if we used the TOSLS to measure high school students' scientific literacy, and a few performed well, but most did not? We might assume that the high performers were more scientifically literate, and this might be the case. However, subsequent research indicates that SAT reading scores are the strongest predictors of TOSLS performance (Shaffer, Ferguson, and Denaro 2019), suggesting that reading comprehension (not scientific literacy) might be driving performance on the TOSLS.

Suppose that undergraduate students who took the TOSLS answered a high proportion of the questions correctly, but a

closer examination suggested that the high-performing students were using test-taking strategies (e.g., word matching, selecting responses based on word length) rather than scientific literacy to answer the questions. If students were not engaging in the types of thinking the TOSLS aims to assess, known as "response process validity," the inferences we made from their scores might not reflect their actual scientific literacy.

Finally, what if we used students' TOSLS scores to make decisions about whether college students should gain admission to certain courses or majors? Although there might be good intentions for this, such as making sure students experienced the right level of challenge in their courses, this could be deeply problematic. The TOSLS was not designed to be used in this way, and using TOSLS results for misaligned purposes could cause unintended harm to students.

Hopefully the TOSLS and examples of its use illustrate why we should avoid referring to any measure as "validated." Validity is not the static property of a measure, but rather a judgment about how confident we can be about the inferences we draw from the resulting data in the context of what is being measured, with whom, in what context, and for what purpose. This example also highlights that no single metric or piece of evidence can be used to argue that a measure is valid. Importantly, it discourages viewing validity as a binary outcome (e.g., valid or not valid) that can be determined by a single statistic. Rather, validity reflects the variety and quality of the evidence that the measure is operating as intended.

Ultimately, scholars must carefully evaluate whether or not a measure has sufficient validity evidence for their intended use. This means assessing existing evidence of validity or gathering new evidence before using the tool for data collection or interpretation of results. Without these considerations, any conclusion drawn from the data may be questionable or misleading. Validity, then, is not an afterthought; it is a prerequisite and essential component of study design in educational research.

How Do You Identify the Best Measure?

Now that you know more about both validity and how we can be misled as scholars when we assume the validity of a measure, we offer guidance on how to identify measures that will best position you to make valid inferences. We present this guidance as a series of steps, recognizing that the actual process is more dynamic, iterative, and messy.

1. **Identify your construct of interest.** This is the most important step because it sets the focus and boundaries for finding a suitable measure. We start by spending time thinking and discussing our construct of interest, what makes it a unique or distinctive idea, and how we know it when we see it. Then we draft both a conceptual and operational definition.
2. **Search the literature.** Although we like to think we are always identifying new and interesting phenomena to study, most of the time the constructs we identify have already been described and a measurement tool has been developed, ideally along with some validity evidence. To find these measures, we typically search Google Scholar with the name of our construct and the word "measure" or "measurement." If we are not sure what search terms to use for our construct, we ask colleagues who do research in that area. For instance, when we started studying negative mentoring in undergraduate research, we consulted a colleague who studied mentoring in the workplace to ask them about scholarly terms for what we were observing. They suggested "negative mentoring" along with "abusive supervision" and "workplace incivility," which allowed us to search more strategically and learn more.
3. **Narrow the focus.** As we noted above, we study research mentoring. When we are deciding on what tools to use in our research, we need to think carefully about what aspect(s) of mentoring we think are most important to interrogate. The construct of mentoring in URSCI can be conceptualized as the quality of the mentoring relationship, including whether the mentee trusts the mentor, feels close to the mentor, likes the mentor, or otherwise finds the relationship satisfying. The construct of mentoring in URSCI also can be conceptualized by its functions. Usually, scholars define mentoring functions as career support (i.e., helping the mentee achieve their professional goals) and psychosocial support (i.e., building mentee confidence and supporting the mentee's personal development). We may be tempted to measure everything about mentoring, but focusing can help with selecting the right measure for the job.
4. **Examine the evidence.** When you find a measure, look for evidence regarding how it was developed and tested. Is there a clear definition of the construct it is supposed to measure? Has it been tested with your population of interest? Has it been tested in the context or situation you are interested in studying? Do the items seem like they are aligned with your construct? Try to put yourself in your research participants' brains. Do you think they would read, understand, and interpret the items accurately? Do you think they would vary in their responses in a way that would be informative for addressing your research question? Would the items be too easy, too hard, or otherwise off the mark?
5. **Be skeptical.** The time to think critically about your measure is before you have invested time, energy, or other resources in collecting data. It is tempting to jump straight to selecting an "off-the-shelf" measure without thinking about validity evidence. Spending some time

and effort reviewing validity evidence up front and selecting measures carefully based on this review will ultimately save a lot of time that might be spent collecting data that is uninformative because of validity concerns.

What If a Measure Is Not Quite Right?

As you follow the tips above, you may realize that an existing measure is not a perfect fit for your study. It might be designed for a different population, use language that does not match your context, or focus on slightly different aspects of the construct of interest to you. Rather than discarding the measure entirely or beginning from scratch, consider adapting the measure in a purposeful and transparent way. It may make sense to alter the response format, such as by converting frequency response options (e.g., from "never" to "every day") to agreement options (e.g., from "strongly disagree" to "strongly agree"). It might be helpful to simplify or clarify item wording for students of different ages or linguistic backgrounds. It might make sense to drop items altogether if they are clearly irrelevant or inappropriate, or reorder items to make them more natural or comfortable for participants to respond to.

Although these types of modifications may seem minor, they can have substantial consequences if made arbitrarily. Measures, especially those with robust validity evidence, are likely to have purposeful design elements, including how items are worded, how response options are formatted, how many items are included, and how responses are scored. Even small adjustments can affect how participants respond, which can affect all three elements of our assessment triangle. That said, making modifications to an existing quality measure will be easier than developing one from scratch. If you find yourself in this situation, be sure to document and justify all modifications, work to ensure modifications were either neutral or helpful for ensuring the validity of the measure, and report the modifications and validation efforts transparently, including acknowledgment of any limitations of these efforts. For example, if a measure has undergone significant word modifications or has limited validity evidence in support of the modifications, consider calibrating claims accordingly. This is not to say that modified measures can not or should not be used. Rather, the limitations of a modified measure and the strength of the validity evidence supporting such modifications should be thoughtfully balanced with the claims being made.

What If You Can Not Find the Right Measure?

In our combined greater than 25 years of educational research experience, there are only a few times we have been unable to find a suitable measure with at least some validity evidence. This is great news, because developing measures from scratch is extraordinarily challenging to do well. Measurement development is labor intensive and methodologically complex. In fact, we think it is some of the most difficult research to accomplish. Experts in educational measurement and assessment strongly discourage DIY (do-it-yourself) measures for all of the reasons we have outlined here (Bandalos, 2018; Flake, Pek, and Hehman 2017; McCoach, Gable, and Madura 2013). Rather than doing it yourself, we strongly encourage collaborating with experts who have experience with measurement development and validity. We collaborated with a psychometrician to develop the MURS (Limeri et al. 2024) and it still took six years from start to finish! Engaging with experts in measurement can strengthen the quality of the research you do and the conclusions that can be drawn from your work. We have compiled resources that we have found helpful for guiding our thinking about measurement development and validity to help you in this journey (Table 1).

Conclusion

In this commentary, we have emphasized the foundational, but sometimes ignored, role of measurement validity in education research. Validity is not a post hoc consideration, but a guiding principle that should inform decisions throughout the research process, beginning well before data collection. Using measures with multiple, robust forms of validity evidence is essential for generating trustworthy results. The phrase "garbage in, garbage out" reminds us that flawed measurement can undermine even the most sophisticated study design and analysis. To support readers in these efforts, we offer recommendations for identifying measures with robust validity evidence and highlight resources that have informed our own thinking about measurement. We hope this article serves as a starting point for further engagement with measurement in educational settings and encourages scholars and practitioners to prioritize validity in their work.

Data Availability

Citations are included in the text.

Institutional Review Board

This is a commentary, and no data were collected from human subjects.

Conflict of Interest

None.

Acknowledgments

This work was supported in part by National Science Foundation grant #2328692 and the Georgia Athletic Association Professorship for Innovative Science Education. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding organizations.

TABLE 1. Resources for Learning about Measurement Development and Validity

Topic	Suggested articles
Measurement development	• American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association. • Bandalos, Deborah L. 2018. *Measurement Theory and Applications for the Social Sciences*. New York: Guilford. • Clark, Lee Anna, and David Watson. 1995. "Constructing Validity: Basic Issues in Objective Scale Development." *Psychological Assessment* 7: 309–319. https://doi.org/10.1037/1040-3590.7.3.309 • Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281–302. https://doi.org/10.1037/h0040957 • DeVellis, Robert F., and Carolyn T. Thorpe. 2021. *Scale Development: Theory and Applications*. Thousand Oaks, CA: Sage. • Haynes, Stephen N., David Richard, and Edward S. Kubany. 1995. "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods." *Psychological Assessment* 7: 238–247. https://doi.org/10.1037/1040-3590.7.3.238 • Hinkin, Timothy R. 1998. "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires." *Organizational Research Methods* 1: 104–121. https://doi.org/10.1177/109442819800100106 • Kane, Michael T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000 • McCoach, D. Betsy, Robert K. Gable, and John P. Madura. 2013. *Instrument Development in the Affective Domain*. Vol. 10. New York: Springer. • Worthington, Roger L., and Tiffany A. Whittaker. 2006. "Scale Development Research: A Content Analysis and Recommendations for Best Practices." *Counseling Psychologist* 34: 806–838. https://doi.org/10.1177/0011000006288127.
Exemplars of measurement development and validation	• Alam, Irfanul, Karen Ramirez, Katharine Semsar, and Lisa A. Corwin. 2023. "Predictors of Scientific Civic Engagement (PSCE) Survey: A Multidimensional Instrument to Measure Undergraduates' Attitudes, Knowledge, and Intention to Engage with the Community Using Their Science Skills." *CBE–Life Sciences Education* 22: ar3. https://doi.org/10.1187/cbe.22- 02-0032 • Beymer, Patrick N., Melissa Ferland, and Jessica Kay Flake. 2022. "Validity Evidence for a Short Scale of College Students' Perceptions of Cost." *Current Psychology* 41: 7937–7956. https://doi.org/10.1007/s12144-020-01218-w • Couch, Brian A., Christian D. Wright, Scott Freeman, Jennifer K. Knight, Katharine Semsar, Michelle K. Smith, Mindi M. Summers, Yi Zheng, Alison J. Crowe, and Sara E. Brownell. 2019. "GenBio-MAPS: A Programmatic Assessment to Measure Student Understanding of *Vision and Change* Core Concepts across General Biology Programs." *CBE–Life Sciences Education* 18(1): ar1. https://doi.org/10.1187/cbe.18-07-0117 • Limeri, Lisa B., Nathan T. Carter, Franchesca Lyra, Joel Martin, Halle Mastronardo, Jay Patel, and Erin L. Dolan. 2023. "Undergraduate Lay Theories of Abilities: Mindset, Universality, and Brilliance Beliefs Uniquely Predict Undergraduate Educational Outcomes." *CBE–Life Sciences Education* 22(4): ar40. https://doi.org/10.1187/cbe.22-12-0250 • Sbeglia, Gena C., and Ross H. Nehm. 2020. "Illuminating the Complexities of Conflict with Evolution: Validation of the Scales of Evolutionary Conflict Measure (SECM)." *Evolution: Education and Outreach* 13: 23. https://doi.org/10.1186/s12052-020-00137-5

References

Alam, Irfanul, Karen Ramirez, Katharine Semsar, and Lisa A. Corwin. 2023. "Predictors of Scientific Civic Engagement (PSCE) Survey: A Multidimensional Instrument to Measure Undergraduates' Attitudes, Knowledge, and Intention to Engage with the Community Using Their Science Skills." *CBE–Life Sciences Education* 22(1): ar3. https://doi.org/10.1187/cbe.22-02-0032

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bandalos, Deborah L. 2018. *Measurement Theory and Applications for the Social Sciences*. New York: Guilford.

Beymer, Patrick N., Melissa Ferland, and Jessica Kay Flake. 2022. "Validity Evidence for a Short Scale of College Students' Perceptions of Cost." *Current Psychology* 41: 7937–7956. https://doi.org/10.1007/s12144-020-01218-w

Clark, Lee Anna, and David Watson. 1995. "Constructing Validity: Basic Issues in Objective Scale Development." *Psychological Assessment* 7: 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Couch, Brian A., Christian D. Wright, Scott Freeman, Jennifer K. Knight, Katharine Semsar, Michelle K. Smith, Mindi M. Summers, Yi Zheng, Alison J. Crowe, and Sara E. Brownell. 2019. "GenBio-MAPS: A Programmatic Assessment to Measure Student Understanding of *Vision and Change* Core Concepts across General Biology Programs." *CBE–Life Sciences Education* 18(1): ar1. https://doi.org/10.1187/cbe.18-07-0117

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52: 281–302. https://doi.org/10.1037/h0040957

DeVellis, Robert F., and Carolyn T. Thorpe. 2021. *Scale Development: Theory and Applications*. Thousand Oaks, CA: Sage.

Flake, Jessica K., Jolynn Pek, and Eric Hehman. 2017. "Construct Validation in Social and Personality Research: Current Practice and Recommendations." *Social Psychological and Personality Science* 8: 370–378. https://doi.org/10.1177/1948550617693063

Glaser, Robert, Naomi Chudowsky, and James W. Pellegrino, eds. 2001. *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academies Press.

Gormally, Cara, Peggy Brickman, and Mary Lutz. 2012. "Developing a Test of Scientific Literacy Skills (TOSLS): Measuring Undergraduates' Evaluation of Scientific Information and Arguments." *CBE–Life Sciences Education* 11: 364–377. https://doi.org/10.1187/cbe.12-03-0026

Haynes, Stephen N., David Richard, and Edward S. Kubany. 1995. "Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods." *Psychological Assessment* 7: 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hinkin, Timothy R. 1998. "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires." *Organizational Research Methods* 1: 104–121. https://doi.org/10.1177/109442819800100106

Kane, Michael T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000

Limeri, Lisa B., Muhammad Zaka Asif, Benjamin H. T. Bridges, David Esparza, Trevor T. Tuma, Daquan Sanders, Alexander J. Morrison, et al. 2019. "'Where's My Mentor?!' Characterizing Negative Mentoring Experiences in Undergraduate Life Science Research." *CBE–Life Sciences Education* 18(4): ar61. https://doi.org/10.1187/cbe.19-02-0036

Limeri, Lisa B., Nathan T. Carter, Riley A. Hess, Trevor T. Tuma, Isabelle Koscik, Alexander J. Morrison, Briana Outlaw, Kathren Sage Royston, Benjamin HT Bridges, and Erin L. Dolan. 2024. "Development of the Mentoring in Undergraduate Research Survey." *CBE—Life Sciences Education* 23(2): ar26. https://doi.org/10.1187/cbe.23-07-0141

McCoach, D. Betsy, Robert K. Gable, and John P. Madura. 2013. *Instrument Development in the Affective Domain*. Vol. 10. New York: Springer.

Sbeglia, Gena C., and Ross H. Nehm. 2020. "Illuminating the Complexities of Conflict with Evolution: Validation of the Scales of Evolutionary Conflict Measure (SECM)." *Evolution: Education and Outreach* 13: 23. https://doi.org/10.1186/s12052-020-00137-5

Shaffer, Justin F., Julie Ferguson, and Kameryn Denaro. 2019. "Use of the Test of Scientific Literacy Skills Reveals That Fundamental Literacy Is an Important Contributor to Scientific Literacy." *CBE–Life Sciences Education* 18(3): ar31. https://doi.org/10.1187/cbe.18-12-0238

Worthington, Roger L., and Tiffany A. Whittaker. 2006. "Scale Development Research: A Content Analysis and Recommendations for Best Practices." *Counseling Psychologist* 34: 806–838. https://doi.org/10.1177/0011000006288127

Trevor T. Tuma
University of Georgia, trevor.tuma@uga.edu

Trevor Tuma is a National Science Foundation STEM education postdoctoral research associate at the University of Georgia. Trained as a tree biotechnologist, he has shifted his research focus to generating knowledge on how educational environments can best support the growth of students. Tuma's research examines the individual, interpersonal, contextual, and organizational variables that support and hinder students' development and success during their educational and research training experiences.

Erin Dolan is a professor of biochemistry and molecular biology, Josiah Meigs Distinguished Professor, and Georgia Athletic Association Professor of Innovative Science Education at the University of Georgia. As a graduate student, Dolan volunteered in K–12 schools, which inspired her pursuit of a career in biology education research. She teaches introductory biology, and her research group, the SPREE Lab, works to delineate features of undergraduate and graduate research that influence students' career decisions.