# A COMPENDIUM OF SCALES for use in the SCHOLARSHIP OF TEACHING AND LEARNING

EARLY CAREER PSYCHOLOGIST COMMITTEE • SOCIETY FOR THE TEACHING OF PSYCHOLOGY •

ECP

EDITED BY:
Rajiv S. Jhangiani
Jordan D. Troisi
Bethany Fleck
Angela M. Legg
Heather D. Hussey

## Acknowledgements

We first thank Jennifer Stiegler-Balfour for her contributions during the conceptualization of this e-book. We also thank the following research assistants for their assistance: Tamaki Sasa, May Reinert, Jonathan Vargas, Natalie Clark, Nadya Hall, Misael Martinez, and Melissa Adolphe from the Pace University SHARPP Lab, and Rebecca Deutschmann and Emilie Lang from the KPU Social & Political Cognition Lab.

Cover image: "Measure a thousand times, cut once" by Sonny Abesamis (CC-BY 2.0)

## Table of Contents

# Chapter 1: Introduction

Rajiv S. Jhangiani[1], Jordan D. Troisi[2], Bethany Fleck[3], Angela M. Legg[4], and Heather D. Hussey[5]

[1]Kwantlen Polytechnic University, [2]Sewanee: The University of the South, [3]Metropolitan State University of Denver, [4]Pace University, [5]Northcentral University

The scholarship of teaching and learning (SoTL) has increased in both prevalence and profile during the past decade (Bishop-Clark & Dietz-Uhler, 2012; Gurung & Landrum, 2015; Gurung & Wilson, 2013). Over this time, SoTL work has become more methodologically rigorous and more accepted by university administrators as valid and valuable products of scholarship. Given its strong empirical foundation and long history of basic research such as cognitive, learning, behavioral, and social, psychology as a discipline is especially well-positioned to lead investigations into practices that enhance the effectiveness of teaching and learning. With a stated mission to "promote excellence in the teaching and learning of psychology," the Society for the Teaching of Psychology (STP) has been at the forefront of this movement within our discipline. STP has supported SoTL by awarding grants (e.g., the SoTL grant), developing demonstrably effective teaching resources (e.g., instructional resource awards), organizing conferences and meetings (e.g., the Annual Conference on Teaching), and effectively disseminating research findings (e.g., publication in its flagship journal *Teaching of Psychology*). This e-book is intended to further support these efforts by providing a valuable resource that facilitates the location, evaluation, selection, and (where necessary) development of psychometrically sound scales for the many traditional areas of focus within SoTL. In doing so, this compendium will achieve the broader goal of raising the scientific standards of evidence-based teaching and learning.

As editors of this e-book, we—the members of the Society for the Teaching of Psychology's Early Career Psychologists (ECP) committee—identified the relevant topic areas and invited well-established SoTL researchers within those areas to contribute chapters. As ECPs, we recognized the need to serve new faculty members and, in particular, the importance of focusing on pedagogy while facilitating SoTL. However, although this e-book is clearly helpful to those just getting started in this area of research, it will be equally valuable to seasoned researchers. SoTL research covers a broad range of topics including critical thinking, metacognition, professor-student relationships, and student perceptions of learning and teaching. This compendium covers each of these topics, along with many others that are at the forefront of SoTL research. Whereas a veteran researcher might be familiar with some of these areas, they will still benefit from learning more about others, as well as potentially new SoTL tools.

## Organization of this E-Book

Organized by topic, this compendium contains scale descriptions, validation information (if available), and references so scholars can examine past research that used each scale. In addition, the authors—each well established within their area of focus—provide advice on choosing appropriate scales, developing scales, and the types of scales the SoTL literature still needs.

The first section of this e-book focuses on the selection, use, development, and validation of scales. In Chapter 2, Regan Gurung discusses several best practices concerning scale use and choice, including identifying relevant psychological variables that influence learning and using published scales instead of cobbling together a non-validated measure. The chapter concludes with a review of criteria for selecting an appropriate scale and (where no published scale is available) writing your own items. In Chapter 3, Andrew Christopher complements the best practices chapter with advice on how to select the right scale, from his perspective as the current editor of *Teaching of Psychology*.

The next two chapters focus specifically on scale development and validation. In Chapter 4, authors Heather Hussey and Tara Lehan provide a brief, accessible guide to the scale development process. They succinctly describe the early stages of scale development such as conducting literature reviews, creating items, pilot testing, and revising items. They also provide an excellent summary of common reliability and validity tests, which will prove particularly useful to anyone new to the scale validation process (or even if you just need a refresher).

Section 1 concludes with a commentary on the state of scale validation in SoTL research (Chapter 5), in which Georjeanna Wilson-Doenges provides some exemplars of ways in which SoTL researchers have adopted the best practices for scale validation while operating within the common constraints of sample size, class composition, and semester length; all of which are perennial issues among SoTL researchers.

The chapters in Section 2 make up the bulk of the e-book and present a topical selection of scale information:

In Chapter 6, Pam Marek, Adrienne Williamson, and Lauren Taglialatela discuss the measurement of student learning and self-efficacy. The authors describe both formative assessments (e.g., classroom assessment techniques) and summative assessments (e.g., writing assignments and standardized tests), before concluding with a review of measures of perceived learning and perceived self-efficacy.

The measurement of critical thinking skills is addressed in Chapter 7 by Eric Landrum and Maureen McCarthy, who review mainstream measures of critical thinking that are specific to psychology as well as broad-based general measures. This chapter concludes with a set of recommendations for how SoTL researchers might balance the desires for efficiency and validity in the measurement of this complex construct.

Measures of student engagement toward coursework at both macro- and micro-levels are reviewed in Chapter 8 by Kevin Zabel and Amy Heger. The latter includes descriptions of measures of student interest, student engagement, as well as ancillary measures such as grit and boredom; whereas the former includes more general measure (e.g., national survey of student engagement).

Lori Simons systematically examines 21 quantitative measures of service learning and civic engagement in Chapter 9, including measures of service impacts on community, faculty perceptions of service, and service impacts on students. Along with distinguishing between scales with and without psychometric evidence, Simons provides advice for educators interested in measuring civic engagement as a student learning outcome.

Students' epistemological beliefs are the focus of Chapter 10, written by Kelly Ku. By providing a theoretical background as well as common measures to assess how students conceptualize the nature of knowledge, Ku addresses questions such as: Do students believe knowledge is relative and changeable or do they view it in absolute terms? How do they view the perspectives and opinions or authorities and when are they able to see ambiguity in knowledge representations?

In Chapter 11, Kristin Layous, S. Katherine Nelson, and Angela M. Legg propose that students' psychological well-being is an essential factor in understanding their experience in the classroom. The authors provide an overview of scales that assess different aspects of well-being, including both positive (e.g., life satisfaction, subjective well-being, meaning in life) and negative (e.g., general and domain-specific stress and anxiety) aspects.

Professor-student relationships are the focus of Chapter 12, contributed by Jenna Meyerberg and Angela M. Legg. These authors begin by providing a conceptual framework for understanding the positive impact on student outcomes of immediacy behaviors, professor-student rapport, and the learning alliance, before reviewing measures of all three constructs.

In Chapter 13, Claire Kirk, Jessica Busler, Jared Keeley, and William Buskist review several approaches to assessing professor efficacy, including by collecting student and peer feedback and evaluating teaching materials, before providing an in-depth examination of the Teacher Behaviors Checklist as an exemplar tool.

The e-book concludes with a thoughtful (and in our minds humorous) chapter by Aaron Richmond, who highlights the types of scales the literature still needs (Chapter 14), including measures of metacognition and learning strategies, scales to assess syllabi and model teaching characteristics, as well as valid behavioral measures to accompany the many self-report SoTL scales. Richmond concludes with a series of five recommendations that represent a call to action for SoTL research.

By highlighting the most prominent scales on a range of topics, this e-book serves as an important reference guide within the scholarship of teaching and learning. Furthermore, the practical advice that each expert author has provided will surely enhance the rigor of scholarly work to follow. We hope that this compendium provides a useful tool for early career psychologist and seasoned researchers alike.

## Suggestions for Reading this E-Book

The organization of this e-book allows readers to pick their own starting point. Whereas the Chapters in Section 1 provide excellent and practical advice on aspects of the scale selection and development process, each chapter in Section 2 provides a standalone account of that topic's validated scales, permitting readers to use each chapter as its own reference guide. In tracking down these existing scales, readers may be pleased to discover that within the references section of each chapter, an asterisk is placed next to the reference for each scale discussed within that chapter. Our hope is that emerging and well-established SoTL researchers alike will find great value in the general and topic-specific guidance within this e-book regarding the best practices in scale development, validation, and use.

For educators, this book can also serve as an excellent supplementary text for courses such as tests and measurements, research methods, and educational assessment. Many of the chapters provide accessible descriptions that graduate and even undergraduate audiences will appreciate. Very few resources exist that provide a "compare and contrast" presentation of a variety of measurement tools, especially within the SoTL literature. The chapters in this e-book provide students with exemplars of the scale development and validation process while offering a current account of how SoTL researchers assess different constructs. As an additional tool, many of our authors also provide pedagogical suggestions for teaching students about the scales they discuss.

Finally, as our authors note in their individual chapters, many gaps still exist in the development of validated scales for SoTL use. Thus, our final suggestion for readers is to take inspiration from the extant literature and take up the challenge of adding to the field's increasing reliance on validated measures. It is our hope that, upon developing and validating these much-needed scales, this e-book will require a second edition to update chapters, add new chapters, and reconsider the state of scale validation and use in SoTL work.

## References

Bishop-Clark, C. & Dietz-Uhler, B. (2012). *Engaging in the scholarship of teaching and learning: A guide to the process, and how to develop a project from start to finish.* Hemdon, VA: Stylus Publishing.

Gurung, R. A. R., & Landrum, R. E. (2015). Editorial. *Scholarship of Teaching and Learning in Psychology, 1*(1), 1-6.

Gurung, R. A. R., & Wilson, J. H. (Eds.). (2013). *Doing the scholarship of teaching and learning: Measuring systematic changes to teaching and improvements in learning.* San Francisco: Jossey-Bass.

# Section 1: Choosing, Using, Developing, and Validating Scales

# Chapter 2: Best Practices in Scale Use in SoTL

Regan A. R. Gurung

University of Wisconsin-Green Bay

"Measure what is measurable, and make measurable what is not so" (Galileo Galilee)

Measurement is at the core of robust scholarship of teaching and learning (SoTL, Gurung & Landrum, 2012). Is my teaching effective? Are my students learning? What predicts student learning? Most empirical projects and quantitative approaches to pedagogical research involve measurement and if there is an existing scale to measure what you want to, why reinvent the wheel? In this chapter I will review some key issues to keep in mind when picking and using scales. I will overview construct validity so pedagogical researchers remember what to look for in a scale and I will touch on best practices in writing your own items.

Before diving in, it is important to clarify some terminology usage. In doing descriptive studies researchers measure many different variables and use a number of different research designs (Bartch, 2013; Schwartz & Gurung, 2012). In descriptive studies, we want to get a picture of what is going on in classroom or teaching (e.g., How many students can list the major approaches to psychology?). In correlational studies we want to measure associations between variables (e.g., Are the students who took more practice quizzes on the chapter better at listing the major approaches to psychology?). In experimental designs we want to collect evidence to see if the changes we implemented, to assignments, lectures, or design, resulted in increases in learning (e.g. Did the group of students who watched my special Intro to Psych video list more of the major approaches to psychology?). For each of these major designs, descriptive, correlational, and experimental, there are a number of ways to measure variables.

Aside from observation or conducting focus groups, which require different forms of measurement, the majority of pedagogical research in psychology involves assessing students' attitudes and behaviors with surveys and measuring their learning with exams, quizzes, and other performance outcomes. Surveys or questionnaires are general terms and each survey can contain many items or questions. Sometimes the researcher generates the questions themselves. Often the researcher uses a preexisting published scale. A survey can hence comprise of many scales, many questions (i.e., never before been used or published), or a combination of both. There are best practices to both the selection and use of scales, and the creation of your own items or questions. I will first cover some best practices regarding scale use and choice and then briefly review pointers for item construction.

## Measure the Usual Suspects: Key Variables in Learning

There is a long history of research on teaching and learning (Gurung & Schwartz, 2012). A wide range of disciplines study the factors influencing learning, with psychology and education playing a major role, although there are also significant studies of the neuroscience of learning (Doyle & Zakrajsek, 2013). The average pedagogical researcher will not be popping students into fMRI machines to observe which nerve cells or areas of the brain fire during thinking, so a review of the biological forms of measurement can be circumvented. In contrast, most psychological SoTL studies use surveys of some form or the other. Although you may need to write some specific questions to get at unique research questions you have, there are a wide variety of existing scales that help measure some common learning variables.

A large body of academic literature identifies key factors influencing learning (Credé & Kuncel, 2008; National Research Council, 2001; Robbins et al., 2004) and what learning techniques work well (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). Study techniques in particular are important because they predict academic performance over and above standardized tests and previous grades (Gurung, 2015). There are a host of key psychological variables that are important as well. For example, it is clear that achievement motivation, goals, social involvement, and perceived social support are all positively associated with students' academic performance (Hattie, 2015; Robbins et al., 2004). In particular, factors such as effort, ability, habits, and self-efficacy are strongly related to academic performance (Credé & Kuncel, 2008; Komarraju & Nadler, 2013). Not surprisingly, current college GPA and other cognitive indicators such as ACT scores and high school GPA also predict learning in a university setting (Komarraju, Ramsey, & Rinella, 2013). Measuring ACT scores and GPA is easy, or at least already done for you. It is getting at the other variables that is the challenge.

The good news is that there are scales to measure each of the different variables reviewed above. A best practice then is to be sure you identify key psychological variables that are relevant to your study and then use the associated scale. Some scales representing common control variables in SoTL research are listed below, with some of them including sample items from the scale:

- Academic Self efficacy (Gredler & Schwartz, 1997; Zajacova, Lynch, & Ependshade, 2005)
    - With the prompt "please answer how stressful these tasks are for you," participants respond to items such as "keeping up with the required readings," "doing well on exams," "participating in class discussions," and "understanding college regulations." Participants also respond the same items with the prompt "how confident are you that you can successfully complete these tasks."
- Self regulation (Brown, Miller, & Lawendowski, 1999)
- Critical thinking and motivation (Valenzuela, Nieto, & Saiz, 2011)
- Academic locus of control (Curtis & Trice, 2013)
    - Participants complete a True or False response to items such as "I have largely determined my own career goals," "There are some subjects in which I could never do well in," "Studying every day is important," and "I am easily distracted."
- Metacognition (Schraw, & Dennison, 1994; Tuncer & Kaysi, 2013; Wells & Cartright-Hatton, 2004)
    - Participants respond to items such as "I have a poor memory," "I need to worry in order to do work well," "I am constantly aware of my thinking," and "it is bad to think certain thoughts."
- Motivated strategies for learning (Pintrich, Smith, Garcia, & McKeachie, 1993)

    - Participants respond to items such as “I prefer class work that is challenging so I can learn new things,” “I think I will be able to learn what I learn in this class in other classes,” “I expect to do very well in this class,” and “I worry a great deal about tests.”
- Depth of processing (Enwistle, 2009)
- Lifelong Learning Scale (Wielkiewicz, & Meuwissen, 2014)
- Procrastination (Tuckman, 1991)
    - Participants respond to items such as “I postpone starting in on things I don’t like to do,” “I delay making tough decisions,” “I get right to work, even on life’s unpleasant chores,” and “when something is not worth the trouble, I stop.”
- Study Behaviors/Process (Fox, McManus, & Winder, 2001; Gurung, Weidert, & Jeske, 2012)
- Textbook Assessment and Utility Scale (Gurung & Martin, 2011)

## Why Should You Use Published Scales?

The easy answer is that it saves you a lot of work. Measurement is not something done casually or quickly. Developing valid and reliable measures is a complex and involving process (Noar, 2003; Sosu, 2013), so the simple reason to use published scales is that the hard work has been done for you. Robust scale development involves multiple studies and iterations. A published scale has been through the peer review process and the associated checks and balances. Furthermore, other researchers will have also used that published scale providing you with additional information about the construct. Correspondingly, you have the use of a scale that should satisfy two important criteria for a good scale: validity and reliability.

Validity and reliability are essential concepts in measurement. How well have you measured your outcomes and predictors? How likely are your measures to provide the same results when used again? Validity in general refers to the extent to which the scale measures what it is supposed to measure (Anastasi, 1988). There are many different forms of validity (e.g., external, statistical, internal) but when using scales we care most about construct validity. Construct validity refers to the idea that a scale is measuring what we think it is (Morling, 2015).

Even when you use published scales, it is prudent to be aware and comfortable with the main forms of construct validity so you can assess the quality of the scale. Whereas some forms of construct validity are subjective in nature, the majority of them are objective in nature and easily assessed by statistical rubrics. Subjective forms of construct validity include face validity and content validity. A scale with good face validity looks like it is measuring what it is supposed to (you can see how subjective this is). Are the items plausible ways to get at the underlying concept? Content validity gets at whether a measure encompasses all parts of the underlying concept. Are the different items getting at all the different parts of concept? To have adequate content validity, a scale should have items at all the different parts of a concept. Often, scales have different sub-components or subscales in order to fully capture concepts.

Objective forms of construct validity include criterion (concurrent and predictive), divergent, and convergent validity. Criterion validity assesses if the scale is related to the outcome it is measuring. If you use a good measure of procrastination (Tuckman, 1991) you want to see it measuring a behavioral measure of the outcome (i.e., the criterion) such as time to turn in an assignment. The relationship of the scale to an outcome can be measured at one testing time (concurrent validity) or to an outcome at a later time (predictive validity). In convergent validity, the scale correlates with similar scales or measures. For example, the measure of procrastination should correlate with measures of conscientiousness (in a negative way). In divergent validity, the scale should not correlate with dissimilar scales or measures. Procrastination need not show a high relationship with extraversion. When you use a published scale, all these forms of validity are suitably established.

Good scales are also reliable. Reliability refers to consistency of measurement (Anastasi, 1988). In the context of scale development and use, two major forms of reliability are important. Test-retest reliability assesses if you will get consistent scores every time you use the measure. If I measure self-efficacy today, will that same self-efficacy scale provide a similar result in a week or a month (assuming no intervention to change self efficacy)? Internal reliability assesses the extent to which your participant provides a consistent pattern of answers. If a scale has 10 items, do participants answer in a consistent pattern across all items even if the items are worded differently? Good scales have high test-retest and internal reliability and these two forms are what you need to look for when selecting measures. Both forms of reliability are measured using forms of the correlation coefficient (*r*). Correlations closer to 1 suggest high reliability. Internal reliability is calculated using a correlation based statistic called Cronbach's alpha (easily generated by statistical packages such as SPSS). The related best practice then is to select measures where test-test reliability and internal reliability are both high, over .75 and .65, respectively.

## How Do You Pick Scales?

Before you take the trouble to write your own items it is well worth your time to use PsycINFO or other databases and see if a scale exists for what you want to measure. This e-book contains the major scales you need and many more exist. Some key points to keep in mind when you select scales:

- Remember that single items can be used to measure concepts but multiple item scales are better (DeVellis, 1991). You cannot calculate internal reliability with a single item.
- Look for scales with high internal reliability (Cronbach's alpha > .65), that have shown criterion, convergent, and divergent validity when possible, and preferably have high face and content validity. You will find this information in a Method section.
- Be sure to use all items from a published scale. Do not pick and chose the items you think best fit your study as scale validity and reliability is based on you using all the items. And if a scale does have subscales, make sure you do not present all subscale items in one group even though they may appear together in the original publication. You should however feel free to mix up the order of the items.
- Some scales, however, such as for measurement of the Big Five personality traits, have short-form versions that provide good alternatives to full versions.

- Be aware of whether the scale is unidimensional (giving you one total score) or multidimensional (giving you different subscale scores). You cannot use a total score if there are subscales.
- When scoring a scale, be sure you have reverse coded items as needed.
- Be cognizant of how long the scale is, as many long scales in a single study can cause participant fatigue and threaten the internal validity of your study (i.e., are there possible alternative explanations for the results/changes in outcome variables?).
- Order effects occur when participants' responses to scales later in the study may not be as reliable as responses early in the study.
- Although using scales exactly as published (same instructions, same response scales) is optimal, you may sometimes have a need to modify scales slightly to suit your purposes. Note that the published validity and reliabilities may no longer hold.

## Best Practices in Writing Your Own Items

If there is no published scale for your purposes you then have to write your own items to get the information you need. Although a full exposition of scale or item construction is beyond the scope of this chapter (see Berk, 2006 for a great treatise on the same), there are some easy key points to keep in mind.

Items can be of three main forms. You can ask open-ended questions (Describe your learning experience in this class?), forced-choice items (Which of the following assignments did you best learn from?), or Likert scale items (How much do you agree with the following on a scale ranging from strongly agree, agree, neither agree nor disagree, disagree, strongly disagree?). Sometimes participants can be asked to respond to a question using a numeric scale anchored by adjectives (Rate the easiness of this course on a scale ranging from 1 = Easy to 5 = Hard). This last type is called a semantic difference format.

Culling together best practices from a variety of sources (Berk, 2006; DeVellis, 1991; Morling, 2015; Noar, 2003) the following are suggestions of strong, concrete guidelines to keep in mind. Your items should be:

- *Clear and concise*: Each statement should convey a specific idea that is easily interpretable by the respondent. Try to limit yourself to about 10 words or less.
- *Devoid of slang or jargon, or double negatives*: Avoid language, words, or terms that are specific to your field or expertise and that may not be meaningful (or known) to the average reader.
- *Unambiguous:* Your meaning should be clear. It is a good idea to have friends read your items and tell you what they think it means. This enables you to confirm your intent.
- *Both positive and negative (to avoid responses sets):* Write some questions so that the accurate answer uses the lower end of the scale while other questions require answers at the other end of the scale. For example both the following items measure high self-esteem, but the second is negatively worded: "I feel good about myself" "At times I feel like a failure."
- *Gender and culturally sensitive:* Items should be applicable to all respondents. Try and write generic statements as it pertains to sex, ethnicity, and culture in general.

- *At an easy reading level:* If you can use simple words, do it. This is not a place to display your impressive vocabulary. As a rule of thumb, use a high school reading level.
- *Not leading or nudging the participant to a certain answer:* Be fair and balanced. Someone reading the item should not be able to guess what the researcher's hypothesis may be.
- *Single-barreled*: Item should only ask one question; each statement should have only one complete behavior, thought, or concept.
- *Have response scales that*:
    - Do not exceed 7 points on the scale.
    - Have each point labeled (e.g., 1 = strongly agree; 2 = agree).
    - Are even numbered (to avoid fence sitting).
    - Do not use numbers on the scale but only the word labels or letter abbreviation (e.g., "SA" for "Strongly Agree").
    - Avoid the "Not Applicable" option.

## Conclusion

Pedagogical researchers of today can be grateful for a wide range of journals that showcase useful scales. Together with *Scholarship of Teaching and Learning in Psychology, Teaching of Psychology,* and *Psychology of Learning and Teaching*, a number of journals across disciplines, such as *The International Journal of the Scholarship of Teaching and Learning,* and the *Journal of Scholarship of Teaching and Learning,* provide scales for use. There are some basic best practices for using scales as described above. Being comfortable with these practices will ensure robust SoTL. Whereas you can make great contributions to the literature and the study of teaching by developing scales to fill in the gaps that exist, that is a whole other ball of wax. There are a significant number of well-validated scales in the published literature. It a good idea for you to seek out a scale measuring the precise conceptual variable you are interested in, and only write items if they are needed.

# References

References marked with an asterisk indicate a scale.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Bartsch, R. A. (2013). Designing SoTL studies-Part I: Validity. In R. A. R. Gurung & J. Wilson (Eds.) *New Directions For Teaching & Learning* (pp. 17-33). doi:10.1002/tl.20073

Berk, R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians.* Sterling, VA: Stylus.

*Brown, J. M., Miller, W. R., & Lawendowski, L. A. (1999). The self-regulation questionnaire. In L. VandeCreek, & T. L. Jackson (Eds.), *Innovations in clinical practice: A sourcebook* (pp. 281–292). Sarasota, FL: Professional Resource Press/Professional Resource Exchange.

Credé, M., & Kuncel, N. R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, *3*, 425-453. doi:10.1111/j.1745-6924.2008.00089.

*Curtis, N. A., & Trice, A. D. (2013). A revision of the academic locus of control scale for college students. *Perceptual & Motor Skills, 116,* 817-829. doi: 10.1037/t32978-000

DeVellis, R. F. (1991). *Scale development: Theory and applications.* Thousand Oaks, CA: Sage.

Doyle T., & Zakrajsek, T. (2013). *The new science of learning: How to learn in harmony with your brain.* Sterling, VA: Stylus.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4-58. doi:10.1177/1529100612453266

*Entwistle, N. J. (2009). *Teaching for understanding at university: Deep approaches and distinctive ways of thinking*. Basingstoke, United Kingdom: Palgrave Macmillan.

*Fox, R. A., McManus, I. C., & Winder, B. C. (2001). The shortened study process questionnaire: an investigation of its structure and longitudinal stability using confirmatory factor analysis. *British Journal of Educational Psychology, 71,* 511-530. doi:10.1348/000709901158659

*Gredler, M. E. & Schwartz, L. S. (1997). Factorial structure of the self-efficacy for self-regulated learning scale. *Psychological Reports, 81,* 5-77. doi:10.2466/PR0.81.5.51-57

Gurung, R. A. R., & Landrum, R. E. (2012). Assessment and the Scholarship of Teaching and Learning. In D. Dunn, S. C. Baker, C. M. Mehrotra, R. E. Landrum, & M. McCarthy (Eds.) *Assessing Teaching and Learning in Psychology: Current and Future Perspectives*.

*Gurung, R. A. R., & Martin, R. (2011). Predicting textbook reading: The textbook assessment and usage scale. *Teaching of Psychology, 38,* 22-28. doi:10.1177/0098628310390913

Gurung, R. A. R., & Schwartz, B. (2009). *Optimizing teaching and learning: Pedagogical Research in Practice.* Wiley Blackwell Publishing. London.

*Gurung, R. A. R., Weidert, J., & Jeske, A. S. (2010). A closer look at how students study (and if it matters). *Journal of the Scholarship of Teaching and Learning, 10,* 28-33.

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology, 1*, 79-91. doi:10.1037/stl0000021

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences*, *25,* 67-72. doi:10.1016/j.lindif.2013.01.005

Komarraju, M., Ramsey, A., & Rinella, V. (2013). Cognitive and non-cognitive predictors of college readiness and performance: Role of academic discipline. *Learning and Individual Differences*, *24,* 103-109. doi:10.1016/j.lindif.2012.12.

Lent, R. W., Brown, S. D., and Larkin, K. C. (1986). Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology 33*, 265–269. doi: 10.1037/0022-0167.33.3.265

Morling, B. (2015). *Research methodology in psychology: Evaluating the world of information (2e).* New York: Norton.

National Research Council. (2000). *How people learn: Brain, mind, experience and school.* Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., & Glaser, R. (Eds.). Washington, DC: National Academy Press.

Noar, S. M. (2003). The role of structural equation modeling in scale development. *Structural Equation Modeling, 10,* 622-647. doi: 10.1207/S15328007SEM1004

*Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, G. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire. *Educational and Psychological Measurement, 53,* 801-813*.* doi: 10.1177/0013164493053003024

Robbins, S., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychological and study skill factors predict college outcome? A meta-analysis. *Psychological Bulletin, 130,* 261–288.

*Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475. doi: 10.1006/ceps.1994.1033

Schwartz, E., & Gurung, R. A. R., (2012). *Evidence-based teaching in higher education.* Washington, DC: American Psychological Association.

*Solberg, V. S., O'Brien, K., Villareal, P., Kennel, R., and Davis, B. (1993). Self-efficacy and Hispanic college students: Validation of the college self-efficacy instrument. *Hispanic Journal of Behavioral Sciences 15,* 80–95. doi: 10.1177/07399863930151004

*Sosu, E. M. (2013). The development and psychometric validation of a critical thinking disposition scale. *Thinking Skills and Creativity, 9,* 107-119. doi:10.1016/j.tsc.2012.09.002

*Tuckman, B. W. (1991). The development and concurrent validity of the procrastination scale. *Educational and Psychological Measurements, 51, 473-480.* doi: 10.1177/0013164491512022

*Tuncer, M. & Kaysi, F. (2013). The development of the metacognitive thinking skills scale. *International Journal of Learning and Development, 3,* 70-76. doi: 10.5296/ijld.v3i2.3449

*Valenzuela, J., Nieto, A. M., & Saiz, C. (2011). Critical thinking motivational scale: a contribution to the study of relationship between critical thinking and motivation. *Electronic Journal of Research in Educational Psychology, 9,* 823-848.

*Wells, A. & Cartwright-Hatton, S. (2004). A short form of the metacognitions questionnaire: properties of the MCQ-30. *Behaviour Research and Therapy, 42,* 385-396. doi: 10.1016/S0005-7967(03)00147-5

*Wielkiewicz, R. M. & Meuwissen, A. S. (2014). A lifelong learning scale for research and evaluation of teaching and curricular effectiveness. *Teaching of Psychology, 41,* 220-227. doi: 10.1177/0098628314537971

*Zajacova, A., Lynch, S. M., & Epenshade, T. J. (2005). Self-efficacy, stress, and academic success in college. *Research in Higher Education, 46,* 677-706. doi: 10.1007/s11162-004-4139-z

# Chapter 3: Selecting the Right Scale: An Editor’s Perspective

Andrew N. Christopher

Albion College

Before I begin to offer my perspectives on selecting the appropriate scale for a research project, it is important that I share some of my academic background because it clearly colors my views in this chapter. In graduate school, I was in a program in which the graduate students in social psychology interacted regularly with graduate students in counseling psychology, cognitive psychology, and neuroscience. Because it was commonplace for such interactions to occur, I figured that was the way researchers in different areas worked; that is, they frequently interacted with each other. Indeed, 16 years after receiving my Ph.D., my best friends from graduate school with whom I still work are from areas of training different than mine. I have learned during the past 16 years that such “cross-area” communication is in fact not as normal as I perceived it to be. Since completing graduate school, I have taught at a small liberal arts college in which faculty in all disciplines regularly interact on scholarly work related to both teaching and research. Some of the best course material I have developed for my industrial/organizational (I/O) psychology class originated from numerous discussions with a colleague from my college’s history department. Some of my own research focuses on individual differences, so it is rooted strongly in a combination of I/O and personality, both of which rely heavily on the use of scales, a practice that is increasingly being scrutinized by people in both areas. With this background in mind, please allow me to provide my insights into selecting the right scale for a research project.

## A (Needlessly) Great Divide

It is not uncommon for there to be great divides between researchers in seemingly related areas of research. For example, it might make intuitive sense for researchers in personnel selection and performance evaluation to collaborate, particularly on applied types of projects. Having worked in both areas, I have been surprised how little collaboration or even simply communication there is between researchers on each side of this process in both academic and applied settings.[1]

Within the realm of educational-type research, Daniel and Chew (2013) drew a distinction between the learning sciences and SoTL, the former of which has “The goal….to better understand the cognitive and social processes that result in the most effective learning, and to use this knowledge to redesign classrooms and other learning environments so that people learn more deeply and more effectively” (Sawyer, 2006, p. xi). Its emphasis is on the application of research in basic areas of psychology, such as personality and cognitive, to educational settings. For example, Christopher, Furnham, Batey, Martin, Koenig, and Doty (2010) examined how work values may be compensatory mechanisms for lower levels of general intelligence. Certainly, such work could be applicable to almost any educational setting. In contrast to the learning sciences, much SoTL work “…typically reports the success of particular methods in a particular context and within a particular level of a specific discipline” (Daniel & Chew, 2013, p. 365). Indeed, SoTL tends to be quite specific in its focus and generalizability of its results. For example, McCabe (2015) examined the extent to which there is a generation effect for learner-created keyword mnemonics and real-life examples when learning neuroscience material in the introductory psychology course. This research is certainly appropriate for teachers of Introductory Psychology.

Daniel and Chew nicely described what they (and I) see as a divide between SoTL and learning sciences, with researchers tending to fall into one camp or the other, but rarely overlapping, despite the enormous desirability and potential for such overlap. In many ways, I see this divide as similar to the notion that researchers tend to get classified, rightly or wrongly, as either "basic" researchers or "applied" researchers (with the former typically being perceived as "better" researchers). Many perceive that there is often no middle ground, which I believe hinders the progress of both types of research.

Much like the divide between personnel selection and personnel evaluation researchers, or the divide between basic and applied researchers, the divide between SoTL and learning sciences researchers is, in my opinion, unfortunate. I believe that this e-book is a large step toward bridging this divide between the learning sciences and SoTL researchers. Indeed, much of the information it contains, particularly in Gurung's (2015) and Wilson-Doenges' (2015) chapters, is actually rooted in the learning sciences. For example, in chapter 2 of this e-book, Gurung listed 10 different scales that could be useful in SoTL research. In reality, all but two of these scales (the Lifelong Learning Scale and the Textbook Assessment and Utility Scale) are rooted in basic individual differences research. Specifically, researchers in the learning sciences have long tended to use scales that have undergone extensive psychometric work. For instance, the Need for Cognition Scale (Cacioppo, Feinstein, Jarvis, & Petty, 1996), which received a great deal of psychometric evaluation, has been used extensively in learning sciences research. At the time of publication of this chapter, a quick PsycINFO search on "need for cognition" and "learning science" revealed 46 hits in the past three years, with articles appearing in a broad spectrum of journals, such as *Learning and Individual Differences, Computers in Human Behavior, Creativity Research Journal*, and *Psychology of Women Quarterly*. A similar search with "scholarship of teaching and learning" substituted for "learning science" revealed but one hit. Of course, this examination relies on a sample of one. However, when I review for journals such as the ones listed previously, it seems as though using psychometrically vetted measurement tools is an expected practice.

In Chapter 5, Wilson-Doenges (2015) describes the current state of psychometric work on SoTL scales. As she notes, such work is becoming more prevalent and more rigorous in SoTL in recent years. It is essential that such work continue and be displayed in prominent SoTL outlets. SoTL is to applied research what the learning sciences are to basic research. This should not be a problem; however, extending the (what I believe to be grossly inaccurate) perception that basic research is more scientific and rigorous that applied research, SoTL work may be perceived as inferior to work in the learning sciences.

Why might psychometric issues receive less attention in SoTL work than in learning science work? I can offer only two reasons, and both are purely speculation. First, perhaps because of the relatively specific focus of its studies, it is difficult to establish the psychometric properties of a measurement tool. For instance, in McCabe's (2015) research presented earlier, the context of her work was limited to one course (Introductory Psychology). Therefore, for her research, psychometrically-established procedures and measures were difficult to come by. For Christopher et al.'s (2010) research on work values, it was much easier to locate psychometrically-tested measures.

In addition to the issue of research specificity, it could be the case that training in the learning sciences focuses more on psychometric issues than does training in SoTL. Similar to training in personality and I/O, both of which have been historically reliant on survey methodologies, learning science researchers may have taken required psychometric coursework in their graduate training. Graduate training in SoTL may focus more on developing classroom teaching and mentoring skills than on more traditional forms of scholarship. In his review of graduate training practices, Buskist (2013) offered a number of challenges and recommendations for the future of such work. Interestingly, training graduate students to perform SoTL was not mentioned. Indeed, SoTL seems like a logical extension of graduate training in teaching, whereas learning sciences implicitly assume scholarship is a part (if not the entirety) of graduate training. Thus, whereas researchers in the learning sciences have a relatively long tradition of emphasizing, at least implicitly, the importance of scale psychometrics, researchers in SoTL are only more recently beginning to more intentionally undertake the task of putting their measures to the psychometric test. Again, there are certainly SoTL articles that focus on issues of psychometrics (see Chapter 5 for some such examples), but in general, SoTL can benefit greatly from following the lead of learning science researchers and their emphasis on scale psychometric properties. In particular, to the extent that the learning sciences is perceived to be "more scientific" than SoTL, here may be an area to trim that perceptive divide.

## My Suggestions

As might be obvious from my tone, I am not happy about the divides between researchers who study similar topics. Here, I will try and suggests ways that SoTL can bridge its divide with the learning sciences and thus profit from work in not only the learning sciences, but in personality and I/O psychology as well.

In his chapter, Gurung (2015) presented a number of excellent arguments for using already-published scales. I cannot agree strongly enough with these suggestions. Most published scales will possess construct validity, or else they almost certainly would not have been published. As Gurung said, such psychometric work is not done casually or whimsically. To establish a measurement tool's psychometric properties is a painstaking process. Any scale that has made it to the point of publication in a respected peer-review journal did not get there by accident. It likely has something to offer and is a relatively "safe" bet, psychometrically speaking. In addition to the guidelines for selecting a published scale that Gurung suggested, allow me to add three other, somewhat interrelated, suggestions.

First, researchers should pay some attention the outlet in which a scale was originally published. Not every scale needs to be published in *Psychological Bulletin* to be a psychometrically sound tool. However, the quality of the outlet in which a scale appears does potentially say something about its psychometric quality.[2] Second, in addition to the original publication that contains the scale, look to see if subsequent work has somehow improved the original scale. For example, the Teacher Behaviors Checklist (TBC; Buskist, Sikorski, Buckley, & Saville, 2002) has been subjected to subsequent psychometrically-focused research (e.g., Keeley, Smith, & Buskist, 2006). The result? A scale that is better, psychometrically-speaking, than it otherwise would be[3]. Finally, do cited reference searches on the publication that contained the original scale and any subsequent research that amended and improved the original scale. If such cited reference searches reveal few or no citations to these sources, it is an indication, albeit an imperfect one, that perhaps the scale is not well-accepted in the scientific community.

With my personality and I/O worldview, I am noticing many scales that seem like they are measuring the same construct[4]. For example, a recent publication examined work ethic and GRIT as predictors of job satisfaction, turnover intention, and job stress (Meriac, Woehr, & Banister, 2015). Work ethic was defined as “a set of beliefs and attitudes reflecting the fundamental value of work” (p. 316). GRIT is the “perseverance and passion for long-term goals” and entails “working strenuously toward challenges, maintaining effort and interest….despite failure, adversity, and plateaus in progress (Duckworth, Peterson, Matthews, & Kelly, 2007, pp. 1087-1088). At first read, work ethic and GRIT may sound like closely related constructs, and in fact, they were correlated (*r* = .44; Meriac et al., 2015). Do we really need both of these measures? Indeed, Meriac and his colleagues suggested that in fact they do differentially predict job satisfaction, turnover intentions, and stress. Furthermore, Meriac and his colleagues found that there is incremental validity in using both of these measures. Indeed, at least within the context of I/O psychology, it does appear that there is value in having both of these measures. Researchers need to consider which one is appropriate to answer a given research question, or if fact it is worth using both of them. Ultimately there is no correct or incorrect choice; rather, it is imperative that some justification for the choice of a scale or scales is needed, particularly when there are seemingly overlapping possible scales to choose from.

Indeed, with so many scales available, look for those that have demonstrated incremental validity in prior research. In my opinion (and my opinion only), incremental validity seems to receive short shrift relative to other forms of validity in establishing the psychometric properties of a scale. The cynical part of me does wonder how many new scales actually predict outcomes above and beyond existing scales. The more-erudite part of me believes that new scales are indeed adding to what’s already available, and if such evidence has not been offered, it is incumbent on the scientific community to offer evidence such is the case.

In Chapter 14, Richmond (2015) describes the issue of relying on self-report data and not actual behavior, something inherent with any research area that relies on scales. In personality and I/O, there is great concern about common method bias. Common method bias occurs when the same general methodology (e.g., reliance exclusively on self-report data) is used to answer a research question. The trend in personality and I/O is to avoid common method bias to the extent possible. Richmond makes a number of excellent suggestions in his chapter on how to go about avoiding common method bias.

No matter what scale a researcher selects, it is helpful to reviewers if authors present an explicit understanding of limitations of a given measure. For example, Wilson-Doenges (2015) mentions the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003) to measure the Big Five. This instrument has amassed a whopping 1253 cited references on PsycINFO as of this writing. Therefore, the scientific community has enthusiastically embraced this scale. But with only 2 items to measure each factor, it is difficult to meet the "gold standard" of a .70 or stronger Cronbach's alpha. However, if there are other scales and measurements that are needed in a study, this may well be an acceptable tradeoff for avoiding participant fatigue. Perhaps reviewers and editors won't agree with this sentiment, but failing to acknowledge this tradeoff is sure to work against the researcher during the peer-review process.

Wilson-Doenges (2015) provides some exemplars of excellent psychometric work in SoTL. Here, I try to complement her suggestions with five examples and brief summaries of psychometric work from personality, I/O, and the learning sciences. These are all sources that I have used in my classes to demonstrate certain psychometric principles. One of the commonalities of each of these five models is that each one provides an excellent conceptual background to the construct(s) being measured in the article.

The first two exemplars are ones I've mentioned previously in this chapter. Duckworth et al. (2007) devised a scale to measure GRIT and conducted six studies to establish its psychometric properties. Across these six studies, they established the factor structure of their measure; its predictive validity by correlating it with lifetime schooling among people of identical age; investigated its incremental validity over the Big Five factors; assessed its convergent and divergent validity by correlating it with GPA and general intelligence; and evaluated forms of reliability including test-retest, internal, and item-total correlations.

Miller, Woehr, and Hudspeth (2002) provide another excellent example of scale development. Across six studies using undergraduate students, U.S. Air Force personnel, and employees from privately-owned companies, these researchers assessed the psychometric properties of the Multidimensional Work Ethic Profile. Specifically, they include discussions of the measure's factor structure, internal reliabilities of the factors, convergent and divergent validity, predictive validity, incremental validity, and test-retest reliability. Within the realm of personality and social psychology, there is Glick and Fiske's (1996) work on ambivalent (hostile and benevolent) sexism, which like the first two exemplars, contains six studies that spotlight a plethora of psychometric consideration in scale development.

The first three exemplars contained six studies. And indeed, thorough scale development does require a great deal of work, as Gurung (2015) emphasized. However, I know of two particularly good sources of psychometric work that contain fewer studies. First, there is Midgley and colleagues' (1998) work on scales to measure different goal orientations (i.e., task-goal, ability-approach, and ability-avoidant orientations). This source is particularly good because it provides a history of the authors' work done on goal orientation measures. This history is organized around major psychometric properties of their measures. It is a particularly good exemplar to use with students in an undergraduate tests and measurements class. Finally, Oleson, Poehlmann, Yost, Lynch, and Arkin (2000) presented two studies that assessed two components of subjective overachievement: self-doubt and concern with performance. In their research, Oleson and colleagues examined the factor structure, internal reliabilities, test-retest reliabilities, convergent validity, and divergent validity of scales to measure self-doubt and concern with performance.

In sum, the choice of scales in SoTL presents us with an opportunity to learn from our colleagues in personality, I/O, and learning sciences, all of which have a long history of scale development. In addition, these areas also have scales that may be of use in SoTL. By integrating what these areas have to offer in terms of psychometric work and resulting scales, we can help bridge intellectual divides that hinder research progress in all of these areas.

# References

References marked with an asterisk indicate a scale.

Buskist, W. (2013). Preparing the new psychology professoriate to teach: Past, present, and future. *Teaching of Psychology, 40,* 333-339. http://dx.doi.org/10.1177/0098628313501047

*Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27-39). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

*Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197-253. http://dx.doi.org/10.1037/0033-2909.119.2.197

Christopher, A. N., Furnham, A., Batey, M., Martin, G. N., Koenig, C. S., & Doty, K. (2010). Protestant ethic endorsement, personality, and general intelligence. *Learning and Individual Differences, 20,* 46-50. http://dx.doi.org/10.1016/j.lindif.2009.10.003

Daniel, D. B., & Chew, S. (2013). The tribalism of teaching and learning. *Teaching of Psychology, 40,* 363-367. http://dx.doi.org/10.1177/0098628313501034

*Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92,* 1087-1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087

*Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528. http://dx.doi.org/10.1002/tl.20071

*Glick, P., & Fiske, S. T. (1996). The Ambivalent Sexism Inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology, 70,* 491-512. http://dx.doi.org/10.1037/0022-3514.70.3.491

Gurung, R. A. R. (2015). Best practices in scale use in SoTL. In R. S. Jhangiani, J. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

*Keeley, J., Smith, D., & Buskist, W. (2006). The Teaching Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33,* 84-91. http://dx.doi.org/10.1207/s15328023top3302_1

McCabe, J. A. (2015). Learning the brain in Introductory Psychology: Examining the generation effect for mnemonics and examples. *Teaching of Psychology, 42,* 203-210. http://dx.doi.org/10.1177/0098628315587617

Meriac, J. P., Slifka, J. S., & LaBat, L. R. (2015). Work ethic and grit: An examination of empirical redundancy. *Personality and Individual Differences, 86,* 401-405. http://dx.doi.org/10.1016/j.paid.2015.07.009

*Midgley, C., Kaplan, A., Middleton, M., Maehe, M. L., Urdan, T., Anderman, L. H., Anderman, E., & Roesner, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology, 23,* 113-131.

*Miller, M. J., Woehr, D. J., & Hudspeth, N. (2002). The meaning and measurement of work ethic: Construction and initial validation of a multidimensional inventory. *Journal of Vocational Behavior, 60,* 451-489. http://dx.doi.org/10.1006/jvbe.2001.1838

*Oleson, K. C., Poehlmann, K. M., Yost, J. H., Lynch, M. E., & Arkin, R. M. (2000). Subjective overachievement: Individual differences in self-doubt and concern with performance. *Journal of Personality, 68,* 491-524. http://dx.doi.org/10.1111/1467-6494.00104

Richmond, A. S. (2015). The missing link(s) of SoTL scales: One researcher’s perspective. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Sawyer, R. K. (Ed.). (2006). *The Cambridge handbook of the learning sciences, preface*. New York, NY: Cambridge University Press.

Wilson-Doenges, G. (2015). The state of scale validation in SoTL research in psychology. In R. S. Jhangiani, J. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Footnotes

[1] Once while working on a personnel selection project, a colleague asked me if there was really any value in personnel evaluation procedures. Thus, what I am calling a “divide” might be better described as “contempt” in some circumstances.

[2] Although a far from perfect indicator of journal quality, one such wide-accepted benchmark is a journal’s impact factor. You can typically find this information for a journal on its publisher’s website and from the Social Sciences Citation Index.

[3] I think that the TBC, although rooted more in SoTL research, could easily be used in learning sciences. Though most of the cited references I found on Keeley et al.’s (2006) psychometric paper were indeed in SoTL-type outlets, this measure has caught on in outlets that might be associated more with the learning sciences, such as *Assessment & Education in Higher Education* and *Review of Educational Research*.

[4] I keep issues of *Personality and Individual Differences* and *Learning and Individual Differences* on my nightstand and read them before going to sleep each night. So, I may be exposed to a wider range of scales than most people and hence my impression of “scale overload” in psychological research.

# Chapter 4: A Primer on Scale Development

Heather D. Hussey and Tara J. Lehan

Northcentral University

## Introduction

One of the primary features of contemporary academic, professional, public, and personal life is a reliance on information and arguments involving numbers. This chapter includes an introduction to the steps involved in developing and validating instruments to collect numerical data for use in the study of teaching and learning (SoTL) research. Independent scholars must be able to evaluate quantitative evidence thoughtfully and critically as well as employ quantitative skills to make contributions in the workplace and to society. To become competent creators and consumers of quantitative research, readers must understand how numerical information is generated, summarized, evaluated, and represented. With these goals in mind, classical and modern test theories, reliability, validity, factor analysis, and other topics are covered in some detail in this chapter. However, it is crucial to note that this chapter is only a starting point for those who are interested in better understanding the use and creation of instruments in SoTL research. Further, many of the discussions lack the depth necessary to provide readers with the knowledge needed to begin to develop their own instruments. Only the most relevant topics are covered at a superficial level due to space limitations.

Although the process of scale creation frequently is iterative, this chapter includes a description of the general steps involved in developing an instrument, including searching the scholarly literature to locate and evaluate preexisting instruments and then determining whether they are adequate or need modification (see Figure 1). If researchers must modify a preexisting instrument or create an entirely new one, they should return to the scholarly literature to ensure that they acquire all of the relevant knowledge on the construct(s) of interest. Next, in conjunction with measurement theories and/or models, researchers use this information to develop a pool of potential items. Following the creation of this pool, researchers often have a panel of experts review the items for clarity and conduct an initial screening of the degree to which they align with construct(s) of interest. Some researchers also conduct a pilot or field test with their instrument with a small sample. They then implement all suggested changes and administer the instrument to a larger sample. Following data collection, researchers analyze the responses. This process often involves factor analysis and assessments of reliability and validity. Based on the findings, they make necessary modifications and repeat the process and/or conduct studies to replicate and/or generalize their findings.

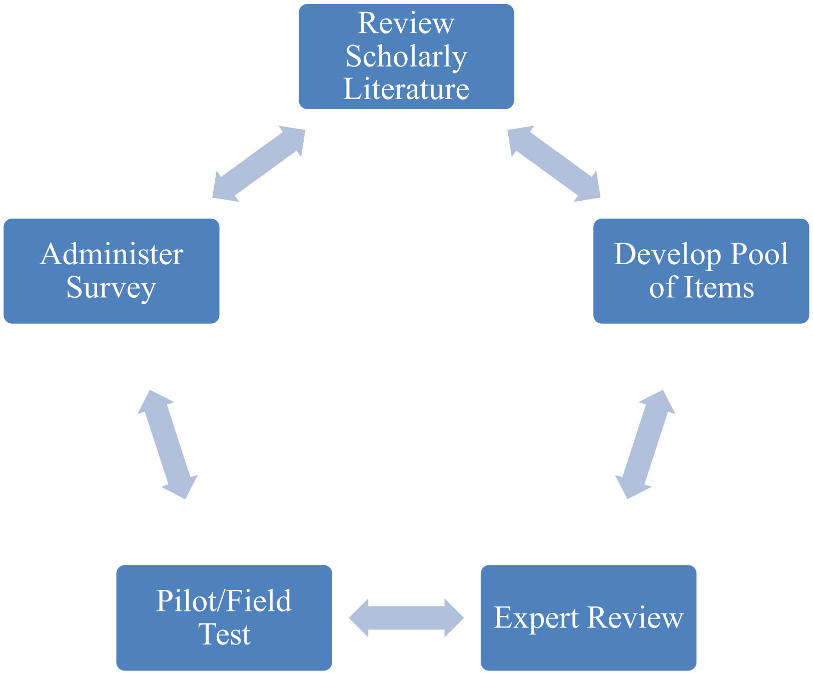

*Figure 1*. Scale creation process depicted as an iterative process. Note: this process should be interpreted loosely, as one might return to any step at any time, depending on the outcome. For example, comments on an expert review might result in returning back to the scholarly literature and so on.)

## Use of Existing Scales

Researchers commonly employ questionnaires to collect data, especially in applied research (Boynton & Greenhalgh, 2004). Whenever possible, researchers should locate and employ previously developed scales with evidence of validity and reliability. One benefit of using existing scales is that there are generally published studies with data that researchers can compare to the data obtained in subsequent studies. When researchers conduct multiple studies using the same scale, they can build a body of literature and theory. Further, they need to devote fewer resources, especially time and energy, when using existing scales. Many of the chapters in this e-book are devoted to sharing such scales that researchers can use in their SoTL research. For example, the Teacher Behaviors Checklist (TBC; Buskist, Sikorski, Buckley, & Saville, 2002) is a popular measure used in SoTL research (see Kirk, Busler, Keeley, & Buskist, 2015). It is discussed further in Chapter 13 of this e-book. However, researchers might find that preexisting scales do not measure certain elements of a construct. For example, Wilson, Ryan, and Pugh (2010) noted that, although there were existing measures of rapport, none specifically addressed professor-student rapport. As a result, they sought to develop such a measure. Also discussed further in Chapter 3 of this e-book, researchers sometimes use scales before there is sufficient research to support their validity and reliability (Christopher, 2015). As such, they might draw seemingly significant conclusions from the application of new scales, only to have the findings of subsequent studies contradict their findings (Cook, Hepworth, Wall, & Warr, 1981). In such cases, it might be appropriate for the researchers to modify existing scales or develop new ones. In either case, it is important for them to return to the scholarly literature to collect as much relevant information as possible on the construct of interest to inform scale development.

## Measurement Theories

There are two major underlying theories upon which quantitative measurement is founded: classical test theory (CTT) and item response theory (IRT). Representing two distinct frameworks, they provide the foundation from which researchers often build instruments. Using these theories, researchers can obtain information about the quality of the items that they develop. Given that there are entire chapters and books devoted to these topics, the purpose of this section is to provide a general overview (see Hambleton & Jones, 1993, and de Champlain, 2010, for an overview and comparison of these theories).

### CTT

CTT describes a set of psychometric procedures used to test scale reliability, item difficulty, and item discrimination (i.e., the extent to which an item helps the research to differentiate between respondents with higher and lower abilities). Many widely available statistical packages produce these statistics. There are several classical test theories, all of which are founded upon the assumption that a raw score is comprised of a true score and a random error (Kline, 2005). In the context of teaching and learning, a true score might reflect a student's true ability. In this case, the unsystematic influence of any factors on the measurement of that ability would be referred to as random error. In addition, the overriding concern of CTT is to manage the random error. The less random error there is, the more the raw score reflects the true score. Due to its relatively weak assumptions, CTT can be used in a wide variety of testing situations (Hambleton & Jones, 1993). However, the main criticism is that statistics that are generated in terms of the observed score, item difficulty, and item discrimination are dependent upon the sample (Fan, 1998).

### IRT

Also known as modern test theory, IRT allows researchers to overcome many of the limitations of CTT. It can be used to model respondent ability using item-level performance, rather than aggregate test-level performance. Many times, the variable that researchers want to study, such as intelligence, cannot be measured directly; therefore, they create a number of items that they believe capture it (Noar, 2009). Such variables are referred to as latent variables. IRT involves establishing a model that specifies the probability of observing each response to an item as a function of the latent trait being measured, which is often knowledge, a skill, or an ability (DeVellis, 2012). The item response function illustrates the relationship between a latent variable and the probability of endorsing an item. It can then be converted into an item characteristic curve, which shows respondent ability as a function of the probability of endorsing the item.

## Item Development

### Wording

High-quality test items are critical to the development of a meaningful instrument. Precise language must be used for a measure to adequately capture the specific construct of interest, yet contain no extraneous content. Researchers should avoid the use of language that might be confusing, such as items that include double negatives, as well as doubled-barreled questions that tap into more than one construct but allow for only one response. In addition, items should be written at a level that is appropriate for the target responses. Various resources exist that provide researchers with information about the grade level at which an item is written, including the readability calculator at https://readability-score.com/. Consider the following questions:

> "To what extent does the teacher utilize examples to explicate convoluted concepts?"
> "How much does the teacher use examples to describe difficult concepts?"

Whereas both questions are tapping into the sample construct, the first one requires that respondents have a higher reading level than the second one. If the intended respondents are sophomores in high school, its reduced readability could potentially impact the responses. For additional considerations to consider when wording questions (e.g., types of scales to use) see Schaeffer & Presser (2003).

### Strategies

An inductive or a deductive approach can be used to generate items. With an inductive approach, there is no theoretical or conceptual framework, whereas there is with a deductive approach. It seems that it is necessary for researchers to establish a clear connection between items and their theoretical domain (Hinkin, 1995). In developing items, researchers often employ a variety of strategies, including pulling from existing measures, interviewing select populations, and using focus groups. See Krause (2002) for a description of a nine-step strategy for developing items. For example, Wilson and colleagues (2010) used students in their upper-level psychology course to generate potential items for their professor-student rapport measure.

## Expert Review and Pilot Testing

Once the instrument is created, researchers often go through review processes before fully administering the instrument. Such processes include having a panel of experts review the items for feedback regarding whether the items are clear and reflect the construct of interest (DeVellis, 2012; Worthington & Whittaker, 2006). Once the scale is created, it should also be pilot-tested on a relatively small sample of the population of interest to determine where there might be wording confusion or redundancy, whether there are completion time issues, and whether response items cover all desired options (van Teijlingen & Hundley, 2001). For example, Van Tassel-Baska, Quek, and Feng (2007) piloted their teacher observation scale and found their 40-item measure could be reduced to 25 items to decrease redundancy, which can impact reliability as discussed below. Researchers might also use cognitive interview techniques with participants after they complete a newly developed instrument to garner feedback about the instrument as well as have participants think aloud while answering each survey question (Ryan, Gannon-Slater, & Culbertson, 2012). These findings should inform modifications to the instrument. Such processes are especially important when special populations are involved (e.g., children, non-native speakers, etc.) (Presser et al., 2004).

## Instrument Administration

### Delivery Options

Once the instrument is complete, it is ready to be fully administered.  However, it is also important to consider the many practical and logistical issues related to administering tests (Nosek, Banaji, & Greenwald, 2002; Wright, 2005). For example, a researcher might wonder whether it is more appropriate to have student participants complete instruments in-person or online (Miller et al., 2002).  Many scholars have opted to collect data online due to benefits such as lower costs and increased access to participants (Miller et al., 2002; Wright, 2005). In addition, researchers have found comparable reliability and validity of instruments administered online versus in-person with paper and pencil (Meyerson & Tryon, 2003; Miller et al., 2002).  Furthermore, researchers have found online samples to approximate the characteristics of traditional samples (Gosling, Vazire, Srivastava, & John, 2004).

### Survey Design

Researchers also need to be cognizant in how the design of the survey can potentially impact participant responses (e.g., completion time and valid responses) (Christian & Dillman, 2004; Couper, Traugott, & Lamias, 2001). For example, Mahon-Haft and Dillman (2010) showed that the design of online surveys, including aesthetics and type of open response format, can negatively impact the quality of data collected (e.g., omissions and shorter responses). This issue is crucially important to reliability and validity, both of which are discussed further below.

### Sample Size

Additional concerns related to instrument administration is adequate sample size and the characteristics of the sample. It is worth noting that although there are varying beliefs in what constitutes an adequate sample size when developing an instrument (e.g., at least 150, 300, varying variable ratios such as 20:1), it is safe to say the larger the sample size, the better (Schmitt, 2011; Williams, Onsman, & Brown, 2010; Yong & Pearce, 2013). Sample size and reliability become critically important when considering their impact on power to detect an effect and the ability to reproduce observed effects (Fraley & Vazire, 2014; Henson, 2001; Simons, 2014). In regard to scale development, sample size is also important in that smaller sample sizes can limit the type of analyses that can be performed (Noar, 2009). It is also important for researchers to consider the demographic characteristics of the participants who are completing the instrument, as test bias and certain ethical issues might be of concern (van de Vijver & Tanzer, 1997). In addition, researchers should attempt to recruit participants who are heterogeneous on the construct (Clark & Watson, 1995).

## Analyze Responses

After researchers administer their instrument to a large group of participants, they begin analyzing their findings. The analyses typically involved in instrument development include factor analysis, reliability estimates, and validity tests. However, it is important to note that these are often not the only analyses involved. Given that there is not nearly enough room in this entire e-book to cover each topic fully, the following discussion touches upon some of the main concepts related to analyzing items for instrument development in order to familiarize readers and guide them toward additional readings.

### Factor Analysis

As discussed further below, researchers cannot directly measure latent variables. Instead, they rely on observable items/variables that are believed to reflect the latent variable and then work to determine how well they have achieved this goal. Often times, measures of reliability are not adequate in informing the dimensionality of an instrument (John & Benet-Martinez, 2000). In other words, an instrument with a high alpha level does not necessarily mean that all the items reflect a single construct or factor (DeVellis, 2012). For example, the Teaching Behaviors Checklist (TBC) (Keeley, Smith, & Buskist, 2006), discussed in more depth later in this e-book, can be used as a single factor instrument for identifying effective teaching behaviors or an instrument with two factors related to evaluating teaching behaviors.

One of the most effective ways to determine how items relate and reflect a latent variable(s) is using factor analysis (Byrne, 2001). Factor analysis is a set of procedures that allow the determination of how many factors there are and how well the items reflect those factors. There are two main types of factor analysis: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), with the former being where most start when first developing or revising an instrument (Yong & Pearce, 2013). Noar (2009) describes the steps of scale development including factor analysis using a structural equation modeling (SEM) technique, which is becoming more popular (Worthington & Whittaker, 2006). Given the space constraints of this chapter and the vast amount of information and considerations related to factor analysis (Schmitt, 2011), readers are encouraged to examine primers (e.g., Williams et al., 2010; Yong & Pearce, 2013), texts (e.g., Thompson, 2004), and researched best practices (Worthington & Whittaker, 2006) for more in-depth information regarding factor analysis in scale development.

As discussed earlier, researchers generally start with a pool of items that they administer to a large group of participants. Through the procedures involved in factor analysis, researchers can determine if their pool of items reflect a single factor or multiple factors (Yong & Pearce, 2013). Going back to the TBC (Keeley et al., 2006) example, the authors used this instrument as an evaluative tool and through factor analysis found all 28 items could be used as a single factor to reflect overall teacher behaviors or separated into two subscales (or two factors) representing caring/supportive teaching behaviors and professional competency/communication skills.

Researchers also use factor analysis for reduction purposes, including limiting the number of items in an instrument as well as the number of factors included in an instrument (Yong & Pearce, 2013). For example, another scale discussed further in Chapter 11 of this e-book is the Satisfaction with Life Scale (SWLS) (Diener, Emmons, Larsen, & Griffin, 1985; see Layous, Nelson, & Legg, 2015). These authors reviewed the literature, found preexisting scales of life satisfaction to be inadequate, and therefore, began the process of developing a new instrument. In the first phase, they developed 48 items related to satisfaction with one's life, which was decreased to 5 items by using factor analysis. Procedures such as these help researchers to find ways to more effectively measure a construct, while also reducing participant fatigue with answering numerous items and increasing the likelihood of replicating findings (Henson & Roberts, 2006).

It is also important to note that statistical programs will only show where and how items load; it is up to the researcher to interpret the findings using the literature and/or a theoretical framework (Henson & Roberts, 2006; Yong & Pearce, 2013). For example, deleting items as suggested by SPSS does not always increase reliability (Cho & Kim, 2015). Depending on the findings of the factor analysis, researchers might go back to the literature and develop additional items to pilot or determine the instrument's structure and assess its reliability and validity (Warner, 2013).

## Reliability

Regardless of whether an instrument being used for SoTL is preexisting, an adapted measure, or newly developed, it is important to assess its reliability. Generally speaking, reliability refers to how consistently an instrument measures a certain construct or variable of interest (John & Benet-Martinez, 2000). Scores on the instrument should not change unless there is a change in the variable. It is not possible to calculate reliability exactly, as there is inevitable measurement error. Therefore, researchers have to estimate reliability.  In doing so, all attempts to remove or control sources of error should be made so as to obtain a truer score of the latent variable (DeVellis, 2012).

The question then becomes, how do we know if an instrument is reliable? The methods for assessing reliability can vary depending on type of data and information sought (John & Benet Martinez, 2000); however, they are all based on the idea that reliability is the proportion of variance of the observed score that can be attributed to the true score of the latent variable being examined (DeVillis, 2012). Many researchers will report a single coefficient alpha regarding reliability (Hogan, Benjamin, & Brezinski, 2000), but that is a limited picture of reliability and not always the best indicator (Cho & Kim, 2015; Cronbach & Shavelson, 2004; John & Benet-Martinez, 2000). For example, Cortina (1993) demonstrated how alpha can increase with the number of items in an instrument despite low average item intercorrelations. Therefore, it is important to examine whether additional reports of item relationships and/or reliability are given (Schmitt, 1996). There is also somewhat of a challenge in determining whether reports of reliability are adequate. For example, most researchers use a cutoff for alpha at .70 or higher as adequate (Hogan et al., 2000; Schmitt, 1996); however, others note .80 or higher to be the recommended levels for alpha (Henson, 2001; Lance, Butts, Michels, 2006). However, higher is not always better as an overly high alpha might be the result of redundancy in the items (Ryan & Wilson, 2014; Streiner, 2003; Tavakol & Dennick, 2011). Ryan and Wilson's (2014) brief version of the Professor-Student Rapport Scale with a Cronbach's alpha of .83 is a good example of acceptable internal consistency that is not so high as to suggest issues with redundancy.

Some of the most common methods of assessing reliability include inter-observer agreement, test-retest reliability, and measures of internal consistency. What follows is a brief discussion of these common ways of assessing reliability as well as additional resources faculty can examine to learn more about these topics. It is important to keep in mind that this discussion is a brief introduction into concepts related to reliability; entire books are devoted to this topic (e.g., see Fink & Litwin, 1995; Litwin, 2003).

### Inter-Observer

Assessing inter-observer agreement comes into play when researchers want to look at the reliability of responses given by two or more raters or judges (Warner, 2013). Depending on the goal of the researcher, the level of agreement between raters could be allowed to vary or not. For example, in a psychology statistics class, two teachers should be in perfect agreement regarding whether a student correctly computed a specific statistical analysis. In contrast, a group of teaching assistants might be asked to come into a classroom and judge students' poster projects for the best designed poster to examine if a certain teaching intervention was effective. In this case, perfect score agreement might not be needed or beneficial. Instead, we might look at how the judges ranked the posters from best to worst. Researchers can also examine percentage of agreement between raters by counting the number of times the raters were in agreement and dividing by the total number of judgments made. However, this method of assessment does not take into account chance levels of agreement between raters and tends to work best when the rated variable is objective rather than subjective. To take into account chance agreements, one would need to calculate Cohen's kappa *(*🐝coefficient (Viera & Garrett, 2005), with coefficients below .40 being poor, between .40 and .59 fair, .60 and .74 good, and.75 and 1.00 excellent (Cicchetti, 1994). Sources of measurement error to be mindful of include judges' background and training regarding the variable of interest and the tool being used as well as prevalence of the finding (DeVillis, 2012; Guggenmoos-Holzmann, 1996; Viera & Garrett, 2005).

### Test-Retest

Reliable instruments should provide constant scores if used to assess a variable at one time point and then again at another time point (so long as theoretically that variable should remain constant across time); however, the error of measurement each time will vary. A correlation could then be computed to assess consistency in scores for the two assessment points.

Although there is some variability in the literature, there appears to be consensus that a retest reliability coefficient of .80 or higher is recommended (Polit, 2014). For example, teachers might be interested in their psychology students' level of satisfaction with their academic advisors and measure said levels at two points in time. Although the measurement of error will differ at each assessment point, the scoring of satisfaction should stay constant if the instrument truly reflects levels of satisfaction. However, it is important to note that changes in scoring from point one to point two can vary for a number of reasons and not necessarily reflect (un)reliability. For example, actual changes in the variable of interest (e.g., an overall increase or decrease in students' level of satisfaction with their advisors) would affect test-retest reliability scores without reflecting unreliability. In addition, test-retest reliability can be affected by participant factors such as purposely trying to answer consistently (or differently) each time they complete the instrument (DeVillis, 2012; Polit, 2014).

### Internal Consistency

As mentioned previously, we are unable to directly measure latent variables; instead, we rely on proxies. If we are to assume the items of an instrument represent the latent variable of interest, then the strength of the relationship between those items should reflect the strength of the measurement of the latent variable (DeVillis, 2012). The most common way to determine whether an instrument is internally consistent is to compute Cronbach's alpha for items that have multiple response options or the Kuder-Richardson formula 20 (KR-20) for items that have dichotomous responses. In addition to computing a coefficient alpha, researchers might include a measure of split-half reliability. Exactly how the items are split (e.g., in the middle, odds vs. evens, etc.) depends on the types of items in the measurement as well as the goals of the researcher. For example, a measure regarding gay, lesbian, bisexual, and transgender students' perceptions of diversity inclusion in curriculum that is split in the middle might have an unbalanced number of survey items focused on lesbian students. In this case, it might make more sense to ensure subgroup perceptions were equally distributed into two halves. Overall, when computing an alpha there are a number of factors researchers need to consider, including the number of items in the instrument and whether there are reverse-scored items as these could affect reliability scores (Warner, 2013).

## Validity

Whether preexisting or newly developed, it is also important to examine the validity of instruments. Generally, validity indicates how well an instrument measures what it claims to measure; however, this metric of an instrument is somewhat more difficult to determine (Warner, 2013). Although some have purported multiple types of validity, the general consensus appears to be around three types: content validity, construct validity, and criterion-related validity (DeVellis, 2012). Further discussion of validity of instruments in the SoTL research is covered in Chapter 5 of this e-book.

### Content Validity

Although there are some differences in how content validity is defined in the literature, “there is general agreement in these definitions that content validity concerns the degree to which a sample of items, taken together, constitute an adequate operational definition of a construct” (Polit & Beck, 2006, p. 490). However, content validity is often left to the judgment of the researcher and/or a panel of experts to ensure only relevant items related to the construct are included (DeVellis, 2012). One issue with stopping at this step is that it is a subjective judgment of whether an instrument has adequate content validity. Some researchers suggest the computation of a content validity index (CVI) to better judge the psychometrics of an instrument (Polit & Beck, 2006; Wynd, Schmidt, & Schaefer, 2003). Generically speaking, this involves having a panel of raters review each item in the measure, rate how relevant each item is to the construct being measured, and then the researcher(s) assess agreement between raters.

### Construct Validity

Construct validity is based on a theoretical framework for how variables should relate. Specifically, “it is the extent to which a measure ‘behaves’ the way that the construct it purports to measure should behave with regard to established measures of other constructs” (DeVellis, 2012, p. 64). Some deem construct validity as one of the most important concepts in psychology (John & Benet-Martinez, 2000; Westen & Rosenthal, 2003). However, there is no agreed upon or simple way to assess construct validity (Bagozzi, Yi, & Phillips, 1991; Westen & Rosenthal, 2003). Instead, many researchers use correlations to show relationships between variables as suggested by the literature, often referred to as convergent and discriminant validity (Messick, 1995). For example, Ryan and Wilson (2014) examined the validity of a brief version of the Professor-Student Rapport Scale by comparing the brief measure to scales that were similar (e.g., Immediacy Scale, to which it should positively relate) and dissimilar (i.e., Verbal Aggressiveness Scale, to which it should negatively relate). There are also computations that can be performed to demonstrate levels of construct validity including confirmatory factor analysis, effect size correlations, and structural equation modeling (Bagozzi et al., 1991; Westen & Rosenthal, 2003).

### Criterion-Related Validity

Criterion-related validity is mainly concerned with practicality versus theoretical underpinnings like construct validity; “it is not concerned with understanding a process but merely with predicting it” (DeVellis, 2012, p. 61). However, many confuse construct validity and criterion-related validity because the same example listed above can be used for either type of validity; the difference is in the intent of the researcher (e.g., to explain or explore variable relationships versus simply identifying and predicting) (DeVellis, 2012). In the case of the latter, there are multiple types of correlations that can be computed to identify specific relationships among variables and in turn, different types of validity. For example, researchers can examine whether the scores on their instrument predict future behaviors (i.e., predictive validity), correlate with scores on another instrument shown to measure the same construct (i.e., convergent validity), and/or do not correlate with instruments that are unrelated to the construct of interest (i.e., discriminant validity; Warner, 2013). In many cases, researchers will want both construct and criterion-related validity due to basing their survey development in theory and striving toward the ability to predict an outcome.

## Conclusions

In sum, scale development is an iterative, rather than a linear process. It is likely that researchers might have to return to earlier steps in the scale-development process at times. In addition, the process should also be deliberate and be guided by a theoretical foundation where appropriate. Only when such a process is employed can a psychometrically sound measure be developed that yields meaningful data. As the saying goes, “a little knowledge is a dangerous thing,” yet this is exactly what was provided in this chapter. The purpose was to familiarize readers with some of the main concepts and processes that researchers must consider when developing or modifying a scale. The hope is that readers might use the easily digestible sources cited in this chapter as a springboard for further skill building related to scale development.

# References

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*(3), 421-458.

Boynton, P. M., & Greenhalgh, T. (2004). Selecting, designing, and developing your questionnaire. *BMJ, 328,* 1312-1315. doi:10.1136/bmj.328.7451.1312

Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27–39). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education*, *44*(1), 109-117. doi:10.1111/j.1365-2923.2009.03425.x

Cho, E., & Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods, 18*(2), 207-230. doi:10.1177/1094428114555994

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly, 68*(1), 57-80. doi:10.1093 /poq/nfh004

Christopher, A. N. (2015). Selecting the right scale: An editor's perspective. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. doi:10.1037/1040-3590.6.4.284

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

Cook, J. D., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The experience of work*. San Diego, CA: Academic Press

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Couper, M. P., Traugott, M. W., & Lamias, M. J. (2001). Web survey design and administration. *Public Opinion Quarterly, 65*, 230-253. doi:10.1086/322199

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391-418. doi:10.1177/0013164404266386

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49*(1), 71-75.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Fink, A., & Litwin, M. S. (1995). *How to measure survey reliability and validity* (Vol. 7). Thousand Oaks, CA: Sage.

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*(10), e109019. doi:10.1371/journal.pone.0109019

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93-104. doi:10.1037/0003-066X.59.2.93

Guggenmoos-Holzmann, I. (1996). The meaning of Kappa: Probabilistic concepts of reliability and validity revisited. *Journal of Clinical Epidemiology, 49*(7), 775-782. doi:10.1016/0895-4356(96)00011-X

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47. doi:10.1111/j.1745-3992.1993.tb00543.x

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement & Evaluation in Counseling & Development*, *34*(3), 177-189.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66*(3), 393-416. doi:10.1177/0013164405282485

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*(5), 967-988. doi:10.1016/0149-2063(95)90050-0

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*(4), 523-531. doi:10.1177/00131640021970691

John, O. P., & Benet-Martinez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H.T. Reis & C.M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 339-369). New York: Cambridge University Press.

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*(2), 84-91. doi:10.1207/s15328023top3302_1

Kirk, C., Busler, J., Keeley, J., & Buskist, W. (2015). Effective tools for assessing characteristics of excellent teaching: The Teacher Behaviors Checklist as exemplar. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Kline, T. J. B. (2005). Classical test theory. In T. J. B. Kline (Ed.), *Psychological testing: A practical approach to design and evaluation* (pp. 91-105). Thousand Oaks, CA. Sage.

Krause, N. (2002). A comprehensive strategy for developing closed-ended survey items for use in studies of older adults. *Journal of Gerontology: Social Science, 57B*(5), S263-S274.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. What did they really say? *Organizational Research Methods, 9*(2), 202-220. doi:10.1177/1094428105284919

Layous, D., Nelson, S. K., & Legg, A. M. (2015). Measuring positive and negative aspects of well-being in the scholarship of teaching and learning. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Litwin, M. S. (2003). *How to assess and interpret survey psychometrics* (Vol. 8). Thousand Oaks, CA: Sage.

Mahon-Haft, T. A., & Dillman, D. A. (2010). Does visual appeal matter? Effect of web survey aesthetics on survey quality. *Survey Research Methods, 4*(1), 43-59. doi:10.18148/srm/2010.v4i1.2264

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Meyerson, P., & Tryon, W. W. (2003). Validating internet research: A test of the psychometric equivalence of internet and in-person samples. *Behavior Research Methods, Instruments, & Computers, 35*(4), 614-620. doi:10.3758/BF03195541

Miller, E. T., Neal, D. J., Roberts, L. J., Baer, J. S., Cressler, S. O., Metrik, J., & Marlatt, G. A. (2002). Test-retest reliability of alcohol measures: Is there a difference between internet-based assessment and traditional methods? *Psychology of Addictive Behaviors, 16*(1), 56-63. doi:10.1037/0893-164X.16.1.56

Noar, S M. (2009). The role of structural equation modeling in scale development. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(4), 622-647, doi:10.1207/S15328007SEM1004_8

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues, 58*(1), 161-176. doi:10.1111/1540-4560.00254

Polit, D. F. (2014). Getting serious about test-retest reliability: a critique of retest research and some recommendations. *Quality of Life Research, 23*, 1713-1720. doi:10.1007/s11136-014-0632-9

Polit, D. F., & Beck, C. T. (2006). The Content Validity Index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489-497.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109-130. doi:10.1093/poq/nfh008

Ryan, K., Gannon-Slater, N., & Culbertson, M. J. Improving survey methods with cognitive interviews in small- and medium-scale evaluations. *American Journal of Evolution, 33*(3), 414-430. doi:10.1177/1098214012441499

Ryan, R. & Wilson, J. H. (2014). Professor-student rapport scale: Psychometric properties of the brief version. *Journal of the Scholarship of Teaching and Learning, 14*(3), 64-74. doi:10.14434/josotl.v14i3.5162

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, 29(1), 65-88. doi:10.1146/annurev.soc.29.110702.110112

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353. doi:10.1037/1040-3590.8.4.350

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304-321. doi:10.1177/0734282911406653

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80. doi:10.1177/1745691613514755

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*(1), 99-103. doi:10.1207/S15327752JPA8001_18

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. doi:10.5116/ijme.4dfb.8dfd

Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263-280. doi:10.9707/2307-0919.1111

Van Tassel-Baska, J., Quek, C., & Feng, A. X. (2007). The development and use of a structured teacher observation scale to assess differentiated best practice. *Roeper Review: A Journal on Gifted Education, 29*(2), 84-92. doi:10.1080/02783190709554391.

van Teijlingen, E., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update*, *35*, 1-4. doi:10.1046/j.1365-2648.2001.01757.x

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistic. *Family Medicine, 37*(5), 360-363.

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Los Angeles, CA Sage.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology, 84*(3), 608-618. doi:10.1037/0022-3514.84.3.608

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine, 8*(3), 1-13.

Wilson, J. H. Ryan, R. G., & Pugh, J. L. (2010). Professor-student rapport scale predicts student outcomes. *Teaching of Psychology, 37*, 246-251. doi:10.1080/00986283.2010.510976

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806-838. doi:10.1177/0011000006288127

Wright, K. B. (2005). Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication 10*(3). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00259.x/full

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research, 25*(5), 508-518. doi:10.1177/0193945903252998

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94.

# Chapter 5: The State of Scale Validation in SoTL Research in Psychology

Georjeanna Wilson-Doenges

University of Wisconsin-Green Bay

In psychology, efforts to operationalize and measure constructs remain a challenging and complex task. Much of this work relies on self-report and other-report ratings that require developing items and scales to tap into complicated psychological constructs. As a field, psychology has high standards for assessing the reliability and validity of such scales. The scholarship of teaching and learning (SoTL) should be held to these same high standards of scale validation in the assessment of learning. As Noar (2003) states, "Across a variety of disciplines and areas of inquiry, reliable and valid measures are a cornerstone of quality research" (p. 622). The purpose of this chapter is to provide a brief overview of the state of scale validation in recent SoTL in psychology and provide some exemplars of ways in which SoTL researchers have employed the best practices for scale validation. It should not go unnoticed that classroom research is often messy, adding a layer of complexity to SoTL research. Due to the lack of total control in this type of research environment, trouble-shooting and work-arounds are part of a SoTL researcher's toolbox. For example, SoTL researchers are often constrained by the sample size and make-up of their classes (where much SoTL is done), as well as the time constraint of the length of a typical semester. Taking the time to assess reliability and validity under these constraints can be challenging and can even require extension of the research over several semesters. This can then lead to issues with reliability and validity over different class samples and instructors, which could hinder the quality of the research. Some examples of ways to maximize scale validation in SoTL research in psychology will also be discussed.

## Method

In order to assess the state of scale validation in recent SoTL in psychology, a content analysis of the last ten issues (volume 40, issue 1 to volume 42, issue 2) of a prominent SoTL journal in psychology, *Teaching of Psychology*, was performed. Only empirical articles were examined. Of the 141 articles assessed, 47 articles used at least one scale in the research. A scale was defined as a variable for which more than one item was used to measure the same construct. Where scales were present, the assessment and reporting of reliability, validity, and the psychometric development of the scales were recorded. Exemplars were also identified during this content analysis.

## Results

### Assessing and Reporting Reliability

#### *Alpha*

Scale validation has several important components, one of them being the internal consistency of the scales used. Internal consistency is most often measured by correlational statistics like Cronbach's alpha (Cronbach, 1951). SoTL in psychology has been consistent in establishing and reporting internal reliability of scales. In the content analysis of recent issues of *Teaching of*

*Psychology*, 36 of the 47 articles (76.6%) where scales were used to assess SoTL outcomes, internal reliability was reported. Thirty-one studies (66.0%) reported Cronbach's alpha values that were all acceptable (> .70; Cronbach, 1951; DeVellis, 1991), whereas five (10.6%) reported at least one alpha value that was below the acceptable .70 value. One example of reporting internal consistency is Maybury's (2013) study of the influence of a positive psychology course on several measures of student well-being using previously established scales and reporting the Cronbach's alpha of each scale and subscale. The alphas in this study ranged from .72 to .89, all above the .70 minimum for acceptable reliability (see Chapter 11 in this e-book by Layous, Nelson, & Legg for more detail of this study). Another example of reporting internal consistency is a study of a team approach to undergraduate research that used items that the researchers developed, rather than a previously established scale (Woodzicka, Ford, Caudill, & Ohanmamooreni, 2015). In this study, the researchers developed a 19-item, 5-point Likert-scale online survey with four subscales. Each subscale had a Cronbach's alpha above acceptable, even with a relatively small sample size (Woodzicka et al., 2015). Reporting the reliability of newly developed items used for the first time is particularly important because of the lack of established history.

Although, the majority of internal consistency values of the scales published in SoTL in psychology examined were above the acceptable minimum (83.3% of the studies reporting acceptable Cronbach's alphas), there were several instances where at least one scale assessed was not. One particular study addressed this issue directly, by assessing reliability and noting that it was not acceptable for one of the subscales and then giving an appropriate rationale for still using the subscale in the research (Boysen, 2015). Specifically, the rationale was that there was a strong theoretical relation of the items in the subscale and that the items showed equal patterns of results; therefore, the researcher kept the subscale in the model even though it was found to be unreliable (Boysen, 2015). Another example of the use of a scale with poor reliability is Boysen, Richmond, and Gurung's (2015) study of model teaching criteria using the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003). Although the Cronbach's alpha reliability coefficients for four of the five subscales were below .70, the authors justified their use because of poor reliability reported in previous studies and the fact that each subscale only had two items, making good reliability less likely (Boysen et al., 2015). These are good examples of how to handle situations when data that have already been collected are not reliable or valid.

Another common occurrence in the reporting of internal consistency is listing the reliability of an established scale from previously published studies rather than reporting the reliability of that scale in the present study. Of the 47 studies that reported scales, 30 studies reported Cronbach's alphas from the current study and six reported internal consistency values from previously published studies, but not the current sample. An example of using previously published internal consistency values is a study looking at changing stereotypes of older adults (Wurtele & Maruyama, 2013), as measured by the Fraboni Scale of Ageism (FSA; Fraboni, Saltstone, & Hughes, 1990). The authors reported support for internal consistency, construct validity, and test-retest reliability from previously published work but did not report any sources of reliability from the current study (Wurtele & Maruyama, 2013). Although using previously established scales with proven reliability and validity is certainly a best practice, reporting continued support for the psychometric properties of a scale furthers researchers' confidence in their use and applicability in diverse contexts. In addition, because Cronbach's alpha is an easily accessible statistic, not reporting a current alpha value may seem like the author is trying to hide a poor reliability value. An example of continued psychometric testing of an already established scale has been the assessment of the Teacher Behaviors Checklist (TBC; Keeley, Smith, & Buskist, 2006) in various studies with reported reliability in these different situations, including adapting to undergraduate teaching assistants (Filz & Gurung, 2013), and correlating the TBC with academic self-concept and motivation (Komarraju, 2013).

A second form of reliability is inter-rater reliability, established when more than one coder observes behaviors or rates qualitative data and assesses their agreement on the resulting rating. Inter-rater reliability also has a strong record in recent published SoTL research in psychology. In last ten issues of *Teaching of Psychology* examined*,* 17 of 22 studies (77.3%) where behaviors or writing assignments were rated as the SoTL outcome, inter-rater reliability was assessed resulting in reliable coding. Although not perfect agreement, most studies had an 80% agreement rate or better. Where there were discrepancies, most studies used a process of consensus on the ratings that were not in agreement when coded individually. A particularly good example of reporting inter-rater reliability and the process of resolution of discrepancies is Lawson and Crane's (2014) study of ideomotor action to increase critical thinking. In the study, both an experimental and control group were asked an open-ended question to provide an explanation for a given phenomenon. Student responses were independently coded by two blind coders for presence of ideomotor action in their response. They achieved excellent inter-rater rater reliability (96%) and then stated that, "disagreements were discussed to consensus" (Lawson & Crane, 2014, p. 54). The practice of discussion and consensus where there are discrepancies in coding is a best practice in the literature with several exemplars of how to report the process (e.g., Daniel & Braasch, 2013; Simon-Dack, 2014).

The last form of reliability, test-retest reliability, is the assessment of correlation of the same outcome over time (e.g., a two to eight week time period) to understand the consistency of the scale over time. Test-retest reliability was assessed far less often in the SoTL in psychology literature examined than the other types of reliability. Only two of the articles in the past ten issues of *Teaching of Psychology* reported test-retest reliability (not including studies solely focused on establishing psychometric properties of the scale used), and those values were from previously published studies, not from the current sample. One example of reporting test-retest reliability as a routine part of the Method section is Wurtele and Maruyama's (2013) study of older adult stereotypes. As discussed earlier in this chapter, the researchers used the FSA scale to test stereotyping, and in their materials section they reported test-retest correlation coefficients from previously published work (Wurtele & Maruyama, 2013). Perhaps because test-retest reliability is used to establish that scales do not change over time, this type of reliability is rarely reported in SoTL work because change is expected over time in so many SoTL projects, making test-retest reliability potentially less relevant.

## Assessing and Reporting Validity

In addition to forms of reliability, another important component of the psychometric properties of scales is validity. As is noted earlier in this e-book (see Hussey & Lehan, 2015), there are several forms of validity (e.g., face, construct, criterion) which constitute ways in which researchers test if variables are actually measuring the constructs they are intended to measure. Although reliability was consistently assessed and reported in SoTL research using scales, validity was far less likely to be assessed and reported. Only eight of 47 (17%) published studies using scales to measure SoTL outcomes in the past ten *Teaching of Psychology* issues (not including scale development-focused articles) assessed any form of validity and reported those results. Two of those eight studies reported previously published validity information rather than validity from the present study (e.g., Wielkiewicz & Meuwissen, 2014). One example of a study that assessed and reported validity in the current study is Buckelew, Byrd, Key, Thornton and Merwin's (2013) study of perception of luck or effort leading to good grades. The researchers adapted a scale of attributions for good grades from a previous study, noting that no published validity information was available for the original scale. Researchers then established validity through a pilot test where they correlated values of their adapted scale to a previously established criterion and found significant positive correlations with the total scale and subscales. This example of establishing criterion validity through a pilot test is one way to ensure valid measures before collecting actual study data. However, it should be noted that validity is truly established over time and through many studies, providing a compelling reason why routinely reporting validity is vitally important despite the dearth of studies assessing it. Validity in its several forms (e.g., content, construct, and criterion) are crucial to researchers' confidence that the data collected are measuring the constructs they were intended to measure. SoTL researchers are urged to incorporate validity testing and reporting into the work they do.

## The Development and Validation of Scales

There have been some great advances in the ways in which scales are developed and validated. The practice of using both exploratory analyses (e.g., principal factor analysis) and then confirmatory factor analysis has greatly enhanced researchers' confidence in the psychometric properties of the scales used to measure outcomes (Noar, 2003). The increasingly common use of structural equation modeling has enhanced scale development by providing easily accessible ways to conduct confirmatory factor analysis and latent variable modeling (Noar, 2003).

There were five studies out of 141 articles published in the last ten issues of *Teaching of Psychology* devoted to the development and validation of a scale. These studies employed some of the most thorough methodology in scale development and validation. One particularly good example of such scale validation is Renken, McMahan, and Nitkova's (2015) initial validation of the Psychology-Specific Epistemological Belief Scale (Psych-SEBS). In this two-part study, researchers first drafted and then refined an item pool using exploratory factor analysis, followed by confirmatory factor analysis and an assessment of internal consistency, test-retest reliability, and convergent validity of the Psych-SEBS. In the second study, researchers assessed the criterion validity by comparing the Psych-SEBS with an established criterion and also tested the incremental validity of the Psych-SEBS above and beyond the variance explained by that criterion. By all measures, the Psych-SEBS 13-item scale is a reliable and valid measure of psychology-specific epistemological beliefs, and the way in which the researchers conducted this scale validation is a model for others in the field. Another example of scale validation is Rogers' (2015) further validation of the Learning Alliance Inventory (LAI). Using correlation and path analysis, Rogers established internal and test-retest reliability, criterion and convergent construct validity, as well as established that the LAI predicted student learning beyond the already established predictors of immediacy and rapport (Rogers, 2015). This study is an exemplar because of the use of the most recent statistical techniques (e.g., path analysis) to thoroughly assess all of the psychometric properties of the LAI scale.

## Discussion

### Reliability and Validity Challenges Particular to SoTL Research

One of the challenges of doing SoTL research is that the classroom is not only the place where data are collected, but it is also the place where teaching, learning, and assessment are happening. Because instructors are often teaching and simultaneously doing SoTL research in their own classes, being mindful not to overburden the students or compromise learning in the pursuit of data collection can create challenges. Sometimes using existing sources of data make the overlap of teaching and research a little less complicated, but such methods of overcoming SoTL research obstacles can come at a cost.

#### *Student Ratings of Instructors*

One of the most common sources of data is student ratings of instructors (SRIs), often used to determine the effectiveness of the teacher and the course (Heckert, Latier, Ringwald, & Silvey, 2006). There is an extensive body of literature assessing the impact of different learning experiences on SRIs, including the influence of immersion scheduling (Richmond, Murphy, Curl, & Broussard, 2015), professor-student rapport (Ryan, Wilson, & Pugh, 2011), humor (Richmond, Berglund, Epelbaum, & Klein, 2015), and student engagement (Handlesman, Briggs, Sullivan, & Towler, 2005). Most of the time the SRI that is used is the one that the researcher's home university uses, which normally does not have a track record of established reliability and validity (although sometimes these SRIs are reliable, e.g. Richmond et al., 2015). In fact, there is a growing literature noting the lack of reliability and validity of SRIs in measuring student satisfaction or learning (Beran & Violato, 2005; Catano & Harvey, 2011; Clayson, 2009; Heckert et al., 2006). Yet, SRIs still remain a very popular measure in SoTL research (and as a determining factor in faculty merit and promotion) despite this fact (Clayson, 2009). However, it is important to note that SRIs are not structured in a way to truly measure learning, but rather students' perceptions of teacher effectiveness. It makes sense to say that an effective teacher enhances student learning, but learning is not the intended target construct of SRIs. In addition, because SRIs are not standardized across universities, comparing SRIs is not possible in their current form (Richmond et al., 2015).

### *Grades*

Another common source of already-existing data in the learning environment is grade point average (GPA). Although GPA is generally thought to be a reliable and valid measure of academic achievement, and therefore a good proxy for learning, that may not necessarily be the case (Imose & Barber, 2015). The literature has established that GPA has good internal consistency, although the test-retest reliability and the comparison of GPA standards across schools have not been established. For example, in one study focused on business majors, four-year cumulative GPA was reliable (Cronbach's alpha = .94), showing excellent internal consistency, but they were not able to assess test-retest reliability or compare across institutions (Bacon & Bean, 2006). The different standards employed by universities also impact comparability. For instance, a 3.5 GPA at one university may make the student in the top 10% of the class, whereas that same GPA would be in the top 25% at another university (Imose & Barber, 2015). In addition, although GPA may be measuring general cognitive ability and conscientiousness, GPA has not been established as a valid measure of learning. For instance, overall GPA had strong predictive validity of individual measures of academic performance, but mainly when that performance was reliably measured and was based on the individual student's ability and effort, not necessarily their learning (Bacon & Bean, 2006).

Student ratings of instructors and GPA are easily accessible outcome measures in SoTL because these variables are already being assessed as part of the university class experience. Because of their accessibility, using already-measured variables reduces the burden on students. However, the accessibility of these sources of data does not mean that they are the best measures of learning. Continuing to establish the reliability and validity of these commonly used measures in SoTL in diverse situations is crucial to ensure confidence that the measures are high quality and are tapping into the constructs we intend to measure.

## Conclusion

Reliability and validity are paramount to researchers across fields to insure that information gathered and reported is measuring what the researcher wants to measure and measuring it well. In the SoTL literature in psychology, internal consistency has been assessed and reported very consistently, while other forms of reliability and validity have not. This is a call for SoTL researchers to be more thorough in their reporting of psychometrics. Publishing the reliability and validity of scales for every new study will elevate the quality of the body of research in SoTL. Assessing the psychometric properties of newly developed scales is also important to provide researchers with novel, reliable, and valid ways to measure SoTL outcomes. Continuing to report the reliability and validity of previously established scales also engenders more confidence that the ways in which we measure SoTL outcomes are psychometrically sound in varied environments and with diverse samples. Particularly important is increasing the focus on establishing the validity of scales, especially due to the finding that the reporting of validity was significantly lower than reliability in recent SoTL in psychology. Lastly, in looking at some of the common ways we measure learning, particularly student ratings of instructors and GPA, we must do better in standardizing and measuring reliability and validity so that we can build a solid track record for the most commonly used measures in the field. This would include studies across institutions with standardized SRI and GPA standards to establish reliability and validity with a wider and more diverse population.  In the other chapters in this e-book, suggestions for reliable and valid ways to measure the most common SoTL outcomes have been collected. Utilizing these scales and furthering the case for their reliability and validity will be vitally important to the future success of SoTL research and our understanding of how students learn.

# References

References marked with an asterisk indicate a scale.

*Bacon, D. R., & Bean, B. (2006). GPA in research studies: An invaluable but neglected opportunity. *Journal of Marketing Education, 28*, 35–42. doi.org/10.1177/0273475305284639

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*, 593-601. doi: 10.1080/02602930500260688

*Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology, 42*, 109-118. doi: 10.1177/0098628315569922

*Boysen, G. A., Richmond, A. S., & Gurung, R. A. R. (2015). Model teaching criteria for psychology: Initial documentation of teachers' self-reported competency. *Scholarship of Teaching and Learning in Psychology, 1*, 48-59. doi: 10.1037/stl0000023

*Buckelew, S. P., Byrd, N., Key, C. W., Thornton, J., & Merwin, M. M. (2013). Illusions of a good grade: Effort or luck? *Teaching of Psychology, 40*, 134-138. doi: 10.1177/0098628312475034

*Catano, R. M., & Harvey, S. (2011). Student impressions of teaching effectiveness: Development and validation of the Evaluation of Teaching Competencies Scale (ETCS). *Assessment & Evaluation in Higher Education, 36*, 701-717. doi: 10.1080/02602938.2010.484879

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*, 16-30. doi:10.1177/0273475308324086

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

*Daniel, F., & Braasch, J. L. G. (2013). Application exercises improve transfer of statistical knowledge in real-world situations. *Teaching of Psychology, 40*, 200-207. doi: 10.1177/0098628313487462

De Vellis, R. F. (1991). *Scale development: Theory and applications* (Applied Social Research Methods Series, Vol. 26). London: Sage.

*Filz, T., & Gurung, R. A. R. (2013). Student perceptions of undergraduate teaching assistants. *Teaching of Psychology, 40*, 48-51. doi: 10.1177/0098628312465864

*Fraboni, M., Saltstone R., & Hughes, S. (1990). The Fraboni scale of ageism (FAS): An attempt at a more precise measure of ageism. *Canadian Journal on Aging, 9*, 56-66.

*Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528. doi.org/10.1002/tl.20071

Heckert, T. M., Latier, A., Ringwald, A., & Silvey, B. (2006). Relation of course, instructor, and student characteristics to dimensions of student ratings and teaching effectiveness. *Journal of Instructional Psychology, 28*, 161-168.

Hussey, H. D., & Lehan, T. J. (2015). A primer on scale development. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Imose, R., & Barber, L. K. (2015). Using undergraduate grade point average as a selection tool: A synthesis of the literature. *The Psychologist-Manager Journal, 18*, 1-11.

*Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behaviors checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*, 84-91. doi: 10.1207/s15328023top3302_1

*Komarraju, M. (2013). Ideal teaching behaviors: Student motivation and self-efficacy predict preferences. *Teaching of Psychology, 40*, 104-110. doi: 10.1177/0098628312475029

*Lawson, T. J., & Crane, L. L. (2013). Dowsing rods designed to sharpen critical thinking and understanding of ideomotor action. *Teaching of Psychology, 41*, 52-56. doi: 10.1177/0098628313514178

Layous, K., Nelson, S. K., & Legg, A. M. (2015). Measuring well-being in the scholarship of teaching and learning. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

*Maybury, K. K. (2013). The influence of a positive psychology course on student well-being. *Teaching of Psychology, 40*, 62-65. doi: 10.1177/0098628312465868

Noar, S. M. (2003). The role of structural equation modeling in scale development. *Structural Equation Modeling, 10*, 622-647.

*Renken, M. D., McMahan, E. A., & Nitkova, M. (2015). Initial validation of an instrument measuring psychology-specific epistemological beliefs. *Teaching of Psychology, 42*, 126-136. doi: 10.1177/0098628315569927

*Richmond, A. S., Berglund, M. B., Epelbaum, V. B., & Klein, E. M. (2015). *a* + ($b_1$)Professor-student rapport + ($b_2$) humor + ($b_3$) student engagement = (*Y*) student ratings of instructors. *Teaching of Psychology, 42*, 119-125. doi: 10/1177/0098628315569924

*Richmond, A. S., Murphy, B. C., Curl, L S., & Broussard K. A. (2015). The effect of immersion scheduling on academic performance and students' ratings of instructors. *Teaching of Psychology, 42*, 26-33. doi: 10/1177/0098628314562674

*Rogers, D. T. (2015). Further validation of the Learning Alliance Inventory: The roles of working alliance, rapport, and immediacy in student learning. *Teaching of Psychology, 42*, 19-25. doi: 10.1177/0098628314562673

*Ryan, R. G., Wilson, J. H., & Pugh, J. L. (2011). Psychometric characteristics of the professor-student rapport scale. *Teaching of Psychology, 38*, 135-141. doi: 10.1177/0098628311411894

*Simon-Dack, S. L. (2014). Introducing the action potential to psychology students. *Teaching of Psychology, 41*, 73-77. doi: 10.1177/0098628313514183

*Wielkiewicz, R. M., & Meuwissen, A. S. (2014). A lifelong learning scale for research and evaluation of teaching and curricular effectiveness. *Teaching of Psychology, 41*, 220-227. doi: 10.1177/0098628314537971

*Woodzicka, J. A., Ford, T. E., Caudill, A., & Ohanmamooreni, A. (2015). A successful model of collaborative undergraduate research: A multi-faculty, multi-project, multi-institution team approach. *Teaching of Psychology, 42*, 60-63. doi: 10/1177/0098628314549711

*Wurtele, S. K., & Maruyama, L. (2013). Changing students' stereotypes of older adults. *Teaching of Psychology, 40*, 59-61. doi: 10.1177/0098628312465867

# Section 2: Scales for Use in the Scholarship of Teaching and Learning

# Chapter 6: Measuring Learning and Self-Efficacy

Pam Marek, Adrienne Williamson, and Lauren Taglialatela

Kennesaw State University

Learning and self-efficacy are closely intertwined. In a meta-analysis of 36 studies in educational settings (from elementary school through college) with a variety of performance measures (e.g., standardized tests, course grades, GPA), perceived self-efficacy positively predicted academic performance (Multon, Brown, & Lent, 1991). Moreover, a burgeoning literature has revealed that learning and self-efficacy both relate to multiple variables such as motivation (Pintrich & DeGroot, 1990), self-regulation (Pintrich, Smith, Garcia, & McKeachie, 1991; Zimmerman & Martinez-Pons, 1988), and metacognitive awareness (Schraw & Dennison, 1994) that promote success.

Compared to most constructs addressed in this e-book, the measurement of learning is unique in several respects. First, the construct of learning is broader than most others addressed. Second, the definition of learning is debated. For example, Barron and colleagues (2015) have pointed out that positioning learning as a change in behavior attributable to experience may be too limiting to apply across disciplines; instead, they suggest that there may be greater consensus with the definition of learning as, "a structured updating of system properties based on the processing of new information" (p. 406). Third, measurement of learning varies greatly within two broad types of assessment: formative and summative.

Formative assessment involves classroom techniques that inform both students and faculty about how well students are grasping concepts in "real time" (Wiliam & Black, 1996). An advantage of this type of assessment is that it provides instant feedback that may potentially highlight ways to enhance the learning process, so students and faculty are able to initiate immediate changes in their behaviors. This type of assessment can scaffold learning for current students, whereas only future students may benefit from end-of-semester evaluations (Angelo & Cross, 1993). Summative assessment involves determining what students have learned in culmination about a particular topic or unit compared to specific criteria at a particular point during in the learning process (Dunn & Mulvenon, 2009). Such assessments include both graded assignments within a course and standardized tests at the end of a program (Wiliam & Black, 1996).

A major distinction between formative and summative assessments relates to how the results are used rather than the actual format or type of assessment. In fact, the same type of assessment may be used as either a formative or summative evaluation, or even for both. For example, rubrics can be used in either formative or summative contexts. Quizzes may also serve multiple purposes. Although instructors often use a quiz as a graded, summative assessment at the end of a unit, if they provide feedback to clarify students' understanding, then the quiz would also serve a formative purpose for future examinations.

Several techniques for assessment of learning, such as exams, presentations, and papers, may not be classified as or evaluated by using traditional scales. These assessments typically vary among courses and among instructors teaching a given course to reflect differences in learning objectives linked to content and skills. Considering the nature of learning, both as a process and an outcome, evaluative tools often stretch beyond traditional scales. Ideally, these measurements of learning are constructed and scored to be consistent with course and assignment goals. Montgomery (2002) discusses the important role that rubrics play in evaluating such assignments. When using a rubric, instructors clearly define criteria for identifying different levels of performance on specific elements of an assignment; students should use these criteria to evaluate their own work as they compose it (Montgomery, 2002; Reddy & Andrade, 2010). (See Mueller, 2014 for detailed information on rubric construction.)

In this chapter, we describe both formative and summative assessments. To illustrate formative assessments, we discuss classroom assessment techniques (CATs) because they are one of the most common tools in this category. We also discuss rubrics used for formative purposes. As examples of summative assessments, we discuss rubrics for writing assignments and standardized tests. We focus on writing because of its prevalence and importance in psychology and because of the value of writing skills in many employment settings. For example, in a national survey, chairs of 278 psychology departments indicated the number of courses completed by at least 80% of their students that emphasized skills specified by the American Psychological Association (APA). Responses indicated that 59% of programs emphasized writing APA style articles, and 68% of programs emphasized writing in other forms in at least four of their classes (Stoloff, Good, Smith, & Brewster, 2015). Regarding the workplace, in a job outlook survey, 73% of employers selected written communication skills as a valued attribute on candidates' resumes (National Association of Colleges and Employers, 2015). As a second type of summative assessment, we chose to focus on standardized tests instead of course exams because of the variability that exists for exams within and between specific courses. In addition, coverage of standardized tests allows us to expand our discussion from classroom evaluation to program evaluation. The chapter concludes with information about measures of perceived learning and perceived self-efficacy.

## Formative Assessments

### Classroom Assessment Techniques

Angelo and Cross (1993) recommended using formative assessments before presentation of new material to assess prior knowledge, during presentation to determine students' level of understanding, and after presentation to strengthen students' learning. It is imperative for the instructor to provide timely feedback from the assessment for the students to benefit fully. There are dozens of frequently recommended CATs that can be used as formative assessments such as Concept Mapping, Quizzes, Student-Generated Exam Questions, Reaction Papers, Polling the Class, Pro and Con Grids, Think-Pair-Share (students are asked to think briefly about an answer to a question provided by the instructor, then pair up with another student to discuss their answers, and finally to share the information with the class), and Jigsaw (teams of students become "experts" within a specific area, and then each member of the original teams joins a new team to teach them about their area of expertise). See Angelo and Cross (1993) for a detailed description of 50 CATs.

Comparisons of classes that repeatedly used CATs with those that did not use CATs yield inconsistent findings regarding their effect on grades. Repeated use of 5-question, multiple-choice, nongraded, formative quizzes was found to increase exam scores in a psychology course (Short & Martin, 2012). However, Cottell and Harwood's (1998) examination of the effect of using multiple CATs (Background Knowledge Probe, Minute Paper, Feedback Form, Directed Paraphrasing, Pro and Con Grid, What did I learn from the exam?, Classroom Assessment Quality Circle Group Instructional Feedback Technique, and Group-Work Evaluation Form) in an accounting class revealed no significant differences between classes in grades for the course or on exams, group projects, or quizzes. Similarly, in a criminal justice class, Simpson-Beck (2011) found no significant differences in grades for the course, on chapter tests, or on a cumulative final exam when using the Muddiest Point, where students were asked to identify the most confusing point from the day's class. Given the variability in using formative assessment (e.g., choice of CAT, course, frequency of administration, quality of implementation), additional studies are needed to determine the best practices of CAT application.

One of the most commonly used CATs is the Minute Paper (also known as the One-Minute Paper or the Half-Sheet Response), a two-question measure that has been described by multiple authors (e.g., Angelo & Cross, 1993; Lom, 2012; Stead, 2005). The first question on the Minute Paper ("What was the most important thing you learned during this class?") requires students to practice identifying the most important points from the lecture. The second question ("What important question remains unanswered?") encourages students to reflect on the overall material to determine what they understood during the class period and what they did not comprehend as completely. This assessment tool is brief and easy to use, and the questions can be adapted to address more focused, course-specific material that can inform an instructor's individual learning objectives. The instructor can quickly read and group the responses and provide feedback either at the beginning of the next class or via email to the entire class or to individual students.

Angelo and Cross (1993) provided an example of the usefulness of this CAT. At the end of each class session, a statistics instructor used a slightly modified version of the Minute Paper by asking students to select the five most important points of the day's class and ask one or two questions. Initially, the instructor found that the class collectively listed approximately 20 main points, some of which were less important details or included partially or completely incorrect information. This discovery prompted the instructor to modify his teaching; at the beginning of each class, he listed many of the points the students had submitted and discussed their relative importance. Additionally, he listed the points he considered most important, compared the lists, and addressed common questions. This approach reduced the students' list of main points from 20 to less than 10 within a month. These results show how the use of a simple CAT can provide information about students' level of understanding and can lead to immediate changes in teaching.

In general, many faculty and students who have used CATs have reported perceived benefits that include better student learning (Anderson & Burns, 2013; Stead, 2005), more active engagement, and an enhanced classroom environment (Soetaert, 1998). However, Cottell and Harwood (2005) found no difference in perceived learning between students in classes who used CATs and classes that did not use CATs. In a review of the Minute Paper, Stead (2005) found that most researchers reported significantly higher test scores for students in courses where this tool was used. Chiou, Wang, and Lee (2014) also found evidence of enhanced learning and reported an additional benefit of reduced course-related anxiety in a statistics course. However, both Angelo and Cross (1993) and Stead caution against the overuse of this assessment to avoid it "being seen as a gimmick" (p. 153, Angelo & Cross) or becoming tedious for both the students and instructor. Also see Angelo and Cross and Stead for further discussion of the advantages and disadvantages of this assessment tool.

### Formative Rubrics

Greenberg (2015) evaluated a rubric (originally designed for summative use) for its potential utility as a formative tool. The rubric is for an APA-style research report and contains 60 learning objectives encompassing content, expression, and formatting. Each outcome is rated on a 4-point scale ranging from *absent* to *achieved*. Students in introductory psychology (*n* = 78) and advanced (*n* = 60) psychology courses were told to use the rubric when creating a writing assignment specific to the course, whereas students in other sections of the same courses were not provided with the rubric (*n* = 68, *n* = 58, respectively). Students who used the rubric scored higher on the writing assignment than students who did not use the rubric, indicating that utilization of the rubric has formative benefits.

Greenberg (2015) also found the rubric to be useful during revision. After students used the rubric prescriptively to prepare their own papers, each student was given a peer's paper and asked to grade it using the rubric. Students were then given the opportunity to revise their own papers; there was significant improvement in paper quality following this revision. Overall, results indicated that rubrics were helpful tools during both the writing and revising phases of paper production.

Lipnevich, McCallen, Miles, and Smith (2014) compared the formative benefit of using detailed rubrics versus exemplar papers. After completing a rough draft, students were randomly assigned to one of three groups: rubric, exemplar papers, or rubric and exemplar papers. Students were instructed to use these materials to revise their papers. Results indicate that all three groups demonstrated significant improvement from first to second draft; however, the rubric [only] group showed the most improvement (Cohen's *d* = 1.54). The rubric has 10 dimensions that follow the basic components of a research paper (e.g., description of research project, study design, study materials). Each dimension is scored on a scale ranging from 1 (*below expectation)* to 3 (*exceeds expectation)*. Examples are provided in the rubric to help distinguish between grade levels within each dimension.

## Summative Assessments

### Summative Rubrics

Although APA lists several standardized assessments designed to measure various aspects of communication (e.g., Collegiate Assessment of Academic Proficiency Writing Essay Test and Writing Skills Test, WorkKeys Foundational Skills Assessment, and Collegiate Level Assessment), and the Association of American Colleges and Universities (AACU) provides a Written Communication rubric (https://www.aacu.org/value/rubrics/written-communication), none focus on communication in psychology, or more globally, on scientific writing. Although writing, specifically scientific writing, is typical in natural and social science curricula, few standardized rubrics to evaluate such work have been developed, tested, and made accessible. In this section, we identify empirically-tested rubrics for summative use in scientific writing that can be easily accessed and implemented or adapted for classroom use in psychology.

Stellmack, Konheim-Kalkstein, Manor, Massay, and Schmitz (2009) created and evaluated the reliability and validity of a rubric used to score APA-style research papers. The researchers focused on only the Introduction section of the rubric; however, updated and expanded rubrics for all sections (e.g., Introduction, Method, and Results/Discussion) are available at http://www.psych.umn.edu/psylabs/acoustic/rubrics.htm. Students were instructed to write an APA-style Introduction incorporating five sources. The rubric encompasses eight dimensions that are fundamental aspects of this type of writing assignment: APA formatting, literature review, purpose of study, study description and hypothesis, overall organization/logical flow, sources, scientific writing style, and composition/grammar/word choice. Students earn up to 3 points for each dimension, and examples of the grade levels within each dimension are provided. For example, if either the study description or the hypothesis is missing, then 1 point would be earned for the study description and hypothesis dimension. Alternatively, if both the description and hypothesis are provided, but either is unclear, then 2 points would be earned. Scores for the Introduction section range from 0 to 24.

Stellmack et al. (2009) evaluated the interrater and intrarater reliability of this rubric. For three graders, interrater reliability was defined both liberally (i.e., scores for a dimension were within 1 point across the three graders) and conservatively (i.e., scores for a dimension were equal across the three graders). Agreement ranged from .90 (liberal) to .37 (conservative). For intrarater reliability, graders re-evaluated a subset of their original papers 2 weeks after the initial assessment. Agreement ranged from .98 (liberal) to .78 (conservative). In such consistency estimates, values at or above .70 are viewed as acceptable (Stemler, 2004).

Beyond psychology-based writing, the Biology Thesis Assessment Protocol (BioTAP) provides a systematic method to assess scientific writing in the biological sciences, and the components of the rubric are applicable across natural and social science disciplines and could be adapted for use within psychology. Reynolds, Smith, Moskovitz, and Sayle (2009) used BioTAP to evaluate undergraduate theses in biology. The rubric has three categories: higher-order writing issues (e.g., "*Does the thesis make a compelling argument for the significance of the student's research within the context of the current literature?"*), mid- and lower-order writing issues ("*Is* *the thesis clearly organized?")*, and quality of scientific work. Across the three categories, there are 13 individual criteria, each scored on a 3-point scale (i.e., *no, somewhat, yes*). The complete rubric and instructional method can be accessed at www.science-writing.org/biotap.html.

Reynolds and colleagues (2009) evaluated the interrater reliability for the nine items that comprised the higher-, mid-, and lower-order issues. The items in the quality of scientific work dimension require specialized knowledge of the paper content, and because the evaluators had different areas of expertise within the biological sciences, this category was not included in the reliability assessments. The percent agreement for individual criteria within the higher, mid-, and lower-ordered issues ranged from .76 to .90, with Cohen's kappa ranging from .41 to .67, indicating sufficient reliability.

Timmerman, Strickland, Johnson, and Payne (2011) created a universal rubric for assessing scientific writing called the Rubric for Science Writing. The rubric was designed to evaluate students' empirical research reports in a genetics, evolution, or ecology course. The rubric is used to evaluate 15 dimensions including those related to an introduction, hypotheses, method, results, discussion, use of primary literature, and writing quality. Each dimension is scored on a 4-point scale ranging from *not addressed* to *proficient*. Similar to Stellmack and colleagues' (2009) rubric, examples of grade levels within each dimension are provided.

Timmerman and colleagues (2011) assessed reliability using generalizability (*g*) analysis (Crick & Brennan, 1984). This assessment determines the amount of variation in rater scores that can be attributed to actual differences in the quality of the paper and separates out variation due to individual raters or rater-assignment interactions (Brennan, 1992). Using this metric, a score of 1.0 indicates that all of the variations in the scores are due to differences in the quality of the student papers and a score of 0.0 indicates that none of the variation in the scores is due to differences in the quality of the paper. Generalizability coefficients at or above .80 are typically considered acceptable (Marzano, 2002). Generalizability analysis for the Rubric for Science Writing indicates that 85% (*g* = .85) of the variation in scores across three biology laboratory papers was attributable to actual differences in the quality of the student work. Therefore, applying the rubric under similar circumstances should produce reliable scores.

## Standardized Assessments

Standardized assessments used to evaluate learning near the end of an undergraduate program in psychology include the Educational Testing Service's (ETS) Major Field Test in Psychology (MFT-P; https://www.ets.org/mft/about/content/psychology), the Area Concentration Achievement Test in Psychology (ACAT-P) by PACAT, Inc. (http://www.collegeoutcomes.com), and the ETS Graduate Record Examination Subject Test in Psychology (GRE-P; https://www.ets.org/gre/subject/about/content/psychology). Each of these assessments measures knowledge about topics and concepts typically taught in psychology courses that are often part of the major. In the most recent data available for each test, the MFT-P has been used by 357 programs (25,895 individuals) and the ACAT-P has been used by approximately 170 programs. In a national survey of psychology programs (Stoloff, 2015), 50 of 278 (18%) programs indicated they used one of these two measures.

The MFT-P is a 2-hr test that includes 140 multiple-choice questions and provides scores for individual students and departmental means on the total test and four subtests: learning (~5–7% of the test), cognition (~9–11%), and memory (~3–5%); sensory and perception (~3–5%) and physiology (~10–12%); clinical and abnormal (~10–12%) and personality (~6–8%); and developmental (~10–12%) and social (~10–12%). Departmental means are also provided for six assessment indicators: memory and cognition; perception, sensation, and physiology; developmental; clinical and abnormal; social; and measurement and methodology. Total score and subscores are reported as scaled scores. The scaled range for the total score is 120 to 200 and for the subscores is 20 to 100. A Comparative Data Guide (https://www.ets.org/s/mft/pdf/acdg_psychology.pdf) is provided by ETS that includes tables of scaled scores and percentiles that can be used to compare total scores and subscores for individual students and to compare means of total scores, subscores, and assessment indicator scores for institutions. The comparative data set includes all U.S. seniors who were tested using the most recent version of the test.

Researchers have examined factors related to performance on the MFT-P. Both overall GPA and psychology GPA were correlated with MFT-P total score and all subscores; most studies found that all SAT scores and the total number of psychology credits completed were correlated with the MFT-P total score and most of the subscores (Dolinsky & Kelley, 2010; Gallagher & Cook, 2013; Stoloff & Feeney, 2002 [for an exception, see Pinter, Matchock, Charles, & Balch, 2014]). However, completion of only a few specific courses was related to MFT-P performance. These specific courses varied somewhat among studies, but included a combination of Abnormal Psychology, Counseling Psychology, Physiological Psychology, Social Psychology, and History and Systems (Dolinsky & Kelley, 2010; Gallagher & Cook, 2013; Stoloff & Feeney, 2002).

The ACAT-P (http://www.collegeoutcomes.com) can be used to assess 13 content areas: abnormal, animal learning and motivation, clinical and counseling, developmental, experimental design, history and systems, human learning and cognition, organizational behavior, personality, physiological, sensation and perception, social, and statistics. The content of this exam is flexible and allows departments to tailor its content by choosing the areas they wish to assess. Departments have the choice of assessing 10 areas (in 120 min), eight areas (in 96 min), six areas (in 72 min), or four areas (in 48 min). The ACAT-P can be used to assess general knowledge of seniors at the end of their program or as a pre-posttest assessment to examine change in knowledge over the undergraduate program. As a national comparison group, mean scores in each content area within psychology are provided for graduating seniors who completed the test during each year for the last 13 years.

The GRE-P (https://www.ets.org/gre/subject/about/content/psychology) has approximately 205 multiple-choice, five-option questions. Students receive an experimental subscore, a social subscore, and a total score. The two scaled subscores can range from 20 to 99 in one-point increments, and the total scaled score can range from 200 to 990 in 10-point increments. The experimental subscore is based on 40% of the test questions and is related to learning (3-5%), language (3-4%), memory (7-9%), thinking (4-6%), sensation and perception (5-7%), and physiological/behavioral neuroscience (12-14%). The social subscore is based on approximately 43% of the test questions and is related to clinical and abnormal (12-14%), lifespan development (12-14%), personality (3-5%), and social (12-14%). The total score is based on questions that contribute to the two subscores along with the remaining test questions that are related to other areas, including general (4-6%) and measurement and methodology (11-13%). Test administration takes 2 hr 50 min. According to ETS (https://www.ets.org/gre/subject/about/content/psychology), research has provided evidence for construct, content, predictive, and external validity of the GRE subject tests.

## Perceived Learning Assessment

### Two Items

Many instructors and researchers have struggled with how to best measure learning in a way that is not based on the specific content within a course. This dilemma has led some individuals to use measures of perceived cognitive learning. For example, Richmond, Gorham, and McCroskey (1987) created their own two-item measure of perceived learning. On this measure, students are asked to indicate on a 10-point scale, ranging from 0 (*nothing*) to 9 (*more than in any other class*), how much they had learned in the class and how much they think they could have learned in the class if they had the ideal instructor. The score on the first item may be taken as a learning score whereas a "learning loss" score can be derived by subtracting the first-item score from the second-item score. Subsequent studies have provided evidence for the test-retest reliability of both the learning (.85) and "learning loss" (.88) scores (McCroskey, Sallinen, Fayer, Richmond, & Barraclough, 1996) as well as the criterion validity ("learning loss" scores were correlated with quiz grade) (Chesebro & McCroskey, 2000) of this perceived learning measure.

### Learning Indicators Scale

Frymier, Shulnian, and Houser (1996) developed an assessment for learner empowerment, which included a learning indicators scale (α reliability = .84) based on nine items related to behaviors that students may engage in to enhance learning. However, Frymier and Houser (1999) suggested that four of the nine items on the original scale were related to communication, and this focus might introduce a potential confound for students who tend to be apprehensive about communication. Thus, Frymier and Houser created a revised learning indicators scale that eliminated this problem. The new scale contained seven items (four new items and three from the original scale) on which students rate how often they engage in given behaviors on a 5-point Likert scale ranging from 0 (*never*) to 4 (*very often*). The scale is reliable (α = .85) and shows both construct validity (scores were positively correlated with instructor nonverbal immediacy, learner empowerment, and state motivation – all constructs typically related to learning) as well as criterion validity (scores were positively correlated with a measure of affective learning and an assignment grade).

## Course Outcomes Scale

Centra and Gaubatz (2005) examined the Course Outcomes Scale on ETS's Student Instructional Report (SIR II; https://www.ets.org/sir_ii/about/) as a measure of perceived student learning. This scale includes five items that are rated on a 5-point scale ranging from 1 (*much more than most courses*) to 5 (*much less than most courses*). The five items address perceptions of learning of course content specifically (2 items) as well as more general learning outcomes (3 items). Other components and scales on the SIR II, including scores on the Overall Evaluation, Student Effort and Involvement Scale, and the Assignments, Exams, and Grading Scale were significant predictors of perceived learning as measured by the Course Outcome Scale. Perceptions of learning measured on course evaluations are related to overall course satisfaction and rating of course instructor.

## Student Estimates

Another measure of perceived learning is students' self-assessment of their performance on an assignment. The type and timing of these self-reports of perceived learning vary. Students' estimates may be made in response to a single question (e.g., provide an overall estimate of expected performance on a graded measure of learning) or on an item-by-item basis on the measure (Schraw, 2009). Estimates can be made prior to (prediction) or after (postdiction) completing the actual knowledge assessment. Moderate correlations between students' estimates and instructors' scores on performance measures have been reported in two meta-analyses; Falchikov and Baud (1989) found a mean correlation of .39 (based on 45 correlation coefficients), and Sitzmann, Ely, Brown, and Bauer (2010) found a mean correlation of .34 (based on 137 effect sizes). There are a number of factors that can influence the strength of the relationship between perceived and performance measures of learning including competence of the learner (experts [Kruger & Dunning, 1999] and higher-performing students [Bol et al., 2005; Hacker et al., 2000] made more accurate self-reports; higher-performing students tended to underestimate their level of performance (Bol et al., 2005), whereas lower-performing students were more likely to overestimate it (Bol et al., 2005; Hacker et al., 2000), delivery mode (stronger correlation in face-to-face and hybrid than in online courses [Sitzmann et al., 2010]), congruence of measures (stronger correlation when the perceived and performance measures were similar [Sitzmann et al., 2010]), and emphasis of self-assessment (stronger correlation when self-report was based on level of knowledge than on gain in knowledge [Sitzmann et al., 2010]). Two additional factors that influence this relationship are practice and feedback.

## Feedback

As previously stated, Angelo and Cross (1993) emphasized the importance of providing feedback about accuracy of content knowledge when using CATs. Feedback regarding performance on summative assessments is also important for learning. Additionally, feedback regarding accuracy when individuals practiced self-assessing knowledge strengthened the relationship between perceived and performance measures of learning (Sitzmann et al., 2010).

Hacker and colleagues (2000) found that self-report accuracy improved with practice only for higher-performing students; lower-performing students did not improve with practice.

In the studies that Sitzmann and colleagues (2010) included in their meta-analysis, use of perceived learning measures as an indicator of learning varied by discipline, ranging from 17% for medical education to 79% for communication. Within the 51 psychology studies that were included, 22% used self-assessment as a measure of learning. However, of the 15 psychology studies included in this meta-analysis that examined the accuracy of self-assessment measures based on similarity between student estimates and instructor-provided scores, 40% found they were accurate, 13% found they were inaccurate, and 47% reported mixed results. Given the lack of evidence for accuracy of this type of perceived measure of learning, Sitzmann et al. (2010) recommended using graded work as indicators of learning.

Although the value and interpretation of perceived learning accuracy measures are a subject of debate (Rovai, Wighting, Baker, & Grooms, 2009), evidence suggests that their accuracy may increase with training. Moreover, perceptions of learning are associated with perceptions of other learning-related constructs. For example, a meta-analysis revealed positive correlations between self-assessments of knowledge and motivation ($\rho = .59$) and self-efficacy ($\rho = .43$, Sitzmann et al., 2010).

## Self-Efficacy

Self-efficacy refers to the extent to which individuals are capable of, or perceive themselves to be capable of, achieving designated outcomes (Bandura, 1977). In this chapter, all measures involve self-reports of perceived self-efficacy. Grounded in Bandura's (1977) cognitive social learning theory, efficacy expectancies (the perceived ability to complete a task successfully) differ from outcome expectancies (the assumption that if an action is completed, then an outcome will be achieved). For example, students' outcome expectancies may be that studying regularly will lead them to do well in a course. The corresponding efficacy expectancy would be their perception of whether they are *capable of* studying regularly.

In an academic context, a meta-analysis revealed that self-efficacy was linked to both enhanced academic performance and persistence (Multon et al., 1991). It has also been linked to self-regulated learning, goal setting, and use of learning strategies (Zimmerman, 2000). One powerful predictor of self-efficacy is performance accomplishments (prior success at a task). Other predictors include vicarious experiences (watching others succeed), verbal persuasion, and interpretation of emotional arousal (Bandura, 1977).

Although self-efficacy may be measured as a global construct (Chen, Gully, & Eden, 2001), its predictive power is greater when it is measured in a specific domain (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996). In this section, we describe instruments for assessing self-efficacy in educational contexts. First, we discuss measures that focus exclusively on academic self-efficacy (e.g., mastery of subject area content and related skills). Second, we review measures that assess perceptions of both academic and social self-efficacy, operationalized by such items as making new friends, talking to university staff, and asking questions in class. Finally, we discuss The Motivated Strategies for Learning Questionnaire (MSLQ, Garcia & Pintrich, 1995; Pintrich et al., 1991; Pintrich, Smith, Garcia, & McKeachie, 1993), a widely-used instrument that measures self-efficacy in conjunction with other motivation and learning-related constructs.

Given the connections between self-efficacy and multiple constructs related to academic success and to academic performance itself, measuring self-efficacy may serve multiple purposes in the classroom. For example, knowing students' current level of self-efficacy may help launch discussions about constructs related to self-efficacy and strategies to improve it. The classroom discussion should include coverage of positive academic behaviors such as recommended study habits (Hoigaard, Kovac, Overby, & Haugen, 2015), self-regulated learning, and goal setting. Improving self-efficacy can increase motivation in current students and may encourage students to become life-long learners. To build realistic self-efficacy beliefs, students should be trained to develop specific skills needed for effective performance (Galyon, Blondin, Yaw, Nalls, & Williams, 2012). An important caveat relates to ensuring that students' individual self-efficacy beliefs are well calibrated with their actual performance (DiBenedetto & Bembenutty, 2013).

## Measures of Academic Self-Efficacy

Among the scales targeted strictly at academic self-efficacy, Chemers, Hu, and Garcia (2001) developed an eight-item, reliable (α = .81) measure on which students rated the extent to which statements applied to them on a 7-point Likert scale ranging from 1 (*very untrue*) to 7 (*very true*). Findings indicated that self-efficacy was positively related to both academic expectations (future performance, meeting goals) and instructor evaluations of classroom performance. Students with higher self-efficacy also tended to have stronger self-rated coping skills relative to expected levels of pressure.

Elias and Loomis (2002) developed the Academic Self-Efficacy Scale (ASES), which differs from others discussed. Instead of ratings of confidence in performing specific tasks, the ASES includes ratings of confidence in successfully completing 18 specific general education and 5 specific physical education courses with at least a grade of B. Additionally, students rate their confidence in achieving 13 academic milestones, such as "earn a cumulative GPA of at least 2.0 (or 3.0) after 2 years of study," "successfully pass all courses enrolled in over the next three semesters," and "graduate." Items are rated on a 10-point Likert scale with answer options ranging from 0 (*no confidence at all*) to 9 (*complete confidence*). The three scale factors all demonstrated acceptable reliability (α = .86 to .94). Self-efficacy for both general courses and academic milestones were positively correlated with need for cognition (NFC; Cacioppo, Petty, & Kao, 1984). In addition, overall self-efficacy and NFC significantly predicted GPA, with self-efficacy serving as a mediator of the relationship between NFC and GPA. Elias and Loomis noted that the pattern of results suggested that enjoyment of academics, as reflected by NFC, enhanced perceptions of self-efficacy.

Another measure focusing on confidence, the Academic Behavioural Confidence (ABC) Scale (Sander & Sanders, 2009), includes 24 items (α = .88). Students indicate confidence in their ability to accomplish each item on a 5-point response scale with *not at all confident* and *very confident* as the anchor points. A four-factor model (grades, verbalizing, studying, and attendance) with 17 items emerged as preferable in confirmatory factor analyses. Using the ABC scale, Putwain, Sander, and Larkin (2013) investigated the relationships among academic self-efficacy, learning-related emotions, and academic success (a standardized, weighted score of assessments during the semester). Findings indicated that confidence in studying abilities at the beginning of the first semester positively predicted performance during that semester and positive learning-related emotions at the beginning of the next semester. The researchers suggested that self-efficacy as it applies to studying may relate to enhanced self-regulated learning, which, in turn, may relate to perceiving difficult tasks as challenges rather than threats.

## Measures of Academic and Social Self-Efficacy

Encompassing social as well as academic self-efficacy, Solberg, O'Brien, Villareal, Kennel, and Davis (1993) developed and validated the College Self-Efficacy Inventory (CSEI) using a Hispanic population, but it is also appropriate for use with other student populations. The inventory included a list of 20 tasks. Students indicated their confidence in completing each task using a scale ranging from 0 (*not at all confident*) to 10 (*extremely confident*). A factor analysis yielded three factors that encompassed 19 of the tasks with factor loadings greater than .50. The three factors related to courses, roommates, and other social-type tasks. Overall reliability (α = .93) and subscale reliability (α = .88 for each factor) were satisfactory. All three self-efficacy factors were negatively correlated with psychological distress as measured by the Brief Symptom Inventory (Derogatis & Melisaratos, 1983) with *r*s ranging from -.44 to -.53.

Gore (2006) reported that correlations of CSEI self-efficacy scores with GPA were notably higher at the end of the second and third semesters (*r*s = .35 and .21, respectively) than they were at the beginning of the first semester (*r*s from .00 to .13). This finding supports the importance of performance accomplishments as an influence on self-efficacy. GPA was more closely associated with course-related self-efficacy than with social self-efficacy.

To examine the effects of both academic and social self-efficacy and stress on academic outcomes, Zajachova, Lynch, and Espenshade (2005) developed a 27-item list of tasks. Students provided two ratings for each task. On the efficacy scale, they rated their confidence in successfully completing the task, using a 0 (*not confident*) to 10 (*extremely confident*) scale. On the stress scale, students rated how stressful each task was, using a 0 (*not stressful*) to 10 (*very stressful*) response scale. Analyses indicated that both scales had the same four factors: interaction at school; performance in class; performance out of class; and managing work, family, and school. Reliabilities for the four factor subscales ranged from α = .72 to .87. For each factor pair, self-efficacy negatively correlated with stress. Findings revealed that self-efficacy was a positive predictor of GPA but was unrelated to retention in the sophomore year. In contrast, stress was negatively, but not significantly, related to GPA.

### The Motivated Strategies for Learning Questionnaire (MSLQ)

Researchers from multiple institutions developed the MSLQ based on social cognitive principles.  As Garcia and Pintrich (1995) explained, the MSLQ has two major sections (motivation and learning strategies) with the motivational section subdivided into three components: expectancy (including the self-efficacy scale), value (e.g., extrinsic and intrinsic orientations), and affect (encompassing test anxiety).  Because beliefs and strategies likely differ across courses, students respond to all 81 items in terms of a specific course using a 7-point Likert scale with answer options ranging from 1 (*not at all true of me*) to 7 (*very true of me*).

The eight items on the self-efficacy scale ($\alpha = .93$, Pintrich et al., 1991) reflect expectancy for success and self-efficacy.  Regarding predictive validity, scores on this scale positively correlated with final grade ($r = .41$); the correlation was stronger for the self-efficacy scale than for any of the other MSLQ motivation or learning strategy scales.  Self-efficacy scores also correlated more strongly with intrinsic motivation ($r = .59$) than with any other scale.  Komarraju and Nadler (2013) reported a significant correlation ($r = .50$) between self-efficacy and effort regulation and found that students with higher self-efficacy were more likely than those with lower self-efficacy to believe that intelligence can change and to adopt mastery goals.

## Conclusion

A well-developed assessment plan encompasses both classroom and program-level assessments of learning based on specific goals.  For example, the measures discussed in this chapter coincide with goals in the APA Guidelines 2.0 (APA, 2013) related to the knowledge base of psychology, communication, and professional development.  Both formative and summative assessments play a distinct role in promoting and measuring student learning; although, educators may not consistently recognize distinctions between the two (Taras, 2002) or realize that the same types of assessments (e.g., rubrics, quizzes) may be used for both formative and summative purposes.

The tools discussed in this chapter to evaluate actual classroom learning (e.g., CATs and rubrics) and program assessment (e.g., standardized examinations) may not be classified as traditional scales.  However, considering the variability in learning goals for both content and skills across courses, it would be challenging to develop scales of actual learning suitable for universal application at the course level.  Moreover, multiple possibilities do exist for meaningful classroom assessment (e.g., examinations, presentations, problem-solving assignments), which, if constructed and evaluated appropriately, may serve as effective measures of learning.  For program evaluation, the use of such standardized assessments can provide a common metric of performance.  Given these existing assessments for evaluating course and program-specific knowledge and skills, it may not be necessary to create generalizable, traditional scales for these purposes.

On the other hand, multiple scales are available to measure students' perceptions of learning and constructs related to learning.  As Rovai and colleagues (2009) noted, educational outcomes are influenced by numerous variables, including factors related to course design and pedagogy as well as students' characteristics and beliefs. Perceived self-efficacy is one such belief discussed in this chapter; it contributes to academic success via boosting engagement, commitment, and persistence (Bandura, 1993). However, as mentioned in Chapter 14, *SoTL Scales: The Case of the Missing Links* (Richmond, 2015), to meaningfully interpret self-reported measures of perceived skill, they should be compared to objective performance measures. Instructors' awareness of the association between perceived self-efficacy and academic performance and of factors contributing to realistic self-efficacy beliefs may contribute to the ultimate goal of student learning.

# References

References marked with an asterisk indicate a scale.

American Psychological Association. (2013). *APA guidelines for the undergraduate psychology major: Version 2.0*. Retrieved from http://www.apa.org/ed/precollege/undergrad/index.aspx

Anderson, D., & Burns, S. (2013). One-minute paper: Student perception of learning gains. *College Student Journal, 47*, 219–227. Retrieved from http://www.projectinnovation.com/college-student-journal.html

Angelo, T., & Cross, K. P. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers*. San Francisco, CA: Jossey-Bass.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215. doi:10.1037/0033-295X.84.2.191

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist*, *28*, 117–148. doi:10.1207/s15326985ep2802_3

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development, 67*, 1206–1222, doi:10.1111/j.1467-8624.1996.tb01791.x

Barron, A. B., Hebets, E. A., Cleland, T. A., Fitzpatrick, C. L., Hauber, M. E., & Stevens, J. R. (2015). Embracing multiple definitions of learning. *Trends in Neurosciences*, *38*, 405–407. doi:10.1016/j.tins.2015.04.008

Bol, L., Hacker, D. J., O'shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education*, *73*, 269–290. doi:10.3200/JEXE.73.4.269-290

Brennan, R. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34. doi:10.1111/j.1745-3992.1992.tb00260.x

*Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of Need for Cognition. *Journal of Personality Assessment*, *48*, 306–307. doi:10.1207/s15327752jpa4803_13

Centra, J. A., & Gaubatz, N. B. (2005). Student perceptions of learning and instructional effectiveness in college courses. Educational Testing Service. Retrieved from http://www.ets.org/Media/Products/perceptions.pdf

Chemers, M., Hu, L., & Garcia, B. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology, 93*, 55–64. doi:10.1037/0022-0663.93.1.55

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods, 4*, 62–83. doi:10.1177/109442810141004

Chesebro, J., L., & McCroskey, J. C. (2000). The relationship between students' reports of learning and their actual recall of lecture material: A validity test. *Communication Education, 49,* 297–301. doi:10.1080/03634520009379217

Chiou, C.-C., Wang, Y.-M., & Lee, L.-T. (2014). Reducing statistics anxiety and enhancing statistics learning achievement: Effectiveness of a one-minute strategy. *Psychological Reports: Sociocultural Issues in Psychology, 115,* 297–310. doi:10.2466/11.04.PRO.115c12z3

Cottell, P., & Harwood, E. (1998). Do classroom assessment techniques (CATs) improve student learning? *New Directions for Teaching and Learning, 75*, 37–46. doi:10.1002/tl.7504

Crick, J. E., & Brennan, R. L. (1984). *General purpose analysis of variance system* 2.2. Iowa City, IA. American College Testing Program.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13,* 595–605. doi:10.1017/S0033291700048017

DiBenedetto, M. K., & Bembenutty, H. (2013). Within the pipeline: Self-regulated learning, self-efficacy, and socialization among college students in science courses. *Learning and Individual Differences*, *23*, 218–224. doi:10.1016/j.lindif.2012.09.015

Dolinsky, B., & Kelley, J. M. (2010). For better or for worse: Using an objective program assessment measure to enhance an undergraduate psychology program. *Teaching of Psychology, 37*, 252–256. doi:10.1080/00986283.2010.510978

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research, & Evaluation, 14,* 1–11. Retrieved from http://pareonline.net/getvn.asp?v=14&n=7

*Elias, S. M., & Loomis, R. J. (2002). Utilizing need for cognition and perceived self-efficacy to predict academic performance. *Journal of Applied Social Psychology*, *32*, 1687–1702. doi:10.1111/j.1559-1816.2002.tb02770.x

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*, 395–430. doi:10.3102/00346543059004395

Frymier, A. B., & Houser, M. L. (1999). The revised learning indicators scale. *Communication Studies*, *50*(1), 1–12. doi:10.1080/10510979909388466

Frymier, A. B., Shulman, G. M., & Houser, M. (1996). The development of a learner empowerment measure. *Communication Education, 45,* 181–199. doi:10.1080/03634529609379048

Gallagher, S. P., & Cook, S. P. (2013). The validity of the Major Field Test in psychology as a programme assessment tool. *Psychology Teaching Review, 19*(2), 59–72. Retrieved from http://www.millersville.edu/~sgallagh/Publications/ptr2013.pdf

Galyon, C. E., Blondin, C. A., Yaw, J. S., Nalls, M. L., & Williams, R. L. (2012). The relationship of academic self-efficacy to class participation and exam performance. *Social Psychology of Education*, *15*, 233–249. doi:10.1007/s11218-011-9175-x

Garcia, T., & Pintrich, P. R. (1995). *Assessing students' motivation and learning strategies: The Motivated Strategies for Learning Questionnaire*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Retrieved from ERIC database. (ED383770)

Gore, P. J. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment*, *14*, 92–115. Retrieved from http://jca.sagepub.com

*Greenberg, K. P. (2015). Rubric use in formative assessment: A detailed behavioral rubric helps students improve their scientific writing skills. *Teaching of Psychology*, *42*, 211–217. doi:10.1177/009868315587618

Hacker, D. J., Bol, L., & Bahbabani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition Learning*, *3*, 101–121. doi:10.1007/s11409-008-9021-5

Høigaard, R., Kovač, V. B., Øverby, N. C., & Haugen, T. (2015). Academic self-efficacy mediates the effects of school psychological climate on academic achievement. *School Psychology Quarterly*, *30*, 64–74. doi:10.1037/spq0000056

Komarraju, M., & Nadler, D. (2013). Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter? *Learning and Individual Differences, 25*, 67–72. doi:10.1016/j.lindif.2013.01.005

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134. doi:10.1037/0022-3514.77.6.1121

*Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students’ use of exemplars and detailed rubrics as formative assessment. *Instructional Science*, *42*, 539–559. doi:10.1007/s11251-013-9299-9

Lom, B. (2012). Classroom activities: Simple strategies to incorporate student-centered activities within undergraduate science lectures. *The Journal of Undergraduate Neuroscience Education, 11,* A64–A71. Retrieved from http://www.funjournal.org/

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, *15*, 249–268. doi:10.1207/S15324818AME1503.2

McCroskey, J. C., Sallinen, A., Fayer, J. M., Richmond, V. P., & Barraclough, R. A. (1996). Nonverbal immediacy and cognitive learning: A cross-cultural investigation. *Communication Education, 45,* 200–211. doi:10.1080/03634529609379049

Montgomery, K. (2002). Authentic tasks and rubrics: Going beyond traditional assessments in college teaching. *College Teaching*, *50*(1), 34–39. doi:10.1080/87567550209595870

Mueller, J. (2014) Rubrics. Available at http://jfmueller.faculty.noctrl.edu/toolbox/rubrics.htm

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, *38*, 30–38. doi:10.1037/0022-0167.38.1.30

National Association of Colleges and Employers (2015). Job outlook*:* The candidate skills/qualities employers want, the influence of attributes. Retrieved from https://www.naceweb.org/s11122014/job-outlook-skills-qualities-employers-want.aspx

Pinter, B., Matchock, R. L., Charles, E. P., & Balch, W. R. (2014). A cross-sectional evaluation of student achievement using standardized and performance-based tests. *Teaching of Psychology, 41,* 20–17. doi:10.1177/0098628313514174

*Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33–40. doi:10.1037/0022-0663.82.1.33

*Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Technical report 91-B-004. Retrieved from ERIC database. (ED338122)

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire. *Educational and Psychological Measurement*, *53*, 801–813, doi:10.1177/0013164493053003024.

Putwain, D., Sander, P., & Larkin, D. (2013). Academic self-efficacy in study-related skills and behaviours: Relations with learning-related emotions and academic success. *British Journal of Educational Psychology*, *83*, 633–650. doi:10.1111/j.2044-8279.2012.02084.x

Reddy, M. Y., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*, 435–448. doi:10.1080/02602930902862859

*Reynolds, J., Smith, R., Moskovitz, C., & Sayle, A. (2009). BioTAP: A systematic approach to teaching scientific writing and evaluating undergraduate theses. *BioScience*, *59*, 896–903. doi:10.1025/bio.2009.59.10.11

Richmond, V. P., Gorham, J. S., & McCroskey, J. C. (1987). The relationship between selected immediacy behaviors and cognitive learning. In M. McLaughlin (Ed.), *Communication Yearbook 10*, (pp. 574-590). Beverly Hills, CA: Sage.

Rovai, A. P., Wighting, M. J., Baker, J. D., & Grooms, L. D. (2009). Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. *Internet and Higher Education*, *12*, 7–13. doi:10.1016/j.iheduc.2008.10.002

*Sander, P., & Sanders, L. (2009). Measuring academic behavioural confidence: The ABC Scale revisited. *Studies in Higher Education*, *34*, 19–35. doi:10.1080/03075070802457058

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33–45. doi:10.1007/s11409-008-9031-3

*Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*, 460–475. doi:10.1006/ceps.1994.1033

Short, F., & Martin, J. (2012). Who wants to be a psychology graduate? Impact of formative multiple-choice review questions on summative assessment performance. *Psychology Learning & Teaching, 11*, 218–227. doi:10.2304/plat.2012.11.2.218

Simpson-Beck, V. (2011). Assessing classroom assessment techniques. *Active Learning in Higher Education, 12*(2), 125–132. doi:10.1177/1469787411402482

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, *9*(2), 169–191. doi:10.5465/AMLE.2010.51428542

Soetaert, E. (1998). Quality in the classroom: Classroom assessment techniques as TQM. *New Directions for Teaching and Learning, 75*, 47–55. doi:10.1002/tl.7505

*Solberg, V. S., O'Brien, K., Villareal, P., Kennel, R., & Davis, B. (1993). Self-efficacy and Hispanic college students: Validation of the College Self-Efficacy Instrument. *Hispanic Journal of Behavioral Sciences*, 15(1), 80–95. doi:10.1177/07399863930151004

Stead, D. R. (2005). A review of the one-minute paper. *Active Learning in Higher Education, 6,* 118–131. doi:10.1177/1469787405054237

*Stellmack, M. A., Kohneim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, A. P. (2009). An assessment of reliability and validity of a rubric for grading APA-style introductions. *Teaching of Psychology*, *36*, 102–107. doi:10.1080/00986280902739776

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation*, *9*(4). Retrieved from http://pareonline.net

Stoloff, M. L., & Feeney, K. J. (2002). The Major Field Test as an assessment tool for an undergraduate psychology program. *Teaching of Psychology, 29*, 92–98. doi:10.1207/515328023TOP2902_01

Stoloff, M. L., Good, M. R., Smith, K. L., & Brewster, J. (2015). Characteristics of programs that maximize psychology major success. *Teaching of Psychology*, *42*, 99–108. doi:10.1177/0098628315569877

*Timmerman, B. E. C., Strickland, D. C., Johnson, R. L., & Payne, J. R. (2011). Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing. *Assessment and Evaluation in Higher Education*, *36,* 509–547. doi:10.1080.02602930903540991

Taras, M. (2008). Summative and formative assessment: Perceptions and realities. *Active Learning in Higher Education, 9,* 172–192. doi:10.1177/1469787408091655

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22,* 537–548. doi:10.1080/0141192960220502

*Zajacova, A., Lynch, S. M., & Espenshade, T. J. (2005). Self-efficacy, stress, and academic success in college. *Research in Higher Education*, *6*, 677–706. doi:10.1007/s11162-004-4139-z

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*, 82–91. doi:10.1006/ceps.1999.1016

*Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, *80*, 284–290. doi:10.1037/0022-0663.80.3.284

# Chapter 7: Measuring Critical Thinking Skills

R. Eric Landrum[1] and Maureen A. McCarthy[2]

[1]Boise State University, [2]Kennesaw State University

Do critical thinking skills exist -- and can they be measured? Clearly articulating the construct of critical thinking is central to measurement yet articulating a clear definition of critical thinking remains elusive. In fact, Halpern (1999) acknowledged the difficulty of defining the construct and she offered a wide range of possible interrelated definitions. She also reflected on similarities of the construct across disciplines including: problem solving, decision making, and cognitive processes (also see Halpern, 1996 for a comprehensive review of the construct of critical thinking). Despite the complexity of defining the construct, we believe that it is both possible and important to measure critical thinking, particularly during this era of increased demand for accountability.

Critical thinking remains one of the most important skills identified as an outcome of a college degree. Not only are critical thinking skills desired in college graduates, but this skill set is beneficial to an educated citizenry. In addition to students, many constituencies have a keen interest in college graduates demonstrating critical thinking skills, including educators (Appleby, 2009; Keeling & Hersh, 2012; Yanchar, Slife, & Warne, 2008), higher education associations (American Association of Colleges & Universities [AAC&U], 2006; 2010), employers (AAC&U, 2008), and the general public (AAC&U, 2005; Baum & Ma, 2007). More recently, the American Psychological Association (APA) reaffirmed the importance of critical thinking skills in the revision of discipline specific guidelines for the undergraduate major (APA, 2013).

More generally, this emphasis on critical thinking as an important outcome of a college degree was emphasized with the publication of *Academically Adrift* by Arum and Roksa (2011a). In their research using the Collegiate Learning Assessment (CLA), they found that a large percentage of students in both two-year and four-year institutions did not demonstrate progress in critical thinking skills at the end of their academic studies. Although the efforts of Arum and Roksa (2011b) have limitations with regard to methodology and the motivation of CLA test-takers, the value of the process is clear; meaningful assessment can provide invaluable feedback to educators, administrators, and to the higher education community.

## Broad Perspectives About Critical Thinking

Scholars have written extensively about critical thinking (Halpern, 1996; Halpern, 2010) as an important skill; however, a comprehensive review and analysis of the construct exceed the scope of this chapter. Some have suggested that critical thinking is developed as a discipline specific skill (Davies, 2013; McGovern, Furumoto, Halpern, Kimble, McKeachie, 1991), whereas others have suggested that critical thinking is developed broadly across many courses. Critical thinking can be described as the act of processing, evaluating, and creating new information rather than merely recalling information (Butler, 2012; Halpern, 2010). In fact, Dunn and Smith (2008) made the argument that writing is a form of critical thinking (see also Preiss, Castillo, Flotts, & San Martin, 2013) and Halpern (1987) suggested that the generation and interpretation of analogies is an activity that clearly demonstrates critical thinking. See Table 1 for additional definitions of critical thinking definitions.

Table 1

*Examples of Critical Thinking Definitions*

"The conscious process a person does when he or she explores a situation or a problem from different perspectives" (French, Hand, Nam, Yen, & Vazquez, 2014, p. 275).

"Challenging a claim or an opinion (either one's own or another person's) with the purpose of finding out what to believe or do" (O'Hare & McGuinness, 2009, p. 123).

"Reasonable and reflective thinking that is focused on deciding what to believe to do" (Norris & Ennis, 1989, p. 1).

"The use of those cognitive skills or strategies that increase the probability of a desirable outcome. It is used to describe thinking that is purposeful, reasoned, and goal-directed—the kind of thinking involved in solving problems, formulating inferences, calculating likelihoods, and making decisions, when the thinker is using skills that are thoughtful and effective for the particular context and type of thinking task" (Halpern, 2003, p. 6, as cited in Butler, 2012).

A second term—psychological literacy—has also been used interchangeably with historical origins dating to the St. Mary's conference in 1991 (McGovern, Furumoto, Halpern, Kimble, McKeachie, 1991). With the re-emergence of psychological literacy (McGovern et al., 2010) emphasized as an important outcome for the major, critical thinking continues to be a central topic in the discussions of psychology educators. In fact, many components of critical thinking are contained in the definition of psychological literacy:

> (a) having a well-defined vocabulary and basic knowledge of the critical subject matter of psychology; (b) valuing the intellectual challenges required to use scientific thinking and the disciplined analysis of information to evaluate alternative courses of action; (c) taking a creative and amiable skeptic approach to problem solving; (d) applying psychological principles to personal, social, and organizational issues in work, relationships, and the broader community; (e) acting ethically; (f) being competent in using and evaluating information and technology; (g) communicating effectively in different modes and with many different audiences; (h) recognizing, understanding, and fostering respect for diversity; and (i) being insightful and reflective about one's own and others' behavior and mental processes. (McGovern, et al., 2010, p. 11)

This conceptualization dovetails nicely with recent national efforts devoted to validating undergraduate education in psychology as a liberal arts degree that affords students opportunities to think critically across multiple career opportunities.

The American Psychological Association revised the *APA Guidelines for the Undergraduate Psychology Major* in 2013, referred to as Guidelines 2.0, which continued to emphasize complex thinking as an important outcome of the major in psychology. In fact, Goal 2 of the Guidelines includes five specific student outcomes:

- use scientific reasoning to interpret psychological phenomena
- demonstrate psychology information literacy
- engage in innovative and integrative thinking and problem solving
- interpret, design, and conduct basic psychological research
- incorporate sociocultural factors in scientific inquiry

If we compare these outcomes to the definitions of critical thinking above, it seems apparent that there is overlap between the definitions of psychological literacy and critical thinking.

## Measures of Critical Thinking

Our review of critical thinking measures is twofold: First, we want to make sure that the mainstream measures of critical thinking are reviewed (albeit briefly) in this chapter. We review critical thinking measures that are specific to psychology as well as broad-based general measures. Second, our review is not to be interpreted as comprehensive. Instead we want to share information about the most common measures of critical thinking. If the reader desires additional details about the measures, we have included an appendix with references for additional information.

### General Measures

For each of the general measures, we provide "quick snippets" about how the measure has been used in published research; this is meant to provide a sampling of the current efforts and is not meant to be comprehensive. For example, the Watson-Glaser Critical Thinking Appraisal test is often cited as one of the most frequently used general measures of critical thinking. More recently Burke, Sears, Kraus, and Roberts-Cady (2014) used the *Watson-Glaser Critical Thinking Appraisal* (WGCTA; Watson & Glaser, 1980) in a between-groups comparison of critical thinking scores across different disciplines. They found that students in a philosophy course improved their critical thinking when measured by the WGCTA. However, this same improvement was not found in the psychology course specifically designed to improve critical thinking skills. These findings may be a reflection of differences in courses, or quite possibly the difficulty in generally measuring the construct of critical thinking.

Macpherson and Owen (2010) also used the WGCTA in a test-retest study to examine development of critical thinking between two cohorts. They experienced difficulty in using the test to detect differences in critical thinking that were not already explained with the subtests of the WGCTA. These findings may reflect the complicated nature of the construct. Further, when Magno (2010) examined the role of metacognition in critical thinking, he used a structural equation model to link metacognition to the WGCTA. The construct is further complicated by findings from Clifford, Boufal, and Kurtz (2004). Using the WGCTA, they found that critical thinking skills were related to personality characteristics, in particular to openness to experience. Thus the construct of critical thinking, and the general measures of critical thinking, make it difficult to accurately measure the important skill.

Similar difficulties in accurately measuring critical thinking are present across other measures. For example, the *Cornell Critical Thinking Test* (CCTT) has been used in a variety of research studies. Recently, Stark (2012) compared the CCTT to a psychology specific test of critical thinking and found increases in the psychology specific test, but that these increases were not reflected in the more general measure using the CCCT. O'Hare and McGuiness (2009) administered a subset of tests from the *California Critical Thinking Skills Test* (CCTST) (Facione, Facione, Blohm, Howard, & Giancarlo, 1998) and *Raven's Advanced Progressive Matrices* (Set 1; Raven, 1965) to psychology undergraduates at Queen's University in Belfast. Using these measures, they found that reasoning skills improved as students progressed from the first to third year in college. The CCTST was also utilized by Feroand colleagues (2010) in a comparison of a small number of nursing students' critical thinking levels to performance on simulated clinical situations in nursing. However, they did not find a correlation between critical skills-based performance and performance on the CCTST. For a more overarching perspective about the challenges facing researchers using the WGCTA and the CCTST, see Schraw and Gutierrez (2012).

The *Halpern Critical Thinking Assessment* (HCTA; Halpern, 2010) is unique in that it relies both on recognition memory (such as completing multiple choice items) as well as recall memory (providing answers to short essays). Another important contribution that researchers have made with the HCTA is that these critical thinking scores have been compared with real-world outcomes, such as a significant negative correlation between HCTA scores and negative life events (Butler, 2012; Butler et al., 2012).

The *Ennis-Weir Test of Critical Thinking* (EWTCT; Ennis & Weir, 1985) is a free-response instrument which requires a written argument in response to a stimulus. The EWTCT was used by Szabo and Schwartz (2011) to examine potential pre-semester to post-semester growth in critical thinking scores using online discussion tools in a face-to-face course. Using this instrument, they concluded that the online supplemental instruction improved the critical reasoning of the pre-service teachers participating in the study.

Pascarella and colleagues (2014) assessed critical thinking in college students using the *Critical Thinking Test* (CTT; American College Testing Program, 1990) to examine how diversity experiences may affect critical thinking at the conclusion of the college experience. They conclude that exposure to diversity increases critical thinking in the students who participated in the study.

### Psychology-Specific Measures

The *Psychological Critical Thinking Exam* (PCTE) developed by Lawson (1999) was utilized by McLean and Miller (2010) as a between groups measure to demonstrate critical thinking differences between courses. This measure also proved useful for Haw (2011) when he administered the PCTE to students in their second and fourth years of instruction to compare advances in critical thinking. Using the PCTE, he concluded that psychology-specific critical thinking skills do improve with additional instruction. Lawson, Jordan-Fleming, and Bodle (2015) recently published an update to the PCTE. Similarly, Muehlenkamp, Weiss, and Hansen (2015) tested the efficacy of problem-based learning instructional techniques, and used scores on the PCTE as pre- and post-outcome measures, demonstrating that students in the problem-based learning condition exhibited higher critical thinking scores at the end of the semester.

A second psychology specific critical thinking test has also been used in a number of studies. *The Critical Thinking in Psychology Test*, developed by Bensley and Baxter (2006), contains an argument analysis test, a methodological reasoning test, and a causal reasoning test; however, this test is unpublished and is not widely available. It has, however, been used in multiple research contexts, such as an instrument used to measure gains after specific critical thinking instruction (Bensley, Crowe, Bernhardt, Buckner, & Allman, 2010) in specific research studies.

## Teaching Critical Thinking Skills

Despite the difficulties with defining the construct and measuring critical thinking, researchers continue to recommend teaching these skills. More specifically, several researchers (Frantz & McCarthy, in press; Lilienfeld, Lohr, & Olatunji, 2008; Wesp & Montgomery, 1998) have recommended that psychology courses offer opportunities for helping students develop these skills by questioning common myths about psychology. For example, Wesp and Montgomery (1998) were able to demonstrate an increase in critical thinking after taking a course designed to decrease beliefs about paranormal activities. Similarly, Lilienfeld and colleagues (2008) designed a course to help students to think critically about psychotherapy effectiveness; in other words, whether the treatment helps more than doing nothing or whether the outcome is due to the placebo effect. They were able to demonstrate improvement in the critical thinking skills of the students enrolled in the course.

In addition to research studies supporting the use of teaching critical thinking as a primary objective of psychology courses, two key texts to aid in designing courses include *The Critical Thinking Companion for Introductory Psychology* (Halonen, 1995) and *Thinking Critically about Critical Thinking* (Halpern, 1996). Both are filled with ideas for hands-on exercises for engaging students in tasks which may help to support the development of critical thinking skills. More importantly, with the availability of these developing measures, psychology educators do not need to guess about the effectiveness of these exercises. Utilizing the techniques available from the scholarship of teaching and learning (SoTL) literature, scholars can measure and document the effectiveness of planned interventions to enhance critical thinking.

## Recommendations

Despite the importance of teaching critical thinking and the attempts to measure this construct, the construct remains difficult to measure efficiently. Ku (2009) identified several key points to consider, including whether an objective multiple-choice format can be used to accurately measure critical thinking. Ku also indicated that it is difficult to measure higher levels of complex reasoning using a multiple-choice format. Although multiple choice testing is certainly an efficient method of measurement, it may be difficult to convince researchers that a multiple-choice format provides an accurate and complete measure of critical thinking.

How do we balance the need for efficient measurement against the complexity of the construct? One solution is to adapt Halpern's (2013) general recommendations for measuring student learning in general. Specifically, measuring student learning should include the following elements (adapted for critical thinking):

1. Multiple, varied measures for critical thinking are necessary because no single measure can capture its complexity.
2. Faculty involvement in all aspects of the measurement of critical thinking and the utilization of critical thinking outcomes is essential for success.
3. Departments should be rewarded for conducting meaningful assessments of critical thinking skills, even when the outcomes of that assessment demonstrate room for improvement.
4. Faculty members and institutions should use the outcomes of critical thinking assessments to improve their teaching and their students' learning, whether that involves curriculum changes, individual faculty changing pedagogical approaches if needed, and so on.
5. Departments should take a value-added approach to the measurement of critical thinking scores over time; that is, strive to understand the critical thinking growth within each student rather than a comparison of different groups of students. Using this approach, all students can demonstrate enhanced critical thinking skills over time.
6. Seek to utilize multiple sources of information about critical thinking from differing perspectives; by identifying overlapping efforts, a convergence of efforts through purposeful coordination may lead to richer sources of data as well as more complete and representative outcomes.

Although an educator might have some indication about the critical thinking skills that are developed during a course, a more thorough understanding is needed. For instance, there are pre-course to post-course studies where researchers examined whether critical thinking changed measurably over the semester, with mixed results (e.g., Stark, 2012). However, we believe that more research is needed regarding critical thinking skills at commencement, and how those skills relate to success after the bachelor's degree. In fact, using the *Halpern Critical Thinking Assessment* (Butler 2012; Butler, et al., 2012), researchers have reported promising outcomes relating critical thinking measures to real-world outcomes.

Perhaps the most integrative measures of critical thinking are reported in the assessment plan of James Madison University (Apple, Serdikoff, Reis-Bergan, & Barron, 2008). Multiple assessments of critical thinking occur not only across courses but also at the conclusion of the psychology major's undergraduate career. Psychology departments should employ the available critical thinking measures more often, and coordinated research efforts on a national scope are needed to maximize the utility of such measures in institution-specific domains. The model provided by Apple and colleagues (2008) is a very good starting point for many departments to consider.

A fully implemented multi-modal multi-method approach includes embedded assessment, nationally standardized tests, cross-sectional and longitudinal studies, and the creation of a national database of test results that may be useful for program review purposes as well as the identification of best practices.

> Without information about learning, there is less learning. Faculty cultures and incentive regimes that systematically devalue teaching in favor of research are allowed to persist because there is no basis for fixing them and no irrefutable evidence of how much students are being shortchanged. (Carey, 2010, p. A72)

In our opinion, an important component of assessment is using the information to inform and revise educational practice. The ultimate goal of testing is the prediction of non-test behavior. The ultimate goal of an undergraduate education in psychology is to impact behaviors, attitudes, and opinions of our students following graduation so that they can create real change in the world, whether that be through their own behavior or through the influence of others. The ability to think critically is a key skill in reaching these goals.

# References

References marked with an asterisk indicate a scale.

American College Testing Program. (1990). *Report on the technical characteristics of CAAP: Pilot year 1: 1988-89*. Iowa City, IA: Author.

American Psychological Association. (2013). *APA guidelines for the undergraduate psychology major: Version 2.0*. Retrieved from http://www.apa.org/precollege/undergrad/index.aspx

Apple, K. J., Serdikoff, S. L., Reis-Bergan, M. J., & Barron, K. E. (2008). Programmatic assessment of critical thinking. In D. S. Dunn, J. S. Halonen, & R. A. Smith (Eds.), *Teaching critical thinking in psychology* (pp. 77-88). Malden, MA: Blackwell Publishing.

Appleby, D. C. (2009). *Essential skills for psychology majors: Those we say they can develop and those employers say they value in job applicants*. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada.

Arum, R., & Roksa, J. (2011a, January 28). Are undergraduates actually learning anything? *Chronicle of Higher Education, 57*(21), A30-A31.

Arum, R., & Roksa, J. (2011b). *Academically adrift: Limited learning on college campuses*. Chicago, IL: University of Chicago Press.

Association of American Colleges and Universities. (2005). *Liberal education outcomes: A preliminary report on student achievement in college*. Washington, DC: Author.

Association of American Colleges and Universities. (2006). *How should colleges prepare students to succeed in today's global economy?* Washington, DC: Hart Research Associates.

Association of American Colleges and Universities. (2008). *How should colleges assess and improve student learning?* Washington, DC: Hart Research Associates.

Association of American Colleges and Universities. (2010). *Raising the bar: Employers' views on college learning in the wake of the economic downturn*. Washington, DC: Hart Research Associates.

Baum, S., & Ma, J. (2007). *Education pays: The benefits of higher education for individuals and society*. Washington, DC: College Board.

Bensley, D. A., & Baxter, C. (2006). *The Critical Thinking in Psychology Test*. Unpublished manuscript, Frostburg State University, Frostburg, MD.

Bensley, D. A., Crowe, D. S., Bernhardt, P., Buckner, C., & Allman, A. L. (2010). Teaching and assessing critical thinking skills for argument analysis in psychology. *Teaching of Psychology, 37*, 91-96. doi:10.1080/0098628100326656

Burke, B. L., Sears, S. R., Kraus, S., & Roberts-Cady, S. (2014). Critical analysis: A comparison of critical thinking changes in psychology and philosophy classes. *Teaching of Psychology, 41*, 28-36. doi:10.1177/0098628313514175

Butler, H. A. (2012). Halpern Critical Thinking Assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology, 26*, 721-729. doi:10.1002/acp.2851

Butler, H. A., Dwyer, C. P., Hogan, M. J., Franco, A., Rivas, S. F., Saiz, C., & Almeida, L. S. (2012). The Halpern Critical Thinking Assessment and real-world outcomes: Cross-national applications. *Thinking Skills and Creativity, 7*, 112-121. doi:10.1016/j.tsc.2012.04.001

Carey, K. (2010, December 17). Student learning: Measure or perish. *Chronicle of Higher Education, 57*(17), A72.

Clifford, J. S., Boufal, M. M., & Kurtz, J. E. (2004). Personality traits and critical thinking skills in college students. *Assessment, 11*, 169-176. doi:10.1177/1073191104263250

Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development, 32*, 529-544. doi:10.1080/07294360.2012.697878

Dunn, D. S., & Smith, R. A. (2008). Writing as critical thinking. In D. S. Dunn, J. S. Halonen, & R. A. Smith (Eds.), *Teaching critical thinking in psychology* (pp. 163-173). Malden, MA: Blackwell Publishing.

*Ennis, R. H., & Millman, J. (2005a). *Cornell Critical Thinking Test, Level X*. Pacific Grove, CA: Midwest Publications.

*Ennis, R. H., & Millman, J. (2005b). *Cornell Critical Thinking Test, Level Z*. Pacific Grove, CA: Midwest Publications.

*Ennis, R.H., & Weir, E. (1985). *The Ennis-Weir Critical Thinking Essay Test*. Pacific Grove, CA: Midwest Publications.

*Facione, P. A., Facione, R. N., Blohm, S. W., Howard, K., & Giancarlo, C. A. F. (1998). *California Critical Thinking Skills Test: Manual* (revised). Millbrae, CA: California Academic Press.

Fero, L. J., O'Donnell, J. M., Zullo, T. G., Dabbs, A. D., Kitutu, J., Samosky, J. T., & Hoffman, L. A. (2010). Critical thinking in nursing students: Comparison of simulation-based performance with metrics. *Journal of Advanced Nursing, 66*, 2182-2193. doi:10.1111/j.1365-2648.2010.05385.x

Frantz, S., & McCarthy, M. A. (in press). Challenging the status quo: Evidence that introductory psychology can dispel myths. *Teaching of Psychology*.

French, B. F., Hand, B., Nam, J., Yen, H.-J., & Vazquez, V. (2014). Detection of differential item functioning in the Cornell Critical Thinking Test across Korean and North American students. *Psychological Test and Assessment Modeling, 56*, 275-286.

Halonen, J. (1995). *The critical thinking companion for introductory psychology*. New York, NY: Worth.

Halonen, J. S. (2008). Measure for measure: The challenge of assessing critical thinking. In D. S. Dunn, J. S. Halonen, & R. A. Smith (Eds.), *Teaching critical thinking in psychology* (pp. 61-75). Malden, MA: Blackwell Publishing.

Halpern, D. F. (1987). Analogies as a critical thinking skill. In D. E. Berger, K. Pezdek, & W. P. Banks (Eds.), *Applications of cognitive psychology: Problem solving, education, and computing* (pp. 305-313). Hillsdale, NJ: Erlbaum.

Halpern, D. F. (1996). *Thinking critically about critical thinking*. Mahwah, NJ: Erlbaum.

Halpern, D. F. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New Directions for Teaching and Learning, 80,* 69-74. San Francisco, CA: Jossey-Bass.

Halpern, D. F. (2003a). Thinking critically about creative thinking. In M. A. Runco (Ed). *Critical creative processes. Perspectives on creativity research* (pp. 189-207). Cresskill, NJ: Hampton Press.

Halpern, D. F. (2003b). *Thought and knowledge: An introduction to critical thinking* (4th ed.). Mahwah, NJ: Erlbaum.

Halpern, D. F. (2013). A is for assessment: The other scarlet letter. *Teaching of Psychology, 40*, 358-362. doi:10.1177/0098628313501050

*Halpern, D. F. (2010). *Halpern Critical Thinking Assessment*. Vienna Test System. Modling, Austria: Schuhfried GmbH. Retrieved from http://www.schuhfried.at/

Haw, J. (2011). Improving psychological critical thinking in Australian university students. *Australian Journal of Psychology, 63*, 150-153. doi:10.1111/j.1742-9536.2011.00018.x

Keeling, R. P., & Hersh, R. H. (2012, July/August). Where's the learning in higher learning? *Trusteeship, 20*(4), 16-21.

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*, 70-76. doi:10.1016/j.tsc.2009.02.001

*Lawson, T. J. (1999). Assessing psychological critical thinking as a learning outcome for psychology majors. *Teaching of Psychology, 26*, 207-209. doi:10.1207/S15328023TOP260311

*Lawson, T. J., Jordan-Fleming, M. K., & Bodle, J. H. (2015). Measuring psychological critical thinking: An update. *Teaching of Psychology, 42*, 248-253. doi:10.1177/0098628315587624

Lilienfeld, S. O., Lohr, J. M., & Olatunji, B. O. (2008). Encouraging students to think critically about psychotherapy: Overcoming naïve realism. In D. S. Dunn, J. S. Halonen, & R. A. Smith (Eds.), *Teaching critical thinking in psychology* (pp. 267-271). Malden, MA: Blackwell Publishing.

Macpherson, K., & Owen, C. (2010). Assessment of critical thinking ability in medical students. *Assessment & Evaluation in Higher Education, 35*, 45-58. doi:10.1080/02602930802475471

Magno, C. (2010). The role of metacognitive skills in developing critical thinking. *Metacognition Learning, 5*, 137-156. doi:10.1007/s11409-010-9054-4

McGovern, T. V., Furumoto, L., Halpern, D. F., Kimble, G. A., & McKeachie, W. J. (1991). Liberal education, study in depth, and the arts and sciences major – Psychology. *American Psychologist, 46*, 598-605.

McGovern, T. V., Corey, L., Cranney, J., Dixon, W. E., Jr., Holmes, J.D., Kuebli, J. E., … & Walker, S. J. (2010). Psychologically literate citizens. In D. F. Halpern (Ed.), *Undergraduate education in psychology: A blueprint for the future of the discipline* (pp. 9-28). Washington, DC: American Psychological Association.

McLean, C. P., & Miller, N. A. (2010). Changes in critical thinking skills following a course on science and pseudoscience: A quasi-experimental study. *Teaching of Psychology, 37*, 85-90. doi:10.1080/0098628100462714

Muehlenkamp, J. J., Weiss, N., & Hansen, M. (2015). Problem-based learning for introductory psychology: Preliminary supporting evidence. *Scholarship of Teaching and Learning in Psychology*. Advance online publication. http://dx.doi.org/10.1037/stl0000027

Norris, S. P., & Ennis, R. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Midwest Publications.

O'Hare, L., & McGuinness, C. (2009). Measuring critical thinking, intelligence, and academic performance in psychology undergraduates. *The Irish Journal of Psychology, 30*, 123-131. doi: 10.1080/03033910.2009.10446304

Pascarella, E. T., Martin, G. L., Hanson, J. M., Trolian, T. L., Gillig, B., & Blaich, C. (2014). Effects of diversity experiences on critical thinking skills over 4 years of college. *Journal of College Student Development, 55*, 86-92. doi:10.1353/csd.2014.0009

Preiss, D. D., Castillo, J. C., Flotts, P., & San Martin, E. (2013). Assessment of argumentative writing and critical thinking in higher education: Educational correlates and gender differences. *Learning and Individual Differences, 28*, 193-203. doi:10.1016/j.lindif.2013.06.004

*Raven, J. (1965). *Progressive matrices*. London, England: H. K. Lewis.

Schraw, G., & Gutierrez, A. (2012). Assessment of thinking skills. In M. F. Shaughnessy (Ed.), *Critical thinking and higher order thinking: A current perspective* (pp. 191-203). New York, NY: Nova Science Publishers.

Stark, E. (2012). Enhancing and assessing critical thinking in a psychological research methods course. *Teaching of Psychology, 39*, 107-112. doi:10.1177/0098628312437725

Szabo, Z., & Schwartz, J. (2011). Learning methods for teacher education: The use of online discussions to improve critical thinking. *Technology, Pedagogy and Education, 20*, 79-94. doi:10.1080/1475939X.2010.5344866

*Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal*. San Antonio, TX: Psychological Corporation.

Wesp, R., & Montgomery, K. (1998). Developing critical thinking through the study of paranormal phenomena. *Teaching of Psychology, 25*, 275-278. doi:10.1080/00986289809709714

Yanchar, S. C., Slife, B. D., & Warne, R. (2008). Critical thinking as disciplinary practice. *Review of General Psychology, 12*, 265-281. doi:10.1037/1089-2680.12.3.265

## Appendix: A Compendium of General Critical Thinking Measures, with Brief Descriptions

| Measure | Brief Description |
|---|---|
| California Critical Thinking Skills Tests | Based on information provided, tasks with increasing difficulty are presented. Separate scale scores available for analysis, interpretation, evaluation, explanation, deductive reasoning, inductive reasoning and a total critical thinking skills score. |
| Cambridge Thinking Skills Assessment | Presents 50 multiple choice questions measuring critical thinking and problem solving skills, including numerical and spatial reasoning, critical thinking, understanding arguments and everyday reasoning. Available online and paper and pencil forms. |
| Collegiate Assessment of Academic Proficiency (CAAP) Critical Thinking Test | Four passages are presented followed by a 32-item multiple choice test which students clarify, analyze, evaluate, and extend arguments. Total score is generated. |
| Collegiate Learning Assessment (CLA) Critical Thinking, Analytic Reasoning, and Problem Solving | Performance and analytic writing tasks are presented that measure a student's ability to evaluate evidence, analyze and synthesize evidence, draw conclusions, and acknowledge alternative viewpoints. |
| Cornell Critical Thinking Test | Students are tested on deduction, credibility, and identification of assumptions; appropriate for grade 5 to grades 12-14. |
| Ennis-Weir Critical Thinking Essay Test | Testing involves getting the point, reasoning and assumptions, offering alternative possibilities and explanations. Used for grade 7 through college. Assesses problem solving, critical thinking, and communication. |
| Halpern Critical Thinking Assessment | Respondents are presented with 25 everyday scenarios, and free responses are constructed; then, the scenarios are presented again requiring a forced choice response. This procedure helps to separate generation and recognition processes. |
| iCritical Thinking | Presented with 14 tasks based on real-world situations, this instrument is completed in 60 minutes and yields a digital literacy certification specific to critical thinking in a technology-enabled digital environment. |
| International Critical Thinking Essay Test | Involves analysis of a writing prompt (identify the elements of reasoning) worth 80 possible points, and assessment of a writing prompt (using analysis and evaluation) worth 20 possible points. |
| Measure of Academic Proficiency and Progress (MAPP) | Addresses reading, writing, mathematics, and critical thinking. The critical thinking sub-score ranges from 100 to 130. Students respond to multiple choice questions requiring evaluation, relevance, and recognition. Student performance is classified as proficient, marginal, or not proficient. |

| | |
|---|---|
| Proficiency Profile | This multiple choice instrument equates to the former Academic Profile, and yields a critical thinking proficiency level (Level I, II, or III). Available in standard form (108 questions) or abbreviated form (36 questions). |
| Watson-Glaser Critical Thinking Appraisal | Students are assessed on decision-making skills and judgment; test takers classified as low, average, or high in critical thinking ability. Using Form S, 40 self-report items are used; higher scores indicate greater critical thinking abilities. |

# Chapter 8: Student Engagement Toward Coursework: Measures, Considerations, and Future Directions

Kevin L. Zabel and Amy Heger

University of Tennessee

Student engagement is the fuel that drives the potential for success in college courses. Just as a car cannot operate without fuel, a lack of student interest or engagement hinders the beneficial impact of class-facilitated experiences. Although multiple operational definitions of student engagement exist, student engagement is broadly defined as the quantity and quality of physical and psychological energy that students invest in the college experience (Astin, 1999) or the extent to which students take part in educationally effective practices (Kuh, 2003). Regardless of construct definition, student engagement relates to academic achievement and several important learning outcomes. For example, student engagement is linked to persistence among first- and second-year college students (Kuh, Cruce, Shoup, Kinzie & Gonyea, 2008), retention and reduced drop-out rates (Finn, 1989), achievement (Newmann, 1992), grade point average (Carini, Kuh, & Klein, 2006), and a plethora of other positive outcomes (see Janosz, 2012 for a review). The importance of student engagement in courses was dramatically showcased in a field experiment where an engagement activity that connected science to students' personal lives increased student interest in class, as well as class performance, especially among students with low expectations of success (Hulleman & Harackiewicz, 2009). Although the importance of student engagement and interest in classroom material seems established, less founded are psychometrically sound and well-validated measures to accurately and reliably assess student engagement and interest toward classroom materials.

Student engagement is an attitude that, like all attitudes, varies among individuals in terms of strength and valence. Past scholars and researchers have utilized a multifaceted operationalization (Fredricks, Blumenfeld, & Paris, 2004; Furlong, Whipple, St. Jean, Simental, Soliz, & Punthuna, 2003; Jimerson, Campos, & Greif, 2003) to define student engagement. Specifically, student engagement consists of affective (e.g., relationships with peers and teachers, emotions), behavioral (e.g., effort, student participation), and cognitive (e.g., investment, personal goals, autonomy) components (Appleton, Christenson, & Furlong, 2008; Fredricks et al., 2004; Furlong et al., 2003; Jimerson et al., 2003). The distinction between affective, behavioral, and cognitive components of student engagement becomes clear when examining how each may importantly impact one another and lead to specific types of long-term consequences. For instance, a reduced sense of belonging within a school or classroom (affective component of student engagement) may lead to withdrawing from school activities (behavioral component), which in turn leads to cognitive perceptions ("school is not important to my self-concept") that have negative long-term consequences.

The multidimensional nature of student engagement has led to divergences in the measurement and operationalization of the construct. Indeed, a lack of a unifying theme in defining and thus measuring the construct is a problem elaborated on by many (e.g., Reschly & Christenson, 2012). Some education researchers have called for a more precise operationalization for measuring student engagement to allow for a refined understanding of how and under what circumstances student engagement predicts learning (Axelson & Flick, 2010). These considerations aside, much extant research has measured student interest toward teachers and inclusion within the *classroom*. However, a dearth of research has examined the measurement of student engagement as it pertains to *classroom material* in particular, especially regarding student engagement in postsecondary education settings. In what follows, we critically review widely-utilized measures relevant to assessing student engagement, broadly defined, toward class material, focusing on their commonalities and distinctions both from pragmatic and psychometric perspectives.

Researchers who study engagement have measured it at both a macro- and a micro-level. The macro-level form of engagement focuses on measuring elements related to investment and effort in school-related activities (e.g., Marks, 2000; Newmann, Wehlage, & Lamborn, 1992; Skinner, Wellborn, & Connell, 1990) and identification or connection broadly with school and academics (Finn, 1993; Finn & Rock, 1997). The micro-level form of engagement focuses on measuring engagement in one particular school-related aspect, such as a particular course or student activity (Handelsman, Briggs, Sullivan, & Towler, 2005). Although macro- and micro-level engagement measures are positively associated, micro-level measures may allow for a more nuanced understanding of engagement toward a particular class and the relevant behavioral, cognitive, and affective factors in that context. For example, faculty can all recall the example of a student who is purported to be engaged in fellow faculty members' classes, but clearly is not in one's own class. This example is simple but illustrates an important point: macro-level measures of student engagement may not be as valid in assessing student engagement toward particular aspects of class material or a particular class. However, macro-level student engagement measures can be especially useful in uncovering a student's engagement with school more generally, which itself can have several positive outcomes. As with any measure, trade-offs exist in using micro, relative to macro, measures. Given our focus on student engagement toward coursework, the bulk of this chapter focuses on micro-level measures.

## Macro-Level Measure

### National Survey of Student Engagement (NSSE)

The main macro-level measure utilized to assess student engagement is the NSSE (Kuh, 2001). The NSSE is used extensively by colleges and universities to evaluate engagement among freshman and senior students. The NSSE consists of a series of college activity (behavior) items directly measuring student engagement, educational and personal growth items, and items regarding opinions about one's school, as well as a variety of other questions. The NSSE positively predicts a variety of learning-related outcomes, including grade point average, critical thinking, and standardized test scores (Carini et al., 2006; Ewell, 2002; Pascarella & Terenzini, 2005). Nevertheless, the NSSE is designed to assess self-engagement broadly toward the college experience, and not particularly toward coursework from specific classes. Furthermore, the NSSE lacks a rigorous theoretical orientation that drives the organization and use of particular items. Indeed, a point of emphasis in the current chapter is that measures of student engagement can be improved by being grounded in theoretical rationale.

## Micro-Level Measures

### Student Interest

Student engagement has been operationalized and measured in a variety of manners. We focus the remainder of this chapter on several micro-level measures utilized in previous research to assess student engagement, broadly defined. Student interest is one such manner in which engagement toward college courses and material has been assessed.

#### *13-Item Student Interest Measure*

One useful measure of student interest is Harackiewicz, Barron, Tauer, Carter, and Elliott's (2000) 13-item measure. The items within this measure focus on interest toward a particular class, the lectures within the class, and the professor teaching the class. This measure has some overlap with a previous 7-item measure of student interest (Harackiewicz, Barron, Carter, Lehto, & Elliot, 1997). However, this 13-item measure, unlike its predecessor (Harackiewicz et al., 1997), differentiates between "catch" and "hold" interest factors. Catch interest factors initially trigger student interest, and may consist of flashy Powerpoint slides, gripping examples, or stimulating teaching methods that lead to initial class enjoyment. Three items assess catch interest (α = .93), including "I don't like the lectures very much (reverse-coded)," "The lectures in this class seem to drag on forever (reverse-coded)", and "I like my professor." Hold interest factors include course elements that maintain student interest over a sustained period of time by activating intrinsic motivation (Harackiewicz et al., 2000). Ten items (α = .72) assess hold interest, such as "I think the course material in this class is useful for me to learn" and "I think the field of psychology is very interesting." Participants respond to items using a 1 (*strongly disagree*) to 7 (*strongly agree*) response range. The 3-item measure of catch interest is labeled as Enjoyment, and the 10-item measure of hold interest is labeled as an Interest Scale (Harackiewicz et al., 2000).

Interestingly, the hold measure of student interest (i.e., the Interest Scale) was a better predictor of performance (i.e., learning) in classes and of the number of subsequent psychology classes students took, suggesting that it might be an especially important dimension to facilitate among students to promote learning outcomes. The catch and hold measures are positively correlated ($r$ = .58), but their operational definitions suggest that they are conceptually distinct dimensions of student interest that should be treated as such in future research (Harackiewicz et al., 2000). Psychometrically, it is important to note that the "catch" interest scale only has three items, and like many other measures of student interest and engagement, lacks proper validation. Future research should examine unique ways that catch and hold interest predict student-related outcomes, as well as means by which they can be obtained through teaching methods and other features of particular classes, keeping in mind that the factors that enhance catch and hold interest may vary as a function class type subject matter and level.

### *Interest as an Emotion*

Student interest can also be operationalized as an emotion (Izard, 1977) that focuses attention and openness to receiving information (Dougherty, Abe, & Izard, 1996). From this perspective, interest is a stable emotion that should influence intrinsic motivation to learn and interest in academic endeavors. The Interest Subscale of the Differential Emotions Scale (Izard, Libero, Putnam, & Haynes, 1993) has been utilized by previous researchers (Bye, Pushkar, & Conway, 2007) to operationalize interest broadly as a dispositional emotion relevant to academic outcomes. The Interest Subscale (α = .75; Bye et al., 2007) is a 3-item measure of the degree to which individuals generally experience the emotion of interest, and consists of the items "How often do you feel so interested in what you're doing that you're caught up in it?," "How often do you feel like what you're doing or watching is interesting?," and "How often do you feel alert, curious, and kind of excited about something?" Participants respond using a 1 (*rarely or never*) to 5 (*very often*) response range. Interest conceptualized this way is positively correlated with intrinsic motivation to learn (Bye et al., 2007). It is important to note, however, that this measure of interest is macro-level, and does not pertain to student interest or engagement to particular aspects of classroom material, though items could easily be adapted to assess the degree to which students experience the emotion of interest specifically toward classes.

## Student Engagement

### *Student Course Engagement Questionnaire (SCEQ)*

Handelsman and colleagues (2005) created the SCEQ to assess student engagement specific to a college course. Contrary to other measures of student engagement (e.g., NSSE), this measure assesses more micro-level class engagement. Participants rate whether each of 23 items are descriptive of them using a 1 (*not at all characteristic of me*) to 5 (*very characteristic of me*) response range. Handelsman and colleagues' exploratory factor analysis revealed that student course engagement was composed of four factors: 1) a *skills engagement factor* (α = .82) that captures engagement by means of practicing skills (example items are "Taking good notes in class" and "Making sure to study on a regular basis"), 2) an *emotional engagement factor* (α = .82) that captures engagement by means of emotional involvement with class material (example items are "Finding ways to make the course interesting to me" and "Applying course material to my life"), 3) a *participation/interaction engagement factor* (α = .79) that captures engagement by means of class participation and interactions with teachers and other students (example items are "Participating actively in small-group discussions," and "Helping fellow students"), and 4) a *performance engagement factor* (α = .76) that captures engagement by means of performance level (example items are "Getting a good grade" and "Doing well on the tests").

Furthermore, Handelsman and colleagues (2005) demonstrated the predictive validity of the SCEQ by finding that particular factors of the multi-faceted engagement construct were positively related to midterm and final exam grades in a mathematics course. Additional studies employing the SCEQ have shown other significant relationships with academic initiatives and outcomes. For example, participation in student management teams (SMTs) increases course engagement, and it is through this engagement that SMTs positively influence course performance (Troisi, 2014). Also, different types of engaged learning activities in college (e.g. internships, undergraduate research) promote particular factors of engagement from the SCEQ (Miller, Rycek & Fritson, 2011).

Overall, given its characteristics, the SCEQ may be a useful measure of student engagement when a particular domain or class is the chief target of interest. Handelsman and colleagues (2005) mention categories from the macro-level NSSE that they anticipated would parallel particular SCEQ factors, but they did not collect data to compare the two scales. Future research should use a longitudinal design to investigate if student scores on the SCEQ across courses converge to predict overall college engagement on the NSSE. Worrisome aspects of the scale, however, need to be highlighted. Specifically, no theoretical foundation was articulated in SCEQ questionnaire item development or regarding a multi-dimensional factor structure. Reasoning was provided post-hoc for each factor and references of supportive research were somewhat sparse. The dependent variables chosen to test patterns for convergent and discriminant validity were one-item measures, which are notorious for being unreliable (e.g., Churchill, 1979; Guilford, 1954), so additional rigorous testing of the scale's psychometric properties should be pursued. Indeed, the lack of rigorous psychometric testing of teaching-related measures is an issue discussed in Chapter 3 of this e-book (Christopher, 2015).

### *Utrecht Work Engagement Scale-Student (UWES-S).*

An additional and more psychometrically rigorous assessment of student engagement has arisen out of the organizational psychology literature. The UWES-S evolved out of research on engagement in the workplace  (e.g., Schaufeli & Bakker, 2003; Schutte, Toppinen, Kalimo, & Schaufeli, 2000), and consists of a three-dimensional, 17-item measure of student engagement (Schaufeli, Salanova, González-Romá, & Bakker, 2002).

Engagement is delineated by three core-dimensions: vigor, dedication, and absorption. Vigor is operationalized as high levels of energy, devoting effort to one's work, and mental resilience and perseverance. Six items are used to assess vigor (e.g., "When I'm doing my work as a student, I feel bursting with energy"). Dedication is operationalized as a sense of dedication, significance, enthusiasm, inspiration, pride, and challenge in one's work. Dedication is assessed by five items (e.g., "I'm enthusiastic about my studies"). Absorption is operationalized as being fully concentrated and deeply engrossed in one's work (Schaufeli et al., 2002, pp. 74). Absorption is assessed by six items (e.g., "When I am studying, I forget everything else around me"). Item responses are recorded on a Likert scale from 0 (*never*) to 6 (*always*). Recent cross-national research has confirmed the sound psychometric properties of the UWES-S (Schaufeli, Martínez, Marques-Pinto, Salanova, & Bakker, 2002; Schaufeli et al., 2002), as well as the generality of the measurement model across cultures for the absorption and vigor subscales. Future work should examine if the dimensions structure is consistent across different academic levels (i.e., elementary, secondary, post-secondary) as well.

Overall, the UWES-S does a solid job of assessing involvement and more personal seeming aspects of the engagement experience with its items addressing investment, persistence, dedication, loss of time, etc. The non-education origin of the engagement scale should not be of concern because its psychometric properties have been investigated rigorously, more so than those of the SCEQ, for instance. Communication across disciplines is often less than satisfactory (Christopher, 2015) and the UWES-S measure of student engagement is an instance in which teaching of psychology researchers could gain from existing measures in the organizational psychology literature.

## Ancillary Measures

### *Grit*

Several measures that may not be considered traditional measures of student engagement may hold utility in tapping into the construct. Grit is one such measure, and is defined as perseverance and passion for long-term goals and interests (Duckworth, Peterson, Matthews, and Kelly, 2007, pp. 1087). The prominent grit measure is Duckworth et al.'s (2007) 12-item, two-factor measure (α = .85). The two factors include Consistency of Interests (6 items α = .84) and Perseverance of Effort (6 items; α = .78). Consistency items include "I often set a goal and decide to pursue a new one" (reverse-coded) and "My interests change from year to year" (reverse-coded). Perseverance of Effort items include "I am a hard worker" and "I have achieved a goal that took years of work." A shorter 8-item version (Duckworth & Quinn, 2009) with the same factor structure as the original grit scale (Duckworth et al., 2007) has also been well-validated (α = .83).

Grit conceptually shares similarities with the student engagement and interest measures previously reviewed, it is distinct in its focus on long-term, consistent effort and perseverance. Although not yet used in the teaching of psychology realm, we believe that measures of grit may hold particular relative advantages over student interest and engagement measures in assessing engagement toward course material in which consistent, long-term engagement is required (e.g., multi-stage assignments such as empirical research papers). Indeed, utilizing measures of grit in combination with measures of student engagement and interest may hold advantages such as aiding in distinguishing students who lack engagement with a particular activity or aspect of a specific course, in comparison to those students who consistently lack engagement toward aspects of courses in general over the long-term.

### *Boredom*

Another worthwhile perspective on engagement is to examine the opposite, the tendency to withdraw from academic activities. Boredom is described as an unpleasant emotional state characterized by low arousal and a lack of stimulation (Mikulas & Vodanovich, 1993). Boredom is understood as a relevant emotion for academic experiences, and has been investigated as one of eight achievement emotions measured by the achievement emotions questionnaire (AEQ; Pekrun, Goetz, & Perry, 2005; Pekrun, Goetz, Titz, & Perry, 2002).

### *Achievement Emotions Questionnaire (AEQ)*

Achievement emotions can arise in different academic settings. Boredom, therefore, is measured with two scales in the AEQ; one specifies it as a learning-related emotion and the other as a class-related emotion. Instructions differ by asking participants to rate on a Likert scale from 1 (*completely disagree*) to 5 (*completely agree*) how the items pertain to feelings they may experience during studying (learning-related) or during class (class-related; each 11 items). Learning-related boredom scale items (α = .93) include "I find my mind wandering while I study" and "Because I'm bored I have no desire to learn." Class-related boredom scale items (α = .92) include "I find this class fairly dull" and "I'm tempted to walk out of the lecture because it is so boring."

Pekrun, Goetz, Daniels, Stupnisky, and Perry (2010) classify boredom as a negative deactivating achievement emotion because it diminishes motivation and can have negative effects on performance. In particular, boredom measured with the AEQ is positively related to attention problems and negatively related to intrinsic motivation to learn, study effort, academic performance scores and other detrimental educational outcomes (Pekrun et al., 2010). Pekrun, Hall, Goetz, and Perry (2014) found that, with a longitudinal study across the academic year, a reciprocal cycle exists between course-related boredom and exam performance, such that boredom negatively impacts subsequent exam performance and exam performance negatively impacts subsequent boredom. Unlike a lack of interest or enjoyment—neutral states characterized by an absence of approach motivation—boredom instigates an avoidance motivation (Pekrun et al., 2010). A desire to escape the situation occurs, which might lead students to disengage from their learning and course-work.

### *Academic Boredom Survey*

The 10-item Academic Boredom Survey (ABS-10) is another scale created to assess boredom (Acee, Kim, Kim, Kim, Hsiang-Ning, Kim, Cho, Wicker, & the Boredom Research Group, 2002). The ABS-10 takes into account whether boredom originated from a task being too hard or too easy by asking participants to recall a situation in which they were over-challenged or under-challenged. Instructions specify to respond on a Likert scale from 1 (*not at all*) to 9 (*extremely*) the extent to which an item is true for each situation. Sample items are "Want to do something else" and "Find the activity dull." Confirmatory factor analysis of the scale revealed a 2-factor solution for instances of being over-challenged (i.e. self-focused factor and task-focused factor) and a 1-factor solution for instances of being under-challenged. These three factors have demonstrated excellent levels of reliability (i.e., α's = .86, .80, & .90, respectively; Acee et al., 2010). However, future research should aim to further validate the measure, due to drawbacks in analyses and assumptions made in initial measure design. Additionally, the results of the ABS-10 compared to the AEQ, while found to be significantly related to boredom, were also related to a greater extent to other achievement emotions (i.e. anger, hopelessness) for self-focused scale items. Overall, boredom research using the AEQ has been most productive in validating a measure of boredom and establishing its harmful relationship for academic performance outcomes (e.g. Pekrun et al., 2010; Pekrun et al., 2014). However, the ABS-10 gives scholars another measurement tool that is relevant for understanding differences in boredom as a result of task characteristics (i.e., being too challenged or not challenged enough).

## Recent Developments

### Student Engagement Scale (SES)

One recent measure of college student engagement that is encouraging with regard to its theoretical base and components is the SES (Gunuc & Kuzu, 2015). The SES is a six-factor measure of student engagement that examines both campus and class engagement and specifically the affective, cognitive, and behavioral components of class engagement. Thus, it is rooted in the tripartite attitude conceptualization of student engagement, and also considers the macro (i.e., campus) and micro (i.e., class) factors of student engagement. The items of the SES were derived through previous measures of engagement and qualitative pilot interviews, as well as based in an underlying theoretical perspective. Reliability for the overall SES is excellent (α = .96), as it is for each of the six smaller factors as well (α's range from .81 to .91). Furthermore, several aspects of the design and implementation of this measure are laudable. Specifically, Gunuc and Kuzu (2015) grounded the measure development in theory and linked it to previous research considering the macro vs. micro distinctions and the multidimensional (e.g., affective, behavioral, and cognitive) nature of engagement. Although research using the SES is in its infancy and further empirical research is needed to validate the measure, we believe it to hold great promise as a student engagement measure for future research.

## Discussion and Future Directions

Our hope is to have provided a range of empirically-validated measures relevant to assessment of student interest and engagement in class material. It is important to highlight, however, that with a lack of consensus on an operational definition of interest and engagement in course material, many researchers have created their own measures for the purpose of measuring interest and engagement relevant to their particular research questions. For example, some researchers create scales where they select only certain items for inclusion from preexisting measures (e.g. Gasiewski, Eagan, Garcia, Hurtado, & Chang, 2012; Langley, 2006; Svanum & Bigatti, 2009), or create their own items (e.g. Marks, 2000; Miller et al., 2011; Taneja, Fiore & Fischer, 2015). We express concern over such practices due to critical deficits in investigation of the psychometric qualities in many of these instances.

Specialized measures should not be overlooked, nonetheless. The ways in which students engage with engineering coursework versus poetry might vary greatly. Limitations of the generalizability of the catchall engagement measures we review here are likely. In fact, some researchers have adapted versions to address specific topics such as the math version of the AEQ (AEQ-M; Pekrun, Goetz, & Frenzel, 2005). We did not discuss in depth interest in domain-specific material, but feel it is necessary to emphasize the importance of such approaches.

Recent efforts (Saddler, Sonnert, Hazari, & Tai, 2012) to investigate females' interest and involvement in the STEM fields (or lack thereof) is just one example of the importance of specialization.

When discussing student interest and engagement, it is vital to call attention to new opportunities for measurement with the constantly changing landscape of the traditional learning environment. The development of a new digital age, where electronics and media have become more versatile and prevalent, has generated a new area of student engagement research (beyond the scope of scales discussed here). Empirical investigations have begun to explore technology's role from concerns of maintaining engagement via new online class platforms, to enhancing engagement via integration of such new technologies during class, to inhibiting engagement via electronic distractions in the lecture hall. One widespread example of how technology advances have permeated educational settings is the usage of audience response systems at universities (Abrahamson, 2006). One version is wireless remotes called clickers that enable students to instantaneously respond to questions during lectures. Clickers are unique from traditional hand-raising as a response system in that they provide anonymity, oftentimes prompting greater participation. Indeed, research has uncovered increases in interest, understanding, and attendance in biology courses with the use of clickers (Preszler, Dawe, Shuster, & Shuster, 2007).

The role of video games, virtual reality, and podcasts in academic courses have also begun to be discussed (Bouta, Retalis & Paraskeva, 2012; Peden & Domask, 2011; Shernoff, 2013). Social media additionally presents another prospective avenue in which students might be more likely to interact with course material. Although using social media (e.g., Twitter) as an educational tool can be beneficial in some contexts (Junco, Heiberger, & Loken, 2011), communicative activities on social media (e.g., Facebook commenting) are negatively associated with college engagement (Junco, 2012). Indeed, constant connectively might have detrimental effects for learning environments, particularly conceivable as a distraction from engaging with the course material to be learned (Sana, Weston, & Cepeda, 2013; Taneja et al., 2015). The effects of technologies on engagement are a ripe area for future investigation.

As is highlighted in this review, diverse measures can be used to examine both micro and macro levels of student engagement. Future researchers should be mindful to choose measures for use based on the theoretical perspective of their research question. Thus, if a researcher's theoretical perspective is based on an affective conceptualization of student engagement toward a particular class or aspect of class, then a student interest (Harackiewicz et al., 2000) or boredom (Acee et al., 2010; Pekrun et al., 2005) measure may be ideal. If a multidimensional conceptualization of student engagement toward a class is required, then the SES (Gunuc & Kuzu, 2015) or the SCEQ (Handelsman et al., 2005) may be ideal measures. Researchers interested in predicting long-term engagement may want to consider the Grit Scale (Duckwoth et al., 2007) as a viable measure. Furthermore, the validation of domain-specific measures of student engagement may be especially fruitful to predict domain-specific outcomes. Finally, measures of student engagement that focus on "catch" elements of student engagement that may lead to "hold" factors of student engagement that mediate student-related outcomes could be particularly productive given enhancements in teaching technology and teaching methods. Overall, a need for multidimensional measures of student engagement to be more theoretically-based and thoroughly validated presents many avenues for future research on the topic.

# References

References marked with an asterisk indicate a scale.

Abrahamson, L. (2006). A brief history of networked classrooms: Effects, cases, pedagogy, and implications. In D. A. Banks (Ed.), *Audience response systems in higher education: Applications and cases* (pp. 1-25). Hershey, PA: Information Science Publishing.

*Acee, T. W., Kim, H., Kim, H. l., Kim, 1., Hsiang-Ning, R. C, Kim, M., Cho, Y., Wicker, F. W., & The Boredom Research Group (2010). Academic boredom in under- and over-challenging situations. *Contemporary Educational Psychology*, *35,* 17- 27. doi:10.1016/j.cedpsych.2009.08.002

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*, 369-386. doi:10.1002/pits.20303

*Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*, 427-445. doi:10.1016/j.jsp.2006.04.002

Astin, A. W. (1999). Student involvement: A developmental theory for higher education. *Journal of College Student Development*, *40*, 518-529.

Axelson, R. D., & Flick, A. (2010). Defining student engagement. *Change: The Magazine of Higher Learning*, *43*, 38-43. doi:10.1080/00091383.2011.533096

Bouta, H., Retalis, S., & Paraskeva, F. (2012). Utilizing a collaborative macro-script to enhance student engagement: A mixed method study in a 3D virtual environment. *Computers & Education*, *58*, 501-517. doi:10.1016/j.compedu.2011.08.031

Bye, D., Pushkar, D., & Conway, M. (2007). Motivation, interest, and positive affect in traditional and nontraditional undergraduate students. *Adult Education Quarterly*, *57*, 141-158. doi:10.1177/0741713606294235

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages. *Research in Higher Education*, *47*, 1-36. doi:10.1007/s11162-005-8150-9

Christopher, A. N. (2015). Selecting the right scale: An editor's perspective. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Churchill Jr., G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16,* 64-73. doi:10.2307/3150876

Dougherty, L. M., Abe, J., & Izard, C. E. (1996). Differential emotions theory and emotional development in adulthood and later life. In C. Magai & S. H. McFadden (Eds.), *Handbook of emotion, adult development, and aging* (pp. 27-41). San Diego, CA: Academic Press.

*Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087–1101. doi:10.1037/0022-3514.92.6.1087

*Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S). *Journal of Personality Assessment*, *91*, 166–174. doi:10.1080/00223890802634290

Ewell, P. T. (2002). *An Analysis of Relationships between NSSE and Selected Student Learning Outcomes Measures for Seniors Attending Public institutions in South Dakota*, National Center for Higher Education Management Systems, Boulder, CO.

Finn, J. D. (1993). *School engagement and students at risk*. Washington, D. C.: National Center for Educational Statistics.

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, *59*, 117-143. doi:10.3102/00346543059002117

Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology*, *82*, 221-234. doi:90/10.1037/0021-9010.82.2.221

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*, 59-109. doi:10.3102/00346543074001059

Fredericks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763–782). New York, NY: Springer.

Furlong, M. J., Whipple, A. D., St. Jean, G., Simental, J., Soliz, A., & Punthuna, S. (2003). Multiple contexts of school engagement: Moving toward a unifying framework for educational research and practice. *The California School Psychologist*, *8*, 99-113. doi:10.1007/BF03340899

Gasiewski, J. A., Eagan, M. K., Garcia, G. A., Hurtado, S., & Chang, M. J. (2012). From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory STEM courses. *Research in Higher Education*, *53*, 229-261. doi:10.1007/s11162-011-9247-y

González-Romá, V., Schaufeli, W. B., Bakker, A., & Lloret, S. (2006). Burnout and engagement: Independent factors or opposite poles? *Journal of Vocational Behavior*, *68*, 165-174. doi:10.1016/j.jvb.2005.01.003

Guilford, J. P. (1954). *Psychometric methods*. New York, NY: McGraw–Hill.

*Gunuc, S., & Kuzu, A. (2015). Student engagement scale: Development, reliability, and validity. *Assessment and Evaluation in Higher Education*, *40*, 587-610. doi:10.1080/02602938.2014.938019

*Handelsman, M. M., Briggs, W. L., Sullivan, N., & Towler, A. (2005). A measure of college student course engagement. *The Journal of Educational Research*, *98*, 184-192. doi:10.3200/JOER.98.3.184-192

*Harackiewicz, J. M., Barron, K. E., Carter, S. M., Lehto, A. T., & Elliot, A. J. (1997). Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social Psychology*, *73*, 1284-1295. doi:90/10.1037/0022-3514.73.6.1284

*Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, *92*, 316-330. doi:10.1037//0022-0663.92.2.316

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*, 1410-1412. doi:10.1126/science.1177067

Izard, C. E. (1977). *Human emotions*. New York: Plenum.

*Izard, C. E., Libero, D. Z., Putman, P., & Haynes, O. M. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology*, *64*, 847-860. doi:10.1037/0022-3514.64.5.847

Janosz, M. (2012). Outcomes of engagement and engagement as an outcome: Some consensus, divergences, and unanswered questions. In S. L. Christenson, A. L. Reschly, & C Wylie (Eds.), *Handbook of research on student engagement* (pp. 695-706). New York, NY: Springer.

Jimerson, S. R., Campos, E., & Greif, J. L. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *The California School Psychologist*, *8*, 7-27. doi:10.1007/BF03340893

Junco, R. (2012). The relationship between frequency of Facebook use, participation in Facebook activities, and student engagement. *Computers & Education*, *58*, 162-171. doi:10.1016/j.compedu.2011.08.004

Junco, R., Heiberger, G., & Loken, E. (2011). The effect of Twitter on college student engagement and grades. *Journal of Computer Assisted Learning*, *27*, 119-132. doi:10.1111/j.1365-2729.2010.00387.x

*Kuh, G. D. (2001). Assessing what really matters to student learning inside the national survey of student engagement. *Change: The Magazine of Higher Learning*, *33*, 10-17. doi:10.1080/00091380109601795

Kuh, G. D. (2003). What we're learning about student engagement from the NSSE: Benchmarks for effective educational practices. *Change: The Magazine of Higher Learning*, *35*, 24-32. doi:10.1080/00091380309604090

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year grades and persistence. *The Journal of Higher Education*, *79*, 540-563. doi:10.1353/jhe.0.0019

Langley, D. (2006). The student engagement index: A proposed student rating system based on the national benchmarks of effective educational practice. *University of Minnesota: Center for Teaching and Learning Services.*

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal*, *37*, 153-184. doi:10.3102/00028312037001153

Mauno, S., Kinnunen, U., & Ruokolainen, M. (2007). Job demands and resources as antecedents of work engagement: A longitudinal study. *Journal of Vocational Behavior*, *70*, 149-171. doi:10.1016/j.jvb.2006.09.002

Mikulas, W. L., & Vodanovich, S. J. (1993). The essence of boredom. *The Psychological Record*, *43*, 3-12. doi:10.3389/fpsyg.2014.01245

Miller, R. L., Rycek, R. F., & Fritson, K. (2011). The effects of high impact learning experiences on student engagement. *Procedia-Social and Behavioral Sciences*, *15*, 53-59. doi:10.1016/j.sbspro.2011.03.050

Newmann, F. M. (1992). *Student engagement and achievement in American secondary schools*. New York, NY: Teachers College Press. doi:10.5860/CHOICE.30-3945

Newmann, F. M., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. In F. M. Newman (Ed.), *Student engagement and achievement in American secondary schools* (pp. 11–39). New York, NY: Teacher's College Press. doi:10.5860/CHOICE.30-3945

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of* *research* (Vol. 2). San Francisco: Jossey-Bass.

Peden, B. F., & Domask, B. Z. (2011). Do podcasts engage and educate college students? In R. L. Miller, E. Amsel, B. M. Kowalewski, B. C. Beins, K. D. Keith, & B. F. Peden (Eds.), *Promoting Student Engagement* (pp. 170-177). Society for the Teaching of Psychology (e-book).

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*, 531–549. doi:10.1037/a0019243

Pekrun, R., Goetz, T., & Frenzel, A. C. (2005). *Achievement Emotions Questionnaire—Mathematics (AEQ-M): User's manual.* Unpublished manual, University of Munich, Department of Psychology.

Pekrun, R., Goetz, T., & Perry, R. P. (2005). *Academic Emotions Questionnaire (AEQ): User's manual*. Munich, Germany: University of Munich, Department of Psychology. doi:10.1016/j.cedpsych.2010.10.002

*Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*, 91-105. doi:10.1207/S15326985EP3702_4

Pekrun, R., Hall, N. C., Goetz, T., & Perry, R. P. (2014). Boredom and academic achievement: Testing a model of reciprocal causation. *Journal of Educational Psychology*, *106*, 696-710. doi:10.1037/a0036006696

Preszler, R. W., Dawe, A., Shuster, C. B., & Shuster, M. (2007). Assessment of the effects of student response systems on student learning and attitudes over a broad range of biology courses. *CBE-Life Sciences Education*, *6*, 29-41. doi:10.1187/cbe.06-09-0190

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 3-20). New York, NY: Springer.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education*, *96*, 411-427.

Salanova, M., Schaufeli, W., Martínez, I., & Bresó, E. (2010). How obstacles and facilitators predict academic performance: The mediating role of study burnout and engagement. *Anxiety, Stress & Coping*, *23*, 53-70. doi:10.1080/10615800802609965

Sana, F., Weston, T., & Cepeda, N. J. (2013). Laptop multitasking hinders classroom learning for both users and nearby peers. *Computers & Education, 62*, 24-31. doi:10.1016/j.compedu.2012.10.003

*Schaufeli, W. B., & Bakker, A. B. (2003). *UWES – Utrecht Work Engagement Scale: Test Manual.* Utrecht, The Netherlands: Department of Psychology, Utrecht University.

Schaufeli, W. B., Leiter, M. P., Maslach, C., & Jackson, S. E. (1996). Maslach Burnout Inventory – General Survey (MBI-GS). In C. Maslach, S. E. Jackson, & M. P. Leiter (Eds.), *MBI manual* (3rd ed.), (pp. 22–26). Palo Alto, CA: Consulting Psychologists Press.

Schaufeli, W.B., Martínez, I., Marques-Pinto, A., Salanova, M., & Bakker, A. (2002). Burnout and engagement in university students: A cross-national study. *Journal of Cross-Cultural Psychology, 33*, 464-481. doi:10.1177/0022022102033005003

*Schaufeli, W.B., Salanova, M., González-Romá, V., & Bakker, A. (2002). The measurement of burnout and engagement: A confirmatory factor analytic approach. *Journal of Happiness Studies, 3,* 71-92. doi:10.1023/A:1015630930326

Schutte, N., Toppinen, S., Kalimo, R., & Schaufeli, W. (2000). The factorial validity of the Maslach Burnout Inventory ‐ General Survey (MBI ‐ GS) across occupational groups and nations. *Journal of Occupational and Organizational Psychology*, *73*, 53-66. doi:10.1348/096317900166877

Shernoff, D. J. (2013). *Optimal Learning Environments to Promote Student Engagement*. New York, NY: Springer.

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology*, *82*, 22-32. doi:0022-0663/90/S00.75

Svanum, S., & Bigatti, S. M. (2009). Academic course engagement during one semester forecasts college success: Engaged students are more likely to earn a degree, do it faster, and do it better. *Journal of College Student Development*, *50*, 120-132. doi:10.1353/csd.0.0055

Taneja, A., Fiore, V., & Fischer, B. (2015). Cyber-slacking in the classroom: Potential for digital distraction in the new age. *Computers & Education*, *82*, 141-151. doi:10.1016/j.compedu.2014.11.009

Troisi, J. D. (2014). Making the grade and staying engaged: The influence of student management teams on student classroom outcomes. *Teaching of Psychology*, *41*, 99-103. doi:10.1177/0098628314530337

# Chapter 9: Measuring Service-Learning and Civic Engagement

Lori Simons

Widener University

During the past three decades, the distribution of academic-based service-learning courses has expanded in undergraduate liberal arts programs in public and private post-secondary institutions (Giles & Eyler, 2013; Marxen, 2003). Institutions of higher education have incorporated academic-based service-learning courses in liberal arts curricula as a way to teach undergraduate students to think critically about the conditions that lead to social and racial disparities in the community and to develop into responsible citizens (Quaye & Harper, 2007). Academic-based service-learning (ABSL) is a pedagogical approach that requires students to connect the course content to the service context through application, reflection and discussion (Eyler & Giles, 1999).

Investigations on ABSL have assessed the impact of service-learning on student learning, community partnerships, and faculty engagement. In fact, the majority of these studies have focused on the benefits from ABSL on student learning outcomes (i.e., critical thinking, civic engagement). For example, students who participated in ABSL reported a deeper understanding of the course content compared to those students who did not participate in ABSL (Litke, 2002, Strage, 2000). Studies have also documented the impacts of ABSL on the development of leadership attributes (Moely, McFarland, Miron, Mercer, & Ilustre, 2002; Vogelesgang & Astin, 2000), interpersonal skills (Eyler, 2000; Moore, 2000), diversity attitudes (Boyle-Baise & Kilbane, 2000; Rockquemore & Shaffer, 2000), and social responsibility and civic engagement (Reinke, 2003).

Much of this research indicates that civic engagement is a service-learning outcome (Eyler & Giles, 1999). Civic engagement a broad term with multiple definitions (Hatcher, 2010). Civic engagement activities are designed to promote socially-responsible leadership in students by having them work with community recipients to solve problems. Examples of these activities include volunteering in a soup kitchen or writing a letter to an elected official (APA, 2015). In addition, civic engagement often refers to civic professionalism, social responsibility, and community engagement (Hatcher, 2010; Steinberg, Hatcher, & Bringle, 2011). Academic endeavors ranging from service-learning, internships, and other forms of experiential learning are used to instill the values and characteristics associated with civic engagement (Association for Experiential Education, 2011).

The scholarship on ABSL has evaluated the impact of service-learning from one of two research methodologies. Quantitative methods have assessed student attitudes before and after service with single, convenient samples. With this method, it is not possible to detect whether the reported changes are attributed to the service experience because the studies lack a comparison group (Payne, 2000; Reinke, 2003; Root, Callahan, & Sepanski, 2002) and measure attitudes instead of skills with either single-item surveys (Rockquemore & Schaffer, 2000), reflective essays (Green, 2001) or ethnographic techniques (Boyle-Base & Kilbane, 2000). Surveys are also used to measure changes in student skills from the beginning to the end of the course. Pretest and posttest surveys are usually administered to students in a course with a service component and compared to those in a course without a service component to measure changes before and after service. This pre-post methodology may gloss over what actually happens within the service experience, because it does not measure changes or detect processes that occur while engaged in service (Boyle-Baise, 2002; Wang & Jackson, 2005).

In contrast, qualitative methods have identified the processes associated with the development of cultural awareness (Rockquemore & Shaffer, 2000; Root et al., 2002) and social, community and civic responsibility (Battistoni, 2002; Eyler, 2009). Qualitative research has been criticized for relying on small samples of White, middle-class students and analytic techniques (i.e., ethnographic, focus groups) that make replication difficult (Boyle-Baise, 2002). The uniqueness of the service context further limits the researchers' ability to generalize the findings (Simons et al., 2012). It is important to recognize that there is a trade-off between external and internal validity in assessment methods of service-learning (Wang, Kelly, & Hritsuk, 2003). However, one area that has received substantially less attention in the literature is service-learning and civic engagement measures (i.e., scales). Although scholars have developed measures to assess service-learning impacts, most of them have constructed single items to assess student learning. In fact, Bringle, Phillips, and Hudson (2004) suggest that the development and implementation of multiple-item measures are necessary if the field of service-learning is to advance. The use of reliable and valid measures to assess service impacts (i.e., civic engagement) may be the next step in the advancement of this area of scholarship. The purpose of this chapter is to describe service-learning and civic engagement measures and distinguish between those with and without psychometric evidence.

## Method

### Search Strategy

A comprehensive search of service-learning measures was conducted using electronic databases search such as *Ebscohost* and *ERIC*. Keywords including service-learning, civic engagement, social responsibility, civic attitudes, knowledge and skills, and community engagement surveys, measures, instruments, and scales were used in this search. The search led to a small number of scholarly publications. A broader search was conducted using *Google Scholar*. This search identified different publication sources, including conference proceedings, books, and journals (i.e., *Teaching and Learning, Higher Education*). Few peer-review or scholarly journals were devoted solely to service-learning and civic engagement. The publication sources identified from the Google search was compared to those identified on the National Service-Learning Clearinghouse (NSLC) website (2015). The NSLC website listed 11 sources, but only seven of them include a peer review process for publication. The peer review process differs and can be more rigorous for journals than for conference proceedings or edited volumes. To limit the scope of the review process, only articles were reviewed that were published in the *Michigan Journal of Community Service-Learning (MJCSL).* The *MJCSL* is a premier journal in the area of service-learning and community engagement. The goals of this journal are to encourage scholarship on community engagement, contribute to the intellectual vigor and academic legitimacy of service-learning and community-engaged scholarship, and expand the knowledge base on academic service-learning (Howard, n.d.).

## Sample

The *MJCSL* published 274 articles from fall 1994 to spring 2015. Each article was examined using an open coding procedure (Creswell, 1994 & 1998) to determine if the article included at least one of the following components: (a) research or pedagogical study; (b) student or program outcomes; (c) quantitative data collection methods; or (d) items, scales or other tools to measure service-learning or civic engagement. A final set of 67 articles was used in the analyses as shown in Appendix A.

## Article Coding

Data (i.e., 67 articles) underwent an item analysis through which patterns were identified and coded using the constant comparative method (Creswell, 2005). Open coding consisted of categorizing the measures according service-learning constructs (Eyler & Giles, 1999) as shown in Table 1. Axial coding consisted of systematically analyzing the measures according to measurement concepts to determine if the construct had acceptable psychometric characteristics (i.e., reliability) (Bringle et al., 2004; Cohen & Swererdlik, 2009). Thirty-eight articles reported on studies that used measures with psychometric evidence. Selective coding consisted of analyzing the data (i.e., 38 articles) according to scale construction and advanced psychometric theories, as also shown in Appendix A (Cohen & Swererdlik, 2009; Furr, 2010). Finally, seven articles with redundant measures, single items, or standardized measures not designed specifically to assess service-learning were removed from the final analysis.

Table 1
*Categories of Service-Learning and Civic Engagement Measures*

| Service-Learning Categories | *n* |
|---|---|
| Civic Attitudes | 4 |
| Civic Mindedness | 1 |
| Community Preferences | 4 |
| Course Evaluations | 5 |
| Community Impacts | 2 |
| Cultural Awareness (Intercultural, Diversity) | 8 |
| Education (Tutoring, Attitudes toward Learning) | 4 |
| Engagement | 5 |
| Integrity | 1 |
| Interpersonal Skills | 3 |
| Leadership | 3 |
| Motivation | 9 |
| Moral Development | 5 |

| | |
|---|---|
| Perceptions | 7 |
| Problem-Solving | 1 |
| Satisfaction | 10 |
| Self-Esteem (Self-Efficacy, Competence) | 8 |
| Service Impacts (Career) | 4 |
| Social Justice | 4 |
| Social Responsibility | 3 |
| Stereotypes (Racism, Ageism) | 5 |
| Student Learning | 10 |
| Tension | 1 |
| Volunteering (Helping) | 3 |
| | |
| *Population that the measure was intended* | |
| Students | 62 |
| Alumni | 2 |
| Faculty | 5 |
| Community Partners | 5 |
| Institutions (Administration) | 1 |

**Note. n* refers to the number of items, scales, and measures reported in the *MJCSL* from 1994-2015. Some measures assess more than one construct or are used with more than one group.

## Results

Twenty-four percent of the articles published in *MJCSL* between 1994 and 2015 included quantitative measures as data collection methods. Out of the 67 articles that met the inclusion criteria, 38 of them included psychometric evidence about the measure. Of these articles, 25 of them reported on measures designed to assess service-learning. As shown in Table 2, 21 measures designed to assess service impacts with supporting psychometric evidence derived from 18 articles represented the final set.

Table 2
*Measures Used to Assess Service-Learning and Civic Engagement*

| *Measure* | *Scale* | *Response Format* | *Psychometric Evidence Reported in the Article* |
|---|---|---|---|
| Scale of Service-Learning Involvement | Single-item<br>4 subscales | 5-point Likert scale | Modest to strong |
| Personal Social Values and Civic Attitudes | Single-item<br>2 subscales | 4-point, Likert-type scale; 5-point Likert scale | Modest to strong |
| Community Service self-Efficacy | Single-item | 10-point | Fair to strong |

| | | | |
|---|---|---|---|
| Scale | | rating | |
| CBO Supervisor Ratings of CbSL Student Performance Evaluations | Single-item | 5-point rating scale | Modest to strong |
| Community Service Involvement Preference Inventory | Single-item<br>4 subscales | 5-point Likert scale | Fair to modest |
| Student Learning from Community Service | Single-item (used with other measures) | Index (mean scores) | Fair |
| Feedback Disposition Instrument | Single-item (used with other measures) | 5-point rating scale | Modest |
| Goal Oriented Instrument | Single-item (used with other measures) | 9-point rating scale | Strong |
| Civic Attitudes and Skills Questionnaire | Multi-item<br>6 scales | 5-point Likert scale | Fair to modest |
| Course Satisfaction Measures | Multi-item<br>4 subscales | 5-point Likert scale | Modest to strong |
| Three Aspects of Engagement | Multi-item<br>4 subscales | 5-point Likert scale | Strong |
| Integrity (this scale was part of a larger questionnaire) | Single-item<br>2 subscales | 6-point Likert scale | Fair |
| Assessment of Community Agency Perceptions | Multi-item<br>4 subscales | 5-point Interview | Fair to Strong |
| The Community Service Preference Scale | Single-item<br>Two-scales | 5-point Likert Scale | Strong |
| The Quality of the Learning Environment | Single-item | 5-point rating Scale | Strong |
| Measure | Scale | Response Format | Psychometric Evidence Reported in the Article |
| Web-Based Faculty Service-Learning Beliefs Inventory | Multi-item<br>4 subscales | 5-point Likert Scale | Fair to strong |
| The Civic Minded Graduate Scale | Multi-item<br>10 subscales | 6-point Likert Scale | Strong |
| Civic Attitudes Measures<br>Valuing Community Engagement<br>Cultural Awareness<br>Seeks Knowledge about Political Social Issues<br>Knowledge of New Orleans | Multi-item<br>6 subscales | 5-point Likert Scale | Modest to strong |

| | | | |
|---|---|---|---|
| Culture and Issues<br>Knowledge of Current Events<br>Cultural Skills | | | |
| Community Engaged Competencies Scale | Single-item | 5-point rating scale | Strong |
| Service-Learning Course Quality | Single-Item | 5-point rating Scale | Modest to strong |
| The Community Impact Scale | Multi-item<br>8 subscales | 6 point rating scale;<br>Interview | Modest to strong |

Note. Psychometric evidence reflects internal consistency estimates. Weak = <.5, fair = .51<.69, modest = .7<.79, and strong = .80<.

## Measures of Service Impacts on Community

### *Assessment of Community Agency Perceptions Interview*

Developed by Miron and Moely (2006), this instrument measures community partner perceptions on four interview scales: (a) Agency Voice gathers information about the supervisor's involvement in planning service activities (i.e., to what extent do you feel that your agency and the University were equal partners); (b) Agency Benefit assesses the perceived benefit to the agency working with the service-learning program (i.e., how were the agency needs met by the student); (c) Interpersonal Relations assesses differences between the agency members and the students (i.e., do you feel the student enjoyed working with others of a different race, social class, or culture); and (d) Perception of the University identifies the supervisor's view of the University (i.e., is the University sensitive to the needs of the surrounding community). Cronbach alpha coefficients range from .66 to .77, and inter-correlations for the four scales range from .33 to .36.

### *Community Impact Scale*

Developed by Srinivas, Meenan, Drogin, and DePrince (2015) to measure the potential impact of school-community partnerships on community partners, this 46-item instrument yields scores on an eight subscales: (a) Overall Experiences (i.e., the community-school partnership was successful); (b) Social Capital (i.e., access to mentors and/or future employers); (c) Skills and Competencies (i.e., ability to work as part of a team); (d) Motivations and Commitments (i.e., commitment to engaging communities); (e) Personal Growth and Self-concept (i.e., compassion and caring for others); (f) Knowledge (i.e., knowledge about relevant social issues); (g) Organizational Operations (i.e., workload and demands on your time); and (i) Organizational Resources (i.e., finances). Cronbach alpha coefficients range from .70 to .94. An item analysis was conducted between survey and interview methods.

### *Community-Based Organizations Supervisor Rating Scale*

Developed by Ferrari and Worrall (2000), this nine-item survey assesses student fieldwork. Community supervisors evaluate students' fieldwork on two subscales, Service (i.e., appearance, site sensitivity) and Work Skills (i.e., attendance, punctuality). The Cronbach's alpha coefficient for Service Skills is .91 and for Work Skills is .87. Content validity was assessed between supervisors' written comments and survey ratings.

## Measures of Faculty Perceptions of Service

### *Community Engaged Scholarship (CES)*

Developed by Jameson, Jaeger, Clayton, and Bringle (2012), this 25-item self-report measure assesses faculty knowledge of and skills in conducing community-engaged scholarship (i.e., skills for fostering community and social change, ability to collaborate with community members in community capacity building endeavors). Cronbach alpha coefficients range from .90 to .95.

### *Web-Based Faculty Service-Learning Beliefs Inventory (wFSLBI)*

Developed by Hou (2010) to assess faculty perceptions of the benefits and barriers associated with community-based service-learning, this measure yields scores on four subscales: (a) PROS-CLS (Benefits Classroom) (i.e., service-learning enriches classroom discussions in my course); (b) PROS-COM (Benefit Community) (i.e., the service my students completed was beneficial to the community); (c) CONS-CLS (Barriers Classroom) (i.e., time constraints interfere with my ability to teach a service-learning course); and (d) CONS-INST (Barriers Institution) (i.e., faculty promotion and tenure policies do not support my service). Cronbach alpha coefficients range from .65 to .91. Confirmatory factory analysis, item discriminant validity, and group comparisons were also conducted to validate the measure.

## Measures of Service Impacts on Students

### *Civic Attitudes Measures*

Developed by Moely and Illustre (2011), this instrument assesses aspects of civic attitudes and engagement on six individual scales: (a) The Valuing of Community Engagement and Service (i.e., I enjoy engaging in community service); (b) Cultural Awareness (i.e., I think it is important for a person to think about his/her racial identity); (c) Seeks Knowledge about Political/Societal Issues (i.e., There is no point in paying attention to national politics); (d) Knowledge of New Orleans Culture and Issues (i.e., I am able to describe communities of color in the New Orleans area); (e) Knowledge of Current Events (i.e., I am well informed about current news events); and (f) Cultural Skills (i.e., I find it difficult to relate to people from a different race or culture). Cronbach alpha coefficients range from .77 to .90. A factor analysis was conducted with other measures of civic attitudes and responsibility.

### *Civic Attitudes and Skills Questionnaire (CASQ)*

Developed by Moely, Mercer, Ilustre, Miron, and McFarland (2002) to measure civic attitudes and skills, this 84-item self-report questionnaire yields scores on six scales: (a) Civic Action (i.e., future service intentions); (b) Interpersonal and Problem-solving Skills (i.e., work cooperatively with others to solve problems); (c) Political Awareness (i.e., awareness of local and national events); (d) Leadership Skills (i.e., ability to lead); (e) Social Justice Attitudes (i.e., attitudes toward poverty and social problems can be solved); and (f) Diversity Attitudes (i.e., attitudes toward diversity and their interests in interacting in culturally different people). Cronbach alpha coefficients range from .69 to .88 and test-retest reliabilities range from .56 to .81. A factor analysis and inter-correlations of the subscales were conducted. Construct validity was also conducted with the Modern Racism Scale (McConahay & Hough, 1976).

### *Civic Minded Graduate (CMG) Scale*

Developed by Steinberg, Hatcher, and Bringle (2011) to assess civic-mindedness (i.e., a disposition to be involved in the community or sense of social responsibility) in college graduates, this 30-item self-report measure yields score on 10 subscales: (a) Knowledge-Volunteer (i.e., I know there are a lot of opportunities to become involved in the community); (b) Knowledge-Academic (i.e., I am confident I can apply what I learned in my classes); (c) Knowledge-Social Issues (i.e., I am prepared to write a letter to the newspaper about a community problem); (d) Skills-Listening (i.e., I am a good listener); (e) Skills-Diversity (i.e., I prefer to work in settings in which I work with people who differ from me); (f) Skills-Consensus Building (i.e., Other students can describe me as a person who can discuss controversial issues); (g) Dispositions-Community Engagement (i.e., I like to be involved in community issues); (h) Dispositions-Self-Efficacy (i.e., I can contribute to improving life in my community); (i) Dispositions-Social Trustee of Knowledge (i.e., I want a career in which I can improve society); and (j) Behavioral Intentions (i.e., I plan to stay current with local and national news after I graduate). Cronbach alpha coefficients range from .85 to 97 and test-retest reliabilities range from .43 to .62. A principal component factory analysis and inter-item correlations were conducted to further assess reliability. Content validity was assessed with interviews and convergent validity was assessed with integrity and social desirability scales.

### *Community Service Self-Efficacy Scale (CSSES)*

Developed by Reeb, Katsuyama, Sammon, and Yoder (1998), this 10-item questionnaire assesses student efficacy for participating in community service (i.e., if I choose to participate in community service in the future, I will make a meaningful contribution). Cronbach's alpha coefficient is .92, inter-item correlations range from .65 to .78, and test-retest reliability is .62. Construct validity was assessed with the Social Responsibility Inventory (Bringle et al., 2004; Reeb et al., 1998).

### *Community Service Involvement Preference Inventory (CSIPI)*

Developed by Payne (2000) to assess student preference for becoming involved in community service, this 48-item inventory yields scores on four preferences: (a) Exploration Involvement

Preference (i.e., commitment to short term service that is convenient for the helper); (b) Affiliation Involvement Preference (i.e., commitment tends to be infrequent and shorter in duration); (c) Experimentation Involvement Preference (i.e., desire to make a difference in the lives of others and to learn more about the community); and (d) Assimilation Involvement Preference (i.e., career and lifestyle decisions based on service to be a responsible citizen). Cronbach alpha coefficients range from .63 to .74.

### *Community Service Preference Scale*

Developed by Moely, Furco, and Reed (2008) and adapted from items created by Moely and Miron (2005) to assess student preferences for typical service activities, this 16-item self-report questionnaire yields scores on four scales: (a) Charity Oriented Experience (i.e., helping those in need); (b) Social Change-Oriented Experience (i.e., changing public policy for the benefit of people); (c) Charity Orientation (i.e., a service placement where you can really become involved in helping others); and (d) Social Change Orientation (i.e., a service placement where you can contribute to social change that affects us all). Cronbach alpha coefficients range from .83 to .90. (Moely & Illustre, 2014).

### *Course Satisfaction Measures*

Developed by Moely, McFarland, Miron, Mercer, and Ilustre (2002), this instrument assesses student views of their courses on four subscales, including; (a) Course Value (i.e., how useful was the material covered in class); (b) Learning about Academic Field (i.e., application of the course concepts, interest in the field,); (c) Learning about the Community (i.e., working with others effectively and seeing social problems in a new way); and (d) Contribution to the Community (i.e., how useful were service activities). Cronbach alpha coefficients range from .74 to .82 (Moely, et al., 2002).

### *Items Used to Measure Integrity*

Developed by Bringle, Hatcher, and Mcintosh (2006), these items assess integrity components (i.e., when I am involved in service, I focus on meeting the immediate need) based on Morton's concept of integrity in students involved in community service. Factor analyses with varimax rotation were conducted on the 10 items. Seven of the 10 factors loaded onto two factors, identity and long-term commitments. The Cronbach alpha coefficient for integrity is .67 and for long-term commitments is .66.

### *Personal Social Values Scale*

Developed by Marby (1998), this nine-item questionnaire yields scores on two subscales: (a) Personal Social Values (i.e., helping others with difficulty); and (b) Civic Attitudes (i.e., adults should give some time for the good of their community). Cronbach alpha coefficients range from .63 to .81. A factor analysis was conducted with the Civic Attitudes subscale and academic benefit questions establishing a single-factor of civic attitudes (Bringle, et al., 2004; Marby, 1998).

### *Scale of Service-Learning Involvement (SSLI)*

Developed by Olney and Grande (1995), this 60-item questionnaire yields scores on three subscales: (a) Exploration (i.e., initial reasons for volunteering); (b) Realization (i.e., continual reasons for volunteering); and (c) Internalization (i.e., volunteering to help solve social problems). Cronbach alpha coefficients range from .70 to .84. Convergent and divergent validity was established with the Intellectual Development Scale and the Moral Development Measure, respectively.

### *Service-Learning Course Quality*

Developed by Moely and Ilustre (2014) and derived from an earlier version developed by Furco and Moely (2006) to assess student views of service-learning courses in terms of having attributes that are considered important. This 12-item self-report measure assesses service-learning attributes on three subscales: (a) Value of Service (i.e., I feel that my service-learning activity was worthwhile); (b) Focus on Service (i.e., the community organization in which I worked was ready to receive service-learning students); and (c) Opportunities for Reflection (i.e., I had opportunities to reflect on my service-learning through discussions with faculty, students and community members. Cronbach’s alpha coefficient for the 12-items is .93 and internal consistency for each scale ranges from .72 to .90.

### *Student Learning from Community Service Instrument*

Developed by Subramony (2000), this 10-item self-report measure assesses students’ perceived effectiveness in meeting service-learning goals. This measure is used in conjunction with the Feedback Disposition Instrument and the Goal Oriented Instrument. The Feedback Disposition Instrument measures student propensity to seek or avoid feedback, and the Goal Oriented Instrument measures students as learning or performance goal-oriented. Cronbach alpha coefficients range from .68 to .81.

### *Three Aspects of Engagement*

Developed by Gallini and Moely (2003), this 27-item self-report questionnaire yields scores on three scales: (a) Community Engagement (i.e., feeling connected to the community); (b) Academic Engagement (i.e., satisfaction with the academic course and university); and (c) Interpersonal Engagement (i.e., the course's influence on students’ ability to work with others effectively). Cronbach alpha coefficients range from .85 to .98.

### *Quality of Learning Environment*

Developed by Bringle, Hatcher, and Muthiah (2010), this 24-item self-report questionnaire measures student perceptions of learning environment components, including the extent to which students experience peer or faculty interaction, course satisfaction, perceived learning, active learning, and personal relevance components that contribute to high quality learning environments (i.e., I have developed a significant relationship with at least one student in this class). Items are added together to produce a composite index. The alpha coefficient for the composite index is .89.

## Discussion

Educators propose that service-learning has a positive effect on student learning (Eyler, & Giles, 1999). Pedagogical scholarship and research studies published in the *MJCSL* have evolved from course assessments to longitudinal, cross-sectional investigations. Data collection methodologies used to assess service-learning have also become more sophisticated over time, despite the fact that few researchers evaluate service impacts with quantitative measures. The instruments used to measure service-learning impacts such as social responsibility and civic engagement have also evolved from single items to multi-item scales.

The results from the current study suggest that civic engagement is directly and indirectly measured as a service-learning outcome. Very little research measured civic engagement directly. Civic engagement is typically measured with scales that were constructed to assess civic attitudes, social responsibility, and community engagement. Civic engagement is also measured with items designed to assess attributes or characteristics of responsible citizenship such as diversity awareness and social justice. Diversity was also cited as a student learning outcome. Student motivations for and perceptions of service scales were used to measure service impacts, but were cited less often compared to items that were used to measure student learning (i.e., questions that assess course objectives) and course satisfaction (i.e., student satisfaction with course and field).

Educators interested in measuring civic engagement as a student learning outcome should begin with constructing or refining existing course objectives. Faculty will need to decide what they want students to learn by the end of the course. One’s discipline will influence this decision making process. For example, a political science professor may want students to demonstrate responsible citizenship; while, a psychology professor may want them to demonstrate cultural competence. Then faculty will need to make sure that the course objectives are aligned with the course outcomes (i.e., student learning). Course objectives should also be explicit. In other words, if demonstrating cultural competence is an outcome then learning about cultural diversity should be an explicit course objective (i.e., compare and contract racial-ethnic-cultural identity models). Faculty will need to ensure that both the course content and service activities are aligned with course objectives and outcomes. For instance, writing a letter to an elected official may increase civic responsibility but not cultural competence. Once faculty align the course content, objectives and outcomes, they will need to focus on assessment of student learning. Table 2 lists service-learning and civic engagement measures. Faculty should decide which outcomes they want to measure and then review Table 2 to identify existing service-learning and civic engagement measures. Faculty should select service-learning or civic engagement measures that align with their course outcomes to demonstrate student learning. In addition, faculty may want to seek other measures that align with the course outcomes. Psychological measures may be helpful in assessing service impacts on student learning; therefore, faculty may want to conduct a search to obtain the particular measure(s). Faculty should develop a multi-item survey that includes service-learning, civic engagement and psychological measures that assess student learning. Faculty may also want to include course satisfaction measures in the survey. Course satisfaction measures are also included in Table 2. Finally, faculty may want to use a multi-method approach using surveys and reflections so they can identify changes in students before, during, and after service.

The measures used to assess service-learning impacts on student learning outcomes differed in terms of the constructs (i.e., civic attitudes), measurements (single-item questions vs. scales) and psychometric evidence (i.e., reliability, validity). While some researchers used standardized instruments or psychological measures, others developed single item or scales with and without such evidence. Very few scales designed to measure service learning had acceptable psychometric data. Out of the 21 service-learning measures that had supporting psychometric evidence, only nine of them were further assessed with advanced psychometric techniques (i.e., confirmatory factory analysis). Therefore, researchers and practitioners who use service-learning measures to evaluate pedagogical impacts may want to also conduct psychometric analyses on the specific measure to control for measurement error and to be assured that they are measuring what they intended (see Lehan & Hussey's (2015) primer on scale development validation in this e-book). Additional service-learning measures should also be constructed and validated using advanced measurement theories and psychometric perspectives.

The current study expands the knowledge-based on measures used to assess service-learning and civic engagement. This is one of the first studies to systematically analyze service-learning and civic engagement measures. Although this study only analyzed quantitative measures, qualitative methods of inquiry are complimentary and should be examined. Qualitative and quantitative data collection methods should also be used to measure service impacts, but only when civic engagement is an explicit course objective. Greater alignment between course objectives and outcomes may be the first step in measuring civic engagement as a service-learning outcome.

# References

References marked with an asterisk indicate a scale.

American Psychological Association. Retrieved July 25, 2015 from http://www.apa.org/education/undergrad/civic-engagement.aspx

Association for Experiential Education. (2011). How it works in higher education. Retrieved October 2, 2011 from https://web.archive.org/web/20110113103425/http://www.aee.org/applications/highered/

Battistoni, R. M. (2002). What is good citizenship? Conceptual frameworks across disciplines. *Introduction to service-learning toolkit, 2nd edition* (pp.179-186). Providence, RI: Campus Compact.

Boyle-Baise. M. (2002). *Multicultural service learning*. New York: Teachers College Press.

Boyle-Baise, M., & Kilbane, J. (2000). What really happens? A look inside service-learning for multicultural teacher education. *Michigan Journal of Community Service Learning, 7*, 54-64.

*Bringle, R.G., Hatcher, J.A., & Muthiah, R.N. (2010). The role of service-learning on the retention of first-year students to second year. *Michigan Journal of Community Service Learning, 16*(2), 38-49.

*Bringle, R.G., Hatcher, J.A., & McIntosh, R.E. (2006). Analyzing Morton's typology of service paradigms and integrity. *Michigan Journal of Community Service Learning, 13* (1), 5-15.

Bringle, R.G., Phillips, M.A., & Hudson, M. (2004). *The Measure of service-learning*. Washington, DC: American Psychological Press.

Cohen. R. J., & Swerdlik, J. (2009). *Psychological testing and assessment: An introduction to tests and measurement, 7th edition*. New York: McGraw-Hill.

Creswell, J. W. (2005). *Educational research, 2nd edition*. Upper Saddle River, NJ: Pearson Prentice Hall.

Creswell, J. W. (1998). *Qualitative inquiry and research design choosing among five traditions*. Thousand Oaks, CA: Sage.

Creswell, J. W. (1994). *Research Design: Qualitative & Quantitative Approaches*. Thousand Oaks, CA: Sage.

Eyler, J. (2009). The power of experiential education. *Liberal Education*, *Fall*, 24-31.

Eyler, J. S. (2000). What do we most need to know about the impact of service-learning on student learning? *Michigan Journal of Community Service Learning, Special Issue*, 11-17.

Eyler, J. S., & Giles, D. E. (1999). *Where's the learning in service-learning?* San Francisco: Jossey-Bass.

*Ferrari, J. R., & Worrall, L. (2000). Assessments by community agencies: How "the other side" sees service-learning. *Michigan Journal of Community Service Learning, 7*, 35-40.

Furr, R. M. (2010). *Scale contraction and psychometric for social and personality psychology*. Los Angeles: Sage.

Furco, A. & Moely, B. E. (2006, April). *A comparative analysis of the impacts of service-learning on students*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA

*Gallini, S. M., & Moely, B. E. (2003). Service-learning and engagement, academic challenge, and retention. *Michigan Journal of Community Service Learning, 10* (1), 1- 14.

Giles, D. E., & Eyler, J. (2013). The endless quest for scholarly respectability in service-learning research. *Michigan Journal of Community Service Learning, 20* (1), *53-64.*

Green, A. E. (2001). "But you aren't white:" Racial perspectives and service learning. *Michigan Journal of Community Service Learning, 8* (1), 18-26.

Hatcher, J. A. (2010). Defining the catchphrase: understanding civic engagement of college students. *Michigan Journal of Community Service Learning, 16* (2), 95-100.

*Hou, S-I. (2010). Developing a faculty inventory measuring perceived service-learning benefits and barriers. *Michigan Journal of Community Service Learning, 16* (2), 78-89.

Howard, J. (Ed) (n.d.). *Michigan Journal of Community Service Learning,* Retrieved June 6, 2015 from http://www.unich.ed/~mjcsl/index.html

*Jameson, J. K., Clayton, P.H., Jaeger, A.J., & Bringle, R. G. (2012). Investigating faculty learning in the context of community-engaged scholarship. *Michigan Journal of Community Service Learning, 18* (2), 40-55.

Hussey, H. D., & Lehan, T. J. (2015). A primer on scale development. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Litke, R. A. (2002). Do all students "get it?"" Comparing students' reflections to course perfomance. *Michigan Journal of Community Service Learning, 8* (2),27-34.

*Mabry. J. B. (1998). Pedagogical variations in service-learning and student outcomes. How time, contact, and reflection matter. *Michigan Journal of Community Service Learning, 5,* 32-47.

Marxen, C. E. (2003). *Meeting NCATE standard through service-learning: Diversity. American Association of College for Teacher Education*, *Issue Brief 3*, Winter, 1-4.

McConhay, & Hough, J. Jr. (1976). Symbolic racism. *Journal of Social Issues, 32* (2), 23-45. doi: 10.1111/j.1540-4560.1976.tb02493.x

*Miron, D., & Moely, B. E. (2006). Community agency voice and benefit in service learning. *Michigan Journal of Community Service, 12(2), 27*-37

*Moely, B. E., & Ilustre, V. (2014). The impact of service-learning course characteristics on university students' learning outcomes. *Michigan Journal of Community Service Learning, 21*(1), 5-16.

*Moely, B. E., & Ilustre, V. (2011). University students' views of public service graduation requirement. *Michigan Journal of Community Service Learning, 17* (2), 43-58.

*Moely, B. E., Furco, A., & Reed, J. (2008). Charity and social change: The impact of individual preferences on service-learning outcomes. *Michigan Journal of Community Service Learning, 15* (1), 37-48.

*Moely, B. E., McFarland, M. Miron, D., Mercer, D., & Illustre, V (2002). Changes in college students' attitudes and intentions for civic involvement as a function of service-learning experiences. *Michigan Journal of Community Service Learning, 9*(1), 18-26.

*Moely, B. E., Mercer, S. H., Ilustre, V., Miron, D., & McFarland, M. (2002). Psychometric properties and correlates of the civic attitudes and skills questionnaire (CASQ): A measure of student's attitudes related to service-learning. *Michigan Journal of Community Service Learning, 8*(2), 15-26.

Moely, B. E., & Miron, D. (2005). College students' preferred approaches to community service: Charity and social change paradigms. In S. Root, J. Callahan, and S.H. Billig (Eds.), *Improving service-learning practice: Research on models to enhance impacts*. Greenwich: CT: Information Age Publishing.

Moore, D. T. (2000). The relationship between experimental learning research and service-learning research. *Michigan Journal of Community Service Learning, Special Issue*, 124-128.

National Service-Learning Clearinghouse. Retried June 6, 2015 from https://gsn.nylc.org/clearinghouse#sthash.1RCNI3Kh.dpuf

*Olney, C., & Grande, S. Validation of a scale to measure development of social responsibility. *Michigan Journal of Community Service Learning, 2,* 43-53.

*Payne, C. A. (2000). Changes in involvement preferences as measured by the community service involvement preference inventory. *Michigan Journal of Community Service Learning, 7*, 41-53.

Quaye, S. J., & Harper, S. R. (2007). Faculty accountability for culturally inclusive pedagogy and curricula. *Liberal Education*, 32-39.

*Reeb, R. N., Katsuyama, R.M., Sammon, J. A., & Yoder, D. S. (1998). The community service self-efficacy scale: Evidence of reliability, construct validity, and pragmatic utility. *Michigan Journal of Community Service Learning, 5,* 48-57.

Reinke, S. J. (2003). Making a difference: Does service -learning promote civic engagement in MPA students? *Journal of Public Affairs Education, 9* (2), 129-137.

Root, S., Callahan, J., & Sepanski, J. (2002). Building teaching dispositions and service-learning practice: A multi-site study. *Michigan Journal of Community Service Learning, 8* (2), 50-59.

Rockquemore, K. A., & Schaffer, R. H. (2000). Toward a theory of engagement: A cognitive mapping of service learning. *Michigan Journal of Community Service Learning, 7,* 14-23.

Simons, L., Fehr, L., Blank, N., Connell, H., Fernandez, D., Georganas, D., Padro, J., & Peterson, V. (2012). Lessons learned from experiential learning: What do students learn from a practicum and internship? *International Journal of Teaching and Learning in Higher Education*, *24* (3), 325-334.

*Srinivas, T., Meenan, C. E., Drogin, E., & DePrince, A. P. (2015). Development of the community impact scale measuring community organization perceptions of partnership benefits and costs. *Michigan Journal of Community Service Learning, 21* (2), 5-21.

*Steinberg, K. S., Hatcher, J. A., & Bringle, R. G. (2011). Civic-minded graduate: A north star. *Michigan Journal of Community Service Learning, 18* (1), 19-33.

Strage, A. A. (2000). Service-learning: Enhancing learning outcomes in a college-level lecture course. *Michigan Journal of Community Service Learning, 7*, 5-13.

*Subramony, M. The relationship between performance feedback and service-learning. *Michigan Journal of Community Service Learning, 7,* 46-53.

Vogelgesang, L. J., & Astin, A. W. (2000). Comparing the effects of community service and service learning. *Michigan Journal of Community Service Learning, 7*, 25-34.

Wang, O. T., Kelly, J. M., & Hritsuk, B. (2003). Does one size fit all?: Challenge of social cognitive development. *Michigan Journal of Community Service Learning, 9 (2),* 5-14.

Wang, Y. & Jackson, G. (2005). Forms and dimensions of civic involvement. *Michigan Journal of Community Service Learning, 11 (2),* 39-48.

## Appendix A: Research Measures for Measuring Service-Learning and Civic Engagement

| *Article Author(s)* | *Publication Date* | *Description* |
|---|---|---|
| Hammond, C. | 1994 | -24 items that measure faculty motivation and satisfaction on three subscales: 1. Personal; 2. Co-curricula, and 3. Curricular motivations |
| Miller, J. | 1994 | -11 items that inquires about student opinions of placement site, personal experience, and academic experience related to service |
| Hesser, G. | 1995 | -Faculty rate the impact of service on 11 Liberal Arts Learning outcomes (i.e., written communication skills) |
| Olney, C. & Grande, S. | 1995 | -[1],[3]The Scale of Service-Learning is a 60-item scale that measures developmental phases of the service-learning model |
| | | -[2]The Defining Issues Test is an objective measure of moral reasoning as defined by Kohlberg's theory |
| | | -[2]The Scale of Intellectual Development is a 102-item that measures intellectual and developmental stages of Dualism, Relativism, and Commitment |
| | | -[2]Measures of Moral Orientation assesses two orientations toward moral decision making as described by Gilligan |
| Greene, D. & Diehm, G. | 1995 | -6 items that measure the degree to which students stereotype older adult nursing home residents |
| Myers-Lipton, S. | 1995 | -[2]The Modern Racism Scale measures racism based on the theory of modern or symbolic racism |
| Kendrick, J. R. | 1996 | -15 items that measure social responsibility and personal efficacy;<br>-9 items that measure values or beliefs from helping others |
| Hudson, W. E. | 1996 | -18 items that measure public policy beliefs; 9 items that measure course impact |
| Eyler, J., Giles, D. E., & Braxton, J. | 1996 | -[2]The National Study of Service-Learning Impacts on Students includes citizenship confidence scales, citizenship values, citizenship scales, and perceptions of social justice. This survey was developed as part of the Measuring Citizenship Project at Rutgers University |
| Miller, J. | 1997 | -Two questions that measure students' perception of power |
| Wade, R.C., & Yarbrough, D. B. | 1997 | -Preceptor teacher survey measures their level of satisfaction with the student teacher and the University;<br>-Student teacher survey measures his satisfaction with the preceptor teacher and service experience |
| Osborne, R. E., Hammerich, S., Hensley, C. | 1997 | -[2]A survey of measures used to assess self-esteem in students. These measures included the Rosenberg Self-esteem scale, the Cognitive Complexity Scale, the Texas Social Behavior Inventory, the Spontaneous Self-Concept measure, and the Remote Associations Test |
| Mabry, B. | 1998 | -[1],[3]22 items that measure personal social values, civic attitudes, perceived course impact on civic attitudes, and perceived academic benefit of service-learning |

| | | |
|---|---|---|
| Reeb, R. N., Katsuyama, R. M., Sammon, J.A., Yoder, D. S. | 1998 | -[1,3]The Community Service Self-efficacy Scale is a 10-item self-report measures that assesses students' confidence in engaging in service;<br>-[2]The Social Responsibility Inventory is a 15-item measure assesses student responsibility for solving social problems |
| Korfmacher, K.S. | 1999 | -10 items that measure alumni satisfaction with the Environmental Science Services Program at Brown University |
| Vernon, A., Ward, K. | 1999 | -A survey of community partners to measure their satisfaction with the program |
| Rockquemore, K.A., Schaffer, R. H. | 2000 | -26 items that measure service-impacts on service-learners |
| Veogelgesang, L. J., Astin, A.W. | 2000 | -Items measured the degree of commitment to promoting racial understanding and activism, academic skills, leadership, and future service were drawn from the longitudinal College Student Survey |
| Ferrari, J. R., Worrall, L. | 2000 | -[1,3]Community based organization supervisor rating of community based service-learininers |
| Payne, C.A. | 2000 | -[1]The description of the Community Service Involvement Preference Inventory (CSIPI); A survey that measures students' level of engagement in service on four subscales |
| Subramony, M. | 2000 | -[1]The Student Learning from Community Service Instrument is a 10-item measure that assesses service-learners' perceived effectiveness in meeting service goals;<br>-[1]Feedback Disposition measures student propensity to seek and avoid feedback;<br>-[1]Goal Oriented Instrument measures the degree to which students are learning or performance goal oriented;<br>-[2]Agency Feedback Instrument measures the amount of feedback that was given to students |
| Jeffers, C. A. | 2000 | -34 items measuring student attitudes toward art and learning of art in the classroom and galleries |
| Abes, E. S., Jackson, G., Jones, S. R. | 2002 | -A survey that measures factors that motivate and deter faculty use of service-learning |
| Moely, B. E., McFarland, M., Miron, D., Mercer, S. H., Ilustre, V. | 2002 | -The Civic Attitudes and Skills Questionnaire (CASQ) measures student civic attitudes and skills on six subscales;<br>-[1]Course satisfaction measures student satisfaction with the course and community on 4 subscales. |
| Moely, B. E., Mercer, S. H., Ilustre, V., Miron, D., McFarland, M. | 2002 | -[1,3]The Civic Attitudes and Skills Questionnaire (CASQ) measures student civic attitudes and skills on six subscales. Psychometric evidence is provided. |
| *Schmidt, A., | 2002 | -[1]71-items measuring tutoring satisfaction; |

| | | |
|---|---|---|
| Robby, M.A. | | -CASQ<br>-SAT/9 Scores<br>-[1]Child and teach evaluations |
| Prins, E. S. | 2002 | -A survey measuring community colleges' purpose, service initiatives, and reasons for engaging in service-learning |
| *Root, S., Callahan, J., Sepanski, J. | 2002 | -[2]Questions measuring teacher efficacy, commitment to teaching, attitudes toward diversity, service ethic, and commitment to future service;<br>-[1]Aspects of service-learning experience scale measures student views of service |
| Sperling, R., Wang, V.O., Kelly, J. M., Hritsuk, B. | 2003 | -An attributional questionnaire measuring students' reason for engaging in service and views of the American Educational system in relation to social inequity |
| *Evangelopous, N., Sidorova, A., Riolli, L. | 2003 | -[1]A survey were used to measure student attitudes and views of usefulness of business statistics. |
| Gallini, S. M., Moely, B. E. | 2003 | -[1]The Three Aspects of Engagement measures community, academic, and interpersonal engagement;<br>-[1]Items measuring academic challenge and retention in students |
| Marchel, C. A. | 2003 | -[2]Items measuring sociocultural awareness, career plans and future service |
| Brody, S.M., & Wright, S. C. | 2004 | -[2]Volunteer Function Inventory;<br>-[2]Self-Expansion Questionnaire |
| Hatcher, J.A., Bringle, R. G., Muthiah, R. | 2004 | -[1]A questionnaire containing items measuring active learning, course satisfaction, faculty interaction, peer interaction, perceived learning, and personal relevance, qualities of service-learning class, academic content, and reflection<br>-[1]The Quality of the Learning Environment measures student views of the academic content and the service context |
| Fensel, L.M., & Peyrot, M. | 2005 | -[2]A survey measuring the quality of service as an undergraduate, current community service, service-related job, and personal responsibility; Items from the Higher Education Research Institute College Student Survey; Social and Personal Responsibility Scale in alumni at community colleges |
| Wang, Y., Jackson, G. | 2006 | -[2]The Student Service-Learning Course Survey measurers student perceptions of civic involvement, social justice and charitable reasons for civic involvement and dimensions of Civic Involvement |
| Bainger, N., Bartholomew, K. | 2006 | -13 items measured organizations' motives to engage in service-learning and their satisfaction with service-learners |
| Miron, D., Moely, B.E. | 2006 | -[1]An interview measuring community partner perceptions on four subscales |
| Bringle, R. G., Hatcher, J.A., | 2006 | -[1]A questionnaire containing items measuring intrinsic and extrinsic motives, leadership, preference for different types of community |

| | | |
|---|---|---|
| McIntosh, R.E. | | service, and integrity;<br>-[1,3]9-items were used to measure integrity. |
| Schnaubelt, T., Stratham, A. | 2007 | -Five items assessed faculty perceptions of service as important at the institution and considered in the tenure and promotion process |
| Banerjee, M., Hausafas, C.O. | 2007 | -A survey measuring faculty perceptions of service-learning as a teaching strategy |
| Moely, B.E., Furco, A., Reed, J. | 2008 | -[1]The Community Service Preference Scale measures students' preference for service that emphasizes charity or social change activities;<br>-[2]HES-LS questionnaire measuring civic attitude, responsibility and career development;<br>-CASQ<br>-Subscales from the Three Aspects of Engagement Scale |
| Bernacki, M., L., Jaeger, E. | 2008 | -[2]The Defining Issues Test measures moral reasoning according to Kohlberg;<br>-[2]The Moral Justification Scale is used to measure moral orientation;<br>-[2]The Service-Learning Outcome Scale measures student perceptions of how their coursework, understanding social problems, problem solving ability, leadership, efficacy, and passion. |
| Clayton, P.H., Bringle, R.G., Senor, B., Huq, J., Morrison, M. | 2010 | -The Transformational Relationship Evaluation Scale assesses the relationship between community partners and University researchers |
| Bringle, R. G., Hatcher, J.A., McIntosh, R.E. | 2010 | -[1]Quality of the Learning Environment is 24-item measuring course satisfaction, faculty interaction, peer interaction, perceived learning, active learning, and personal relevance |
| Barney, S.T., Corseer, G.C., White, L.H. | 2010 | -[2]The Community Attitudes to Mental Illness Scale is a 40-item measure assessing attitudes towards those with a mental illness; |
| Hou, S-I. | 2010 | -[1,3]The Web-based Faculty Service-Learning Beliefs Inventory assesses faculty views of the benefits from and barriers to implementing service on four subscales |
| Seider, S.C., Gillmor, S.C., Rabinowicz, S.A. | 2010 | -[2]The Protestant Ethic Measure student beliefs that individual hard work leads to success. |
| Bowman, N., Brandenberger, J.W., Mick, C.S., Smedley, C.T. | 2010 | -[2]Situation Attributions for Poverty Scale measures beliefs about poverty;<br>-[2]Openness to Diversity items measure the degree to which students are interested in learning about those who differ from them;<br>-[2]Responsibility for Improving Society assess how much personal responsibility one perceives for taking action to help others;<br>-[2]Empowerment View of Helping measures beliefs about whether people can overcome their problems with assistance from others; |

| | | |
|---|---|---|
| | | -[2]Belief in a Just World measures the beliefs that good things happen to good people;<br>-[2]Social Dominance Orientation measures preferences for and acceptance of inequality across social groups;<br>-[2]Self-generating View of Helping measures the beliefs that individuals are only able to help themselves overcome their problems |
| Steinberg, K. S., Hatcher, J. A., Bringle, R. G. | 2011 | -[1,3]The Civic Minded Graduate Scale is a 30-item self-report measuring civic-mindedness on 10 subscales |
| Moely, B. E., Ilustre, V. | 2011 | -[1,3]The Civic Attitude Measures assessed student attitudes, knowledge and skills for community engagement;<br>-CASQ |
| Mills, S. D. | 2012 | -The Tension Survey asks students and community partners to indicate the extent to which they have experienced any of the briefly described tension |
| Jameson, J.K., Jaeger, A. J., Clayton, P.H., Bringle, R. G. | 2012 | -[1]The Community Engaged Scholarship Competencies Scale is designed to assess new faculty views of and comfort with implementing community engaged scholarship |
| Neihaus, E., Crain, L. K. | 2013 | -[2]The National Survey of Alternative Break measures service engagement, community engagement, community/staff interaction, student learning, emotional and physical challenged, social issues, reflection, journaling and orientation |
| Moely, B. E., Ilustre, V. | 2013 | -[1]The Service-Learning Course Quality Indicators measures characteristics of high quality service-learning;<br>-CASQ; Civic Attitudes, Knowledge and Skills |
| Soria, K.M., Thomas-Card, T. | 2014 | -[2]The Student Experience in the Research University Survey contains 600 items that measure student satisfaction, academic engagement, campus climate, research experiences, and civic/community engagement |

|  |  |  |
|---|---|---|
| *Moely, B. E., Ilustre, V | 2014 | -The Service-Learning Course Quality Indicators;<br>-Preferences for Charity- and Social Change-oriented Service;<br>-Course Satisfaction Measures: Learning about the Community, Academic Learning, Satisfaction with the University. |
| De Leon, N. | 2014 | -[2]The Cultural Quotient Scale is a 20-item measure that assesses cultural intelligence, knowledge, strategy, action and drive;<br>-[2]The Intercultural sensitivity Scales is a 24-item measure that assesses intercultural communication and competence. |
| Srinivas, T., Meenan, C.E., Drogin, E., DePrince, A. P. | 2015 | -[1]The Community Impact Scale measures the impact of service on community partners. |
| *Russell-Stamp, M. | 2015 | -Web-based Faculty Service-Learning Beliefs Inventory measures faculty perceptions of the benefits and barriers associated with service. |
| *Mitchell, T.D., Richard, F.D., Battistoni, R.M., Rost-Banik, C. Netz, R. Zakoske, C. | 2015 | -The Civic Minded Professional scale measures the degree to which alumni are involved in service-related work or careers. |

Note. [1]Denotes scales with psychometric evidence that were specifically designed to measure service impacts, [2]denotes scales with psychometric evidence that were not specifically designed to measure service impacts, [3]denotes scales with advanced psychometric data, and *indicates either articles with redundant measures or single item measures (i.e., one question) that were excluded from the final analysis.

# Chapter 10: Measuring Individual Differences in Epistemological Beliefs

Kelly L. Y. Ku

Hong Kong Baptist University

Is knowledge non-changing and absolute, or is it ever evolving and tentative? Students in the same classroom may hold different views of knowledge. Such views may affect how they accept, evaluate, and interpret what is being learned. Understanding how learners view knowledge is now fundamental to education, because there has been a rising body of evidence on the important role of mature epistemological beliefs towards academic and intellectual performances (see for example, Inglis & Mejia-Ramos, 2009; Ku, Lai & Hau, 2014; Kuhn & Park 2005; Qian & Alvermann, 1995; Rodríguez & Cano, 2006; Schommer-Aikins & Hutter, 2002).

Derived and subsumed under the broad philosophical study of *epistemology* (the study of what knowledge is), the term *epistemological belief* is used to describe an individual's representation of the nature of knowledge. Since the original work of Perry (1970), different teams of researchers have put forth somewhat different emphases in conceptualizing and measuring epistemological beliefs. Two major approaches have emerged; these include the developmental approach represented by the work of Kuhn's (see Kuhn & Park, 2005; Kuhn, 1991) and King and Kitchener's (1994), as well as the multi-facet model as signified by Schommer's (1990, 2004) work.

This chapter begins with an outline on the theoretical frameworks of epistemological beliefs. It goes on to review a list of instruments grounded in the discussed frameworks and closes with issues and suggestions in examining epistemological beliefs.

## Theoretical Conceptions of Epistemological Beliefs

Epistemology refers to the broad philosophical theory of knowledge. Psychologists and educators adopt an empirical approach in studying *personal* epistemology on the other hand, aimed at assessing individuals' subjective understanding of what knowledge is like. It is generally held that individuals with a more mature understanding of knowledge tend to view knowledge as independently constructed and complex in its structure, and realize that conclusions to any problem might be tentative in nature. These representations are called epistemological beliefs, or epistemic beliefs.

Existing research on epistemological beliefs is mostly derived from the longitudinal studies of Perry conducted in the 70s. It was observed that college students moved through a sequence of epistemic development typically from taking a dualistic (e.g. only one view can be correct) to a relative (e.g. both views can be correct) stance of knowledge (Perry, 1970). This process indicated an intellectual growth from a simple and absolute understanding of knowledge to an integrative one where conflicting views are allowed and valued. It was believed that students eventually grew to develop a commitment to use evidence in justifying alternative points of view and thus understood that not all views carry equal strength.

The construct of epistemological beliefs is not a straightforward one. Over the years, to continue the work of Perry's, researchers have adopted two major approaches. These include examining the developmental sequence of epistemological beliefs as well as identifying a system of independent beliefs about knowledge that an individual might hold simultaneously.

## The Developmental Approach of Epistemological Beliefs

Kuhn (Kuhn & Park, 2005; Kuhn, 1991) defined a learner's epistemological beliefs as falling into one of four levels: *realist*, *absolutist*, *multiplist*, and *evaluativist* (see Table 1). In the earliest level of realist, children attain knowledge through direct "copying" of the reality. For example, teachers may observe a child of age 3 believing everyone experiences the same event in the same way. If the child enjoyed a birthday party, he or she would believe everyone else enjoyed the party. While children begin to observe the existence of different views in the absolutist stage, they resolve conflicts of understanding through simple assertion of a correct view over incorrect ones. Children at this age often argue about things they disagree on in an either-or-logic(e.g., "I am right, therefore you must be wrong"). Adolescents, on the other hand, believe everyone has the right to hold different opinions. This characterizes the third level, thus all opinions are equally correct and critical thinking is unnecessary. In the final level, the mature learner recognizes that some opinions are better supported by evidence and are therefore more correct than others. At this stage, the learner recognizes the importance of critical thinking. Children's development of the theory of mind underlies the core of the development of epistemological understanding (Kuhn & Park, 2005). That is to say, children's representation of knowledge advances as they begin to realize how they and others come to know and reason differently. Kuhn's model therefore describes a generic sequential developmental trend associating age and levels of epistemological understanding.

Table 1

*Development of Epistemological Understanding*

| Level | Age Group | Concept of Knowledge |
|---|---|---|
| Realist | Children before age 4 | Knowledge is a replica of external reality. |
| Absolutist | Late childhood to pre-adolescence | Knowledge is discrete pieces of fact that are either correct or incorrect replica of reality. |
| Multiplist | Adolescence | Knowledge is individual's self-chosen opinions, which are competing claims without discrimination. |
| Evaluativist | Adulthood | Knowledge is claims with different strengths determined by reasons and evidence. |

Sharing similar features with Kuhn's developmental model, but distinctive in its stronger emphasis on the means of justification used by learners, King and Kitchener (1994) proposed the *Reflective Judgment Model* (see Table 2). The model comprises three levels: *pre-reflective*, *quasi-reflective*, and *reflective*. Learners at early stages dictate that justification of an opinion is either unnecessary or is referenced solely to an external agent of authority. At later stages, the strength of various types of evidence (e.g., anecdotes, scientific, historical, etc.) and criteria (e.g., meaningfulness, utility, consistence, etc.) are used to validate an opinion in the process of justification. Unlike Kuhn's model, King and Kitchener (1994) put an emphasis on epistemological growth of adolescence to adulthood in their model. Though without clear age specification, it is believed that starting the period of adolescence allows learners to grasp the importance to use evidence in supporting their beliefs about the world, and that the various means of justification are acquired inclinations that can be cultivated through learning and experience. Thus, it is loosely assumed for those with higher education backgrounds to operate at more advanced epistemological stages.

Table 2

*Stage of Reflective Judgment in Epistemological Understanding*

| Stage | Concept of Knowledge | Role of Justification |
|---|---|---|
| Pre-reflective | Knowledge is first viewed as based on personal experience, and later as obtained from external authorities. | Justification is either unnecessary or is made with reference to an authority. |
| Quasi-reflective | Any personal claim supported by different reasons can be knowledge; it is a matter of a person's own perspective. | Relying first on subjective or idiosyncratic reasons then to using a variety of evidence |
| Reflective | Knowledge is constructed and supported by reasons and evidence, with an aim of arriving at a well-informed and complete understanding of an issue. | Different perspectives and sources of evidence are compared and evaluated for consistency, meaningfulness and coherence. |

## The Multiple-Facet Approach of Epistemological Beliefs

Epistemological beliefs are also conceptualized as a set of personal theories about various facets of knowledge. Beginning with Schommer's (1990) work, five facets were identified. These include *simple knowledge* (knowledge as isolated versus interrelated), *certain knowledge* (knowledge as static versus evolving), *omniscient authority* (the source of knowledge as passed down from higher agent versus the source of knowledge as challengeable), *quick learning* (learning takes place quickly or not at all versus learning takes place gradually), and *innate ability* (intellectual ability as fixed entity versus ability as acquired). The simplicity and certainty facets pertain to beliefs about structure and form of knowledge; whereas, omniscient authority and quick learning concern the origin of knowledge and process of knowing.

Based on Schommer's (1990) work, Hofer (2000) later proposed a refined four-factor model of epistemological beliefs. They include *certainty of knowledge*, *simplicity of knowledge*, *source of knowledge* and *justification for knowing* (justification based on personal opinion, authority, or the use of multiple evidences). In this refined version, Hofer integrated King and Kitchener's (1994) emphasis of justification into Schommer's original model and put forth the factor *justification for knowing,* which examines a person's active inclination in using reasons and evidence in knowledge construction.

## Measuring Epistemological Beliefs

Existing instruments provide measurements for assessing development of beliefs, generic beliefs, and discipline- or issue-specific beliefs about knowledge.

### Measuring Development of Epistemological Beliefs

#### *Epistemological Development*

Kuhn, Cheney, and Weinstock's (2000) instrument consists of 15 forced-choice items representing five judgment domains (see Table 3). Each item describes a pair of contrasting claims. The respondent needs to decide if one claim is right (i.e., a response of the absolutist level) or whether both could be right (i.e., a response of the multiplist level). If the latter is chosen, the learner has to decide whether one claim could be more right than the other (i.e., a response of the evaluativist level).

Table 3

*Instrument of Epistemological Development (Kuhn et al., 2000)*

| Judgment Domain | Sample Item |
|---|---|
| Aesthetic | *Robin thinks the first painting they look at is better.*<br>*Chris thinks the second painting they look at is better.* |
| Value | *Robin thinks lying is wrong.*<br>*Chris thinks lying is permissible in certain situations.* |
| Personal Taste | *Robin says warm summer days are nicest.*<br>*Chris says cool autumn days are nicest.* |
| Social World | *Robin has one view of why criminals keep going back to crime.*<br>*Chris has a different view of why criminals keep going back to crime.* |
| Physical World | *Robin believes one book's explanation of what atoms are made up of.*<br>*Chris believes another book's explanation of what atoms are made up of.* |

The scoring essentially includes counting the dominating response conforming at the absolutist, multiplist, or evaluativist level for each judgment domain. It was reported that more people have demonstrated multiplist thoughts in the domains of personal taste and aesthetic domains, and evaluatist thoughts in the domains of social and physical worlds (Kuhn et al., 2000). Convergent validity was supported by studies comparing results of this instrument with other similar ones with 70% - 80% reported compatibility in terms of identified epistemological levels (Kuhn et al., 2000). This is a straight-forward instrument that is easy to understand and administer. It can be administered in a group setting with an estimated completion time of 15 minutes. It has been found suitable for 5th – 12th graders as well as adult respondents (Kuhn &

Park, 2005). The merit of the instrument also includes profiling the pattern of a person's epistemological beliefs across domains using a more economical method compared to the traditional approach of interviewing the participant. The data depicted by this model are particularly useful in addressing the theoretical question of whether personal epistemology is domain-generic or specific. The limitation of this instrument includes the relatively limited statistical analysis that can be done with the forced-choice data generated. Ahola (2009) discussed the potential of this instrument and how qualitative components can be integrated to better capture individuals' reasoning underlying their choices, allowing more sophisticated analysis.

### *Reasoning About Current Issues Test (RCI)*

The RCI (King and Kitchener, 1994; see Wood, Kitchener, & Jensen, 2002 for description and format of the test) is a test created based on the Reflective Judgment Model. The test asks test-takers' opinions about controversial authentic issues (e.g., artificial sweeteners, federal debt, global warming). Test-takers are asked to rate, on a five-point continuum from very dissimilar to very similar, how alike his or her opinion is on each issue compared to ten provided statements. The statements are claims about justifications that match different stages of the Reflective Judgment Model, such as "Researchers disagree because they are really studying different facets of the issue and the best ways to address one facet of the issue are different than the best ways to address other facets" and "Researchers arrive at different conclusions because the evidence itself is complex and they examine it from several perspectives. They arrive at a decision by synthesizing their knowledge, experiences, and expert opinions." In the last section, test-takers  select up to three statements that best match their own thinking about the issue. The RCI score is based solely on the choices made in this section, whereas the rating serves only to probe test-takers' personal judgments on particular issues (King & Kitchener, 2004). In other words, the scoring focuses not on assessing factual knowledge or cognitive skills in making reflective judgments, but the assertions that a person holds about how judgments are made. The results therefore reveal more about epistemic inclination than intellectual ability.

The RCI can be administered on a computer (see http://www.reflectivejudgment.org/index.cfm) and in paper-and-pencil format. The reported Cronbach's alphas of the RCI ranged from .50s to .70s (Duell & Schommer-Aikins, 2001; King & Kitchener, 2004). A cross-sectional analysis of over 9,000 students of undergraduate, graduate, and professional education programs reported that the RCI is able to discriminate between students across schooling after controlling for co-variants (Wood et al., 2002), confirming a progressive developmental trend as predicted by the Reflective Judgment Model.

### Measuring Multiple Epistemological Beliefs

Instruments using variations of Schommer's (1990) model are listed. All three of them are self-report questionnaires in paper-and-pencil format utilizing five-point Likert-scales (except for Hofer's, 2004, which uses a 7-point Likert-scale), with the degree of agreement measured reflecting sophistication of an individual's epistemological beliefs.

### *Epistemological Questionnaire (EQ)*

The 62-item EQ (Schommer, 1990) is constructed based on Schommer's five-facet model of epistemological beliefs. Four validation studies utilizing different population samples with different educational and work experiences were conducted; the reported internal consistency reliability ranged from .50s to .70s (Schommer, 1993; 1998). Group comparisons revealed sensitivity of the EQ to age, gender, educational level, and cultural differences. Adults and female students were more likely to believe in ability as acquired as opposed to predisposed (Schommer, 1993; 1998). Asian students were found to be extremely consistent in agreeing that authorities are not to be criticized as opposed to those of the West (Chan & Elliot, 2000).

The EQ has reported a somewhat mixed factor structure despite its popularity. The proposed five-factor structure has not been consistently replicated in later studies (Braten & Stromso, 2005; Clarebout, Elen, Luyten, & Bamps, 2001; Duell & Schommer-Aikins 2001; Schraw, Bendixen, & Dunkle, 2002; Wood & Kardash, 2002). Researchers tend to conclude that the low to modest internal consistency might be caused by the rather ambiguous nature of the construct of personal epistemology itself. Additionally, the inconsistency of respondents' ratings of some items could reflect that respondents hold contrasting beliefs that (e.g. *the only thing that is certain is uncertainty itself*) do not lend themselves easily to empirical research (Qian & Alverman, 1995).

### *Epistemological Belief Inventory (EBI)*

The EBI (Schraw et al., 2002) is a refined version of the EQ in an attempt to enhance its psychometric properties. This is a shorter instrument with a more stable factor structure generating up to 60% of variance (Schraw et al., 2012). It consists of 32 items and yields five factors with reliabilities that typically range from .50 to .65.

The original 62-tiem EQ and the EBI are amongst the most popular multi-facet instruments of personal epistemology; however, both of them are not free from measurement problems. Yet in comparison, the EBI is slightly more favorable as the authors have made an effort in this refined 32-item version to remove redundant items, and rephrase ambiguously worded items. Such effort lead to improved clarity in the comprehension of the items, an enhanced structural stability, as well as reliabilities set within an acceptable range.

### *Discipline-Focused Epistemological Beliefs Questionnaire (DFEBQ)*

The FEDBQ (Hofer, 2000) is a domain-specific questionnaire with items similar to those of Schommer's (1990) EQ. It contains 18 items, assessing four factors corresponding to Hofer's revised model of epistemological beliefs, using a 7-point Likert scale. Respondents are asked to focus on a specific academic discipline (e.g., Science or Psychology) as they rate each item.

Similar factor structures were produced across disciplines. The results that the science-focused version reported less mature beliefs than those of the psychology-focused version seem to support domain specificity to some extent.

## Measurement Issues

Empirical research generally supports the positive relationship between sophisticated epistemological beliefs and a number of desired learning outcomes. For instance, researchers have found correlations between sophisticated epistemological beliefs and two-sided reasoning about controversial social issues (Kardash & Scholes, 1996; Mateos et al., 2011; Schommer-Aikins & Hutter, 2002), positive changes in scientific conception (Sinatra, Southerland, McConaughy & Demastes, 2003), and academic achievement (Rodríguez & Cano, 2006).

Self-report instruments measuring multiple epistemological beliefs are more widely used in empirical studies for their amenability to relatively more complex statistical analyses (Schraw, 2012), as compared to those measuring the development of epistemological understanding. In particular, the popularity of the EBI has been growing in the past years because of its domain generic nature and its advantage of measuring all five facets of the original Schommer's model using only half of the EQ's items. Although measurement problems have been persistent in empirical epistemological research, multi-facet questionnaires still contribute to the field because of their strength for allowing an otherwise fuzzy and complex construct to be broken down into agreeable facets. Examination of interrelationships among these facets and their unique and combined variances in predicting teaching and learning variables is also made possible. For instance, Chan and colleagues (2011) found that even after controlling for cognitive ability, individuals who hold the belief that knowledge is certain and absolute performed more poorly in two-sided thinking tasks than those who recognized that knowledge is changeable. This finding supported the unique contribution of epistemological beliefs to reasoning beyond a person's overall intellectual capacity.

Another issue to take note of is that instruments using domain generic beliefs might not be adequate in capturing epistemic attitudes that are specific to a particular context or discipline. Likewise, those measuring domain specific beliefs suffer from limited generalization of findings. Given the lack of a unified conceptualization of personal epistemology and psychometric inconsistencies in factor structure across studies with different instruments, researchers should take caution with result interpretation.

## Conclusion

In recent years, alternate theories have been put forth that are worth taking note of for future studies. In particular, the field has called for more consideration when interpreting what constitutes sophisticated and naïve believers using generic measurements of beliefs about knowledge. For instance, researchers such as Elby and Hammer (2001) have proposed the epistemological resources theory that challenged the simple either-or classification of a learner's belief as naïve or sophisticated. They hold that the correct sophisticated understanding of the discipline science as changing and tentative might not be productive when it comes to understanding some specific and fundamental scientific phenomenon, such as that living organisms evolve. This theory was found to be supported with some preliminary empirical evidence in a study of Ku and colleagues (2014), in which sophistication of epistemological beliefs were made salient in predicting students' argumentation thinking only under an experimental condition where the students were prompted to consider the ambiguous information. Likewise, Bromme, Pieschl and Stahl (2010) revealed in their research that a person's sophisticated beliefs were activated in "reflective" task contexts and not in others. In other words, a learner might approach a task in a "naïve" manner if that enhances effectiveness in task completion, despite his or her ability to adopt a more sophisticated approach. In addition, Hofer (2001) has added metacognitive processes in the multi-facet model by emphasizing beliefs about knowledge as a function of metacognitive knowledge, and metacognitive monitoring and judgments. More empirical testing of these models will enrich general understanding of personal epistemology.

The chapter discussed the developmental and multi-facet approach in theorizing epistemological understanding. The scope, strength, and limitation of self-reports instruments measuring beliefs about knowledge, as well as objective assessments of epistemological development were reviewed. The previous section also highlighted what is not adequately captured by existing measurements. More attention should be paid to the interplay of beliefs about knowledge and other contextual factors in future studies. Lastly, it is suggested to consider using more than one form of instrument to triangulate results, and when applicable, more sophisticated statistical analyses to enhance reliability of findings.

# References

References marked with an asterisk indicate a scale.

Ahola, S. (2009). Measurement issues in studying personal epistemology. *Psychology and Society, 2*(2), 184-191.

Bråten, I., & Strømsø, H. I. (2004). Epistemological beliefs and implicit theories of intelligence as predictors of achievement goals. *Contemporary Educational Psychology*, *29*(4), 371-388. doi:10.1016/j.cedpsych.2003.10.001

Bromme, R., Pieschl, S., & Stahl, E. (2010). Epistemological beliefs are standards for adaptive learning: A functional theory about epistemological beliefs and metacognition. *Metacognition and Learning, 5*(1), 7-26. doi:10.1007/s11409-009-9053-5

Chan, K.W. & Elliott, R.G. (2000). Exploratory study of epistemological beliefs of Hong Kong teacher education students: Resolving conceptual and empirical issues. *Asia-Pacific Journal of Teacher Education, 28*, 225-234. doi:10.1080/713650691

Clarebout, G., Elen, J., Luyten, L., & Bamps, H. (2001). Assessing epistemological beliefs: Schommer's Questionnaire revisited. *Educational Research and Evaluation*, *7*(1), 53–77. doi:10.1076/edre.7.1.53.6927

Duell, O. K. & Schommer-Aikins, M. (2001). Measures of people's beliefs about knowledge and learning. *Educational Psychology Review, 13*(4), 419-449. doi:10.1023/A:1011969931594

Elby, A. & Hammer, D. (2001). On the substance of a sophisticated epistemology. *Science Education*, *85* (5), 554-567. doi:10.1002/sce.1023

*Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Psychology, 25*, 353-405. doi:10.1006/ceps.1999.1026

Hofer, B. K. (2001). Personal epistemology research: Implications for learning and teaching. *Educational Psychology Review, 13*(4), 353-383. doi:10.1023/A:1011965830686

Inglis, M., & Mejia-Ramos, J. P. (2009). The effect of authority on the persuasiveness of mathematical arguments. *Cognition and Instruction*, *27*(1), 25-50. doi:10.1080/07370000802584513

Kardash, C. M., & Scholes, R. J. (1996). Effects of preexisting beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology, 88*(2), 260-271. doi:10.1037/0022-0663.88.2.260

*King, P. M. & Kitchener, K. S. (1994). *The development of Reflective Judgment in adolescence and adulthood*. San Francisco, USA: Jossey Bass.

King, P.M., & Kitchener, K.S. (2004). Reflective judgment: Theory and research on the development of epistemic assumptions through adulthood. *Educational Psychologist, 39*(1), 5-18. doi:10.1207/s15326985ep3901_2

Ku, K. Y. L., Lai, C. M., & Hau, K. T. (2014). Epistemological beliefs and the effect of
authority on argument-counterargument integration: An experiment. *Thinking Skills and Creativity, 13*, 67-79. doi:10.1016/j.tsc.2014.03.004

Kuhn, D. (1991). *The skills of argument*. Cambridge, UK: Cambridge University Press.

Kuhn, D., & Park, S.H. (2005). Epistemological understanding and the development of
intellectual values. *International Journal of Educational Research, 43*, 111-124. doi:10.1016/j.ijer.2006.05.003

Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development, 15*, 309-328. doi:10.1016/j.ijer.2006.05.003

Mateos, M., Cuevas, I., Martı´n, E, Martı´n, A., Echeita, G., & Luna, M. (2011). Reading to write an argumentation: The role of epistemological, reading and writing beliefs. *Journal of Research in Reading, 34*(3), 2011, 281-297. doi:10.1111/j.1467-9817.2010.01437.x

Perry, W. G., Jr. (1970). *Forms of intellectual and ethical development in the college years: A scheme*. New York: Holt, Rinehart, and Winston.

Qian, G., & Alvermann, D. (1995). Role of epistemological beliefs and learned helplessness in secondary school students' learning science concepts from text. *Journal of Educational Psychology, 87*(2), 282-292. doi:10.1037/0022-0663.87.2.282

Rodríguez, L. & Cano, F. (2006). The epistemological beliefs, learning approaches and study orchestrations of university students. *Studies in Higher Education, 31*, 617–636. doi:10.1080/03075070600923442

*Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*(3), 498-504. doi:10.1037/0022-0663.82.3.498

Schommer, M. A. (1993). Epistemological development and academic performance among secondary schools. *Journal of Educational Psychology, 85*(3), 406-411. doi:10.1037/0022-0663.85.3.406

Schommer, M. A. (1998). The influence of age and schooling on epistemological beliefs. *British Journal of Educational Psychology, 68*, 551-562. doi:10.1111/j.2044-8279.1998.tb01311.x

Schommer-Aikins, M. (2004). Explaining the epistemological belief system: Introducing the embedded systemic model and coordinated research approach. *Educational Psychologist*, *39*(1), 19-29. doi:10.1207/s15326985ep3901_3

Schommer-Aikins, M., & Hutter, R. (2002). Epistemological beliefs and thinking about everyday controversial issues. *The Journal of Psychology, 136*(1), 5–20. doi:10.1080/00223980209604134

*Schraw, G., Bendixen, L. D., & Dunkle, M. E. (2002). Development and validation of the Epistemic Belief Inventory (EBI). In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 261-275). Mahwah, NJ: Lawrence Erlbaum.

Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching, 40*(5), 510-528. doi:10.1002/tea.10087

*Wood, P.K., Kitchener, K.S., & Jensen, L. (2002). Considerations in the design and evaluation of a paper-and-pencil measure of reflective thinking. In B. Hofer and P. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 277-294). Mahwah, NJ: Lawrence Erlbaum.

Wood, P., & Kardash, C. (2002). Critical elements in the design and analysis of studies of epistemology. In B. K. Hofer, & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge* (pp. 231-261). Mahwah, NJ: Lawrence Erlbaum.

# Chapter 11: Measuring Well-Being in the Scholarship of Teaching and Learning

Kristin Layous[1], S. Katherine Nelson[2], and Angela M. Legg[3]

[1]California State University, East Bay, [2]Sewanee: The University of the South, [3]Pace University

Students' achievement and learning do not simply reflect their latent abilities or their conscientiousness. As instructors, we want to know the X-factor that could enhance students' learning experience, as well as the negative factors that could hinder it. We propose that students' psychological well-being—including positive aspects like having a global sense that one's life is good and negative aspects like crippling stress and anxiety—is an important factor in understanding students' experience in the classroom, including their learning, growth, motivation, and ultimate grade. For example, subjective well-being—one's global life satisfaction, frequency of positive emotions, and infrequency of negative emotions (Diener, Suh, Lucas, & Smith, 1999)—is predictive of success in multiple life domains, including relationships, health, and work (Lyubomirsky, King, & Diener, 2005). This chapter will provide instructors with an overview of scales to assess different aspects of well-being and illustrate how instructors can incorporate these measures in their scholarship of teaching and learning (SoTL).

## Positive Aspects of Well-Being

Some researchers subset positive aspects of well-being into hedonic and eudaimonic components, with the former being the "subjective well-being" construct described earlier (e.g., life satisfaction, positive and negative emotions) or just plain old "happiness," and the latter being the degree to which one has a sense of meaning or purpose in life (e.g., Ryff, 1989). In practicality, hedonic and eudaimonic well-being are highly overlapping (Kashdan, Biswas-Diener, & King, 2008), but scales exist to measure the conceptually distinct constructs and we include a variety of options here. Notably, although various aspects of well-being have been linked with academically related outcomes among college students (see below), research on well-being and student performance, engagement, motivation, and learning is not as prevalent as might be expected, and could be a ripe area for future research.

### Life Satisfaction

The Satisfaction With Life Scale (SWLS; Diener, Emmons, Larsen, & Griffin, 1985), the most widely used life satisfaction measure, assesses respondents' current satisfaction and has been linked with academic achievement among college students (Borrello, 2005; Lepp, Barkley, & Karpinsky, 2014). The SWLS consists of five questions (e.g., "In most ways my life is close to my ideal," "I am satisfied with my life"), which are rated on 7-point Likert-type scales (1 = *strongly disagree*, 7 = *strongly agree*). Validation studies have shown that the SWLS comprises a single factor and possesses high internal reliability (Cronbach's α = .87) and high test-retest reliability (*r* = .82; Diener et al., 1985). Recent evidence suggests that just using one item, "In general, how satisfied are you with your life?," had similar patterns with other related variables as using the rest of the scale, and it alone might be sufficient to capture a quick picture of life satisfaction (4-item scale from 1 = *very satisfied* to 4 = *very dissatisfied*, reverse-scored; Cheung, & Lucas, 2014; also see Hussey & Lehan's (2015) chapter of this e-book for additional information on scale validation).

One of the simplest ways for instructors to incorporate well-being scales into classroom research is to administer trait versions at the beginning of a course and then use them to predict course grades or other outcomes throughout the course (e.g., paper grades, attendance). In one such study, researchers administered the SWLS on the first day of an introductory psychology course and found that life satisfaction was positively related to students' final grades (Borrello, 2005). Surprisingly little research has taken this approach, so SoTL researchers could contribute greatly to the literature on well-being and academic achievement. In addition, exposing students to actual scales could be a valuable learning tool as they can see first-hand what it means when a research article says that participants reported their life satisfaction.

## Positive and Negative Emotions

In addition to knowing how students feel about their lives in general, instructors might also want to know how positively or negatively students have felt over the past week, past few days, today, or even "right now." Indeed, in accordance with the broaden-and-build theory (Fredrickson, 2001), positive emotions at one time point predicted school-related personal resources (i.e., a composite of academic self-efficacy, and study-related hope and optimism) at a subsequent time point among university students (Ouweneel, Le Blanc, & Schaufeli, 2011). To assess positive and negative emotions, we recommend the Affect-Adjective Scale (AAS; Diener & Emmons, 1984; Cronbach's α = .89 and .84, respectively), the Modified Differential Emotions Scale (mDES; Fredrickson, Tugade, Waugh, & Larkin, 2003; α = .79 and α = .69, respectively), or the Positive and Negative Affect Schedule; (PANAS; Watson, Clark, & Tellegen, 1988, αs > .84). The question stems of any of the three scales can all be altered to assess the time in the course of interest (e.g., "right now" vs. "past week"). Each scale assesses people's experience of positive and negative emotions over a specified time period, but each has unique attributes that might make it better or worse for your purposes.

For example, the AAS and PANAS are largely composed of high-arousal emotions (e.g., happy, joyful, distressed, irritable), whereas the mDES includes both low and high arousal emotions (e.g., interested, alert, curious). Both the AAS and the PANAS assess the extent to which people have felt a certain way (*not at all* to *very much* or *extremely*), whereas the mDES asks the frequency with which people have felt a certain way (*never, hardly, some of the time, often,* and *most of the time*). The former may assess intensity of emotions rather than frequency of occurrence, whereas the latter may miss the intensity of the emotion, but capture frequency. Both the AAS and the PANAS assess one emotion at a time, whereas a limitation of the mDES is that it lists three similar emotions at once (e.g., amused, fun-loving, and silly appear together), thus introducing confusion through triple-barreled questions (e.g., Gehlbach, 2015). The number of items for each measure differ widely, with the AAS including 9 items, the mDES 23 items, and the PANAS 44 items, and the measures also differ on number of scale anchors, with the PANAS and the mDES both using 5-point scales and the Affect-Adjective using a 7-point scale.

Instructors could take the pulse of the class by asking students how they feel “right now” before an important class activity or exam or they could ask how students have felt during the “past few weeks” or “past week” when they are about to turn in an important and laborious paper. In either instance, instructors could explore whether positive or negative emotions relate to performance. In addition, SoTL researchers can use novel methodology to explore positive and negative emotions and their relationship to college achievement. For example, recent innovations in positive and negative emotion research include exploring the within-person variability in the experience of emotions. First, researchers can assess emodiversity, the degree to which a person experiences a variety of emotions throughout the week (or past few days, or past month, etc.; see http://www.emodiversity.org for an equation; Quoidbach et al., 2014). Second, researchers can explore the standard deviation of positive or negative emotions throughout multiple measurement periods throughout the week to assess the degree to which people experience fluctuations in extremity of positive and negative emotions (e.g., high levels of positive emotions on one day, and then possibly none at all the next day; Gruber, Kogan, Quoidbach, & Mauss, 2013). These relatively new ways of evaluating positive and negative are a ground-breaking way of exploring emotions and college achievement.

### Subjective Well-Being Composite

Researchers commonly assess overall well-being by averaging participants’ life satisfaction and frequency of positive and (reverse-scored) negative emotions to represent the theoretical tripartite structure of well-being (Diener et al., 1999). If the three constructs are measured on different Likert scales (e.g., 1-5 vs. 1-7), researchers can transform scores to z-scores for each scale and average the standardized scores. Additionally, although life satisfaction and positive and negative emotions are typically highly correlated (Pavot, Diener, Colvin, & Sandvik, 1991), researchers should explore the correlations in their own data before combining. Although researchers have proposed more complicated ways of combining the three constructs (Busseri & Sadava, 2010), averaging them is the most common approach.

In addition to subjective well-being predicting course outcomes, SoTL researchers could also use subjective well-being and the other similar constructs as dependent variables. Research has now demonstrated a host of simple and brief activities that boost well-being (i.e., positive activities; Lyubomirsky & Layous, 2013; Sin & Lyubomirsky, 2009) and reduce the negative effects of threat on academic achievement (i.e., self-affirmation and belonging interventions; Cohen & Sherman, 2014; Walton & Cohen, 2011). Many of these activities take about 10-15 minutes and could easily be administered within a class session or online as homework. Furthermore, instructors could also test the effects of intervention-induced changes in well-being on subsequent exams or class assignments by administering the intervention, measuring well-being as a manipulation check and mediator, and then assessing performance as a behavioral outcome.

## Happiness

Rather than represent happiness with the subjective well-being composite, some researchers recommend to simply ask people if they are happy to tap whatever that person thinks it means to be happy. One such face valid measure is the 4-item Subjective Happiness Scale (SHS; Lyubomirsky & Lepper, 1999), which asks participants to consider how generally happy they are, how happy they are relative to their peers, (1 = *less happy*, 7 = *more happy*), and the extent to which a description of a "very happy" and a "very unhappy" person, respectively, characterizes them (1 = *not at all*, 7 = *a great deal*; Cronbach's α's > .79). Although researchers should take caution before dropping items from a validated scale, recent research suggests that negatively worded and reverse-scored items contribute to poor reliability (Gehlbach, 2015) and recent research on the SHS in specific suggests that dropping the fourth item improves scale reliability (O'Connor, Crawford, & Holder, 2014; also see Wilson-Doenges, 2015 for more discussion of this issue).

Among high school students in Hong Kong, student scores on the SHS were related to their perceptions of their school's quality and their own grades (Kashdan & Yuen, 2007). Similarly, like the SWLS, the SHS taken at the beginning of an introductory psychology course was also related to final grades (Borrello, 2005). In addition to exploring how well-being relates to final grades, researchers could also explore whether well-being relates to exam performance or possibly just relates to final grades due to perseverance on homework assignments and class participation. Additionally, researchers could measures stress or anxiety (see Negative Aspects of Well-Being in this chapter) to explore whether well-being simply buffers the negative effects of stress on academic performance or is uniquely related.

## Eudaimonic Well-Being

The Questionnaire for Eudaimonic Well-Being (QEWB; Waterman et al., 2011) was developed to assess people's well-being derived from the development and fulfillment of their individual potential (as opposed to just assessing how people generally feel about their lives, regardless of the source, like the life satisfaction and happiness scales). The scale includes 21 items (e.g., "My life is centered around a set of core beliefs that give meaning to my life"), which are rated on a scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*). Moreover, the scale demonstrates strong reliability, with Cronbach's α = .86. The QEWB also includes questions addressing personal identity and growth. Indeed, one study found that well-being (measured with the QEWB) significantly mediated the association between identity development (e.g., commitment making, identification with commitment) and internalizing symptoms, externalizing symptoms, and health-risk behaviors (Ritchie et al., 2013). Accordingly, instructors may wish to administer this scale in their classrooms in the context of discussing identity development in emerging adulthood.

### Meaning in Life

The Meaning in Life Questionnaire (MLQ; Steger, Frazier, Oishi, & Kaler, 2006) is one of the most widely used scales assessing life meaning. It includes 10 items to assess the presence of (e.g., "My life has a clear sense of purpose") and search for (e.g., "I am always looking for something that makes my life feel meaningful"; 1 = *absolutely untrue,* 7 = *absolutely true*) meaning in life. Both subscales demonstrated strong reliability (Cronbach's αs = .86 and .87, respectively), as well as convergent and discriminant validity. Demonstrating the applicability across a wide range of demographics, both subscales were largely unrelated to age, gender, race, and religion (Steger et al., 2006).  Instructors could contrast the MLQ and the QEWB with the SWLS and SHS to illustrate the various ways in which researchers assess well-being (see also the next few scales).

If the focus of the course is on well-being, instructors could do a more elaborate study of how these types of courses affect student well-being. For example, one study measured students' happiness, life satisfaction, self-actualization, hope, and search for and presence of meaning in life at the beginning and end of a semester-long course on positive psychology (Maybury, 2013). Throughout the course, students completed course activities regarding their personal values and character strengths, gratitude, and optimism. Over the course of the semester, students reported gains in hope, self-actualization, life satisfaction, happiness, and search for meaning, but not presence of meaning. Accordingly, these class results may offer instructors the opportunity to discuss what makes for a happy life and the difference between *searching* for meaning in life and *feeling* that life is already meaningful.

### Psychological Well-Being

In contrast to the singular focus on meaning in life, the Scales of Psychological Well-Being (PWB; Ryff, 1989) conceptualizes well-being as including multiple facets. The PWB was originally developed as a 120-item instrument encompassing 6 subscales representing each facet (20 items per scale): self-acceptance (e.g., "I like most aspects of my personality"), positive relations with others (e.g., "I know that I can trust my friends, and they know they can trust me"), autonomy (e.g., "I am not afraid to voice my opinions, even when they are in opposition to the opinions of most people"), environmental mastery (e.g., "In general, I feel I am in charge of the situation in which I live"), purpose in life (e.g., "I have a sense of direction and purpose in life"), and personal growth (e.g., "I think it is important to have new experiences that challenge how you think about yourself and the world"). The length of this scale makes it cumbersome for use within the classroom, but instructors could present a subset of items to contrast the PWB with the aforementioned scales. Additionally, this scale is included in the Midlife in the United States study (MIDUS), a national longitudinal study of health and well-being, so interested students could request use of this data for an independent research project on how well-being relates to a host of physical health and demographic variables (http://www.midus.wisc.edu/).

## Psychological Flourishing

The Mental Health Continuum (MHC; Keyes, 2002) was developed as a tool for measuring mental health as a state of flourishing, rather than merely the absence of disease. Accordingly, this measure includes three subscales: emotional well-being (6 items; e.g., "How much of the time in the last 30 days did you feel full of life?"; 1 = *none of the time,* 5 = *all;* Cronbach's α = .91), psychological well-being (18 items adapted from Ryff, 1989; e.g., "I like most parts of my personality"; 1 = *disagree strongly,* 7 = *agree strongly;* α = .81), and social well-being (15 items; e.g., "I feel close to other people in my community"*;* 1 = *disagree strongly,* 7 = *agree strongly*; α = .81). A 14-item short-form of the MHC (Lamers, Westerhof, Bohlmeijer, ten Klooster, & Keyes, 2011) was recently developed, including shortened versions of each of the three subscales, which also demonstrated strong reliability (αs > .74 for the three subscales, α = .89 for the total MHC-SF). To diagnose mental health, averages for the three subscales are calculated, and those who score in the upper tertile are considered to be flourishing and those who score in the lower tertile are considered to be languishing.

## Psychological Need Satisfaction

Based on self-determination theory (Ryan & Deci, 2000), the Balanced Measure of Psychological Needs (BMPN; Sheldon & Hilpert, 2012) is an 18-item scale that measures people's feelings of autonomy (e.g., "I was free to do things my own way"), competence (e.g., "I took on and mastered hard challenges"), and connectedness (e.g., "I felt close and connected to other people who are important to me"). Each item is rated on a scale ranging from 1 (*no agreement*) to 5 (*much agreement*). Each subscale demonstrates strong reliability, Cronbach's αs > .78.

The BMPN offers a number of advantages for classroom use. Students can complete this relatively short scale quickly, leaving plenty of time for discussion and other activities in class. In addition, the items reflect general feelings rather than domain-specific satisfaction, rendering the scale applicable to students across diverse experiences. Finally, the scale demonstrates strong reliability and validity as three independent scales or as a single scale representing overall psychological need satisfaction, which provides instructors with a variety of ways to use this scale in the classroom. For example, an instructor could administer only the connectedness subscale when discussing relationships in a course on social psychology but could administer the entire scale when discussing self-determination theory in a course on motivation.

## Optimism

Optimism is the global belief that good things will happen (i.e., generalized outcome expectancies; Scheier & Carver, 1985). The revised life orientation test (LOT-R; Scheier, Carver, & Bridges, 1994) is a 6-item measure designed to assess trait optimism. Respondents are asked to indicate their degree of general agreement "over the past year" with statements such as "I'm always optimistic about my future," using a 5-point response scale (1 = *strongly disagree*, 5 = *strongly agree*; Cronbach's α > .78). To measure state optimism (Kluemper, Little, & DeGroot, 2009), change the question stem to "over the past week" and re-word individual items to indicate current state of optimism with statements such as "Currently, I'm optimistic about my future," using a 5-point scale (1 = *strongly disagre*e, 5 = *strongly agree*). College students high in trait optimism were less likely to see their education as a source of stress in their lives (Krypel & Henderson-King, 2010), more likely to expect success (Robbins, Spence, & Clark, 1991), but were no more or less likely to achieve good grades (close to zero correlation; Robbins et al., 1991; see also Rand, Martin, & Shea, 2011).

### Hope

Distinct from optimism, hope typically relates to positive feelings about specific goals and planning to meet goals rather than generalized expectancies (Snyder et al., 2002). The 12-item Trait Hope Scale (THS; Snyder et al., 1991) includes two subscales to measure hope via agency (i.e., goal directed determination; e.g., "I energetically pursue my goals) and pathway (i.e., planning of ways to meet goals; e.g., "I can think of many ways to get the things in life that are important to me"). Each subscale includes four items, which are rated on a scale from 1 (*definitely false*) to 8 (*definitely true*). The scale demonstrated good reliability in a college student sample (agency subscale: Cronbach's αs > .71, pathways subscale: αs > .63, total scale: αs > .74). Hope positively predicts academic success in college (Snyder et al., 2002), and hope, but not optimism, positively correlates with grades for first-year law students, controlling for admissions test scores and undergraduate grades (Rand et al., 2011). Instructors may want to highlight the differences between hope and optimism and engage students in a class discussion about why hope, but not optimism, is related to academic achievement.

## Negative Aspects of Well-Being

Negative aspects of well-being can range from feelings of sadness, tension, low self-efficacy, and learned helplessness. Although many measures of clinical symptomology exist to assess students' depressive symptoms or anxiety disorder criteria, for the purposes of this chapter, we chose to focus on non-clinical measures of stress and anxiety that generalize well to students. General feelings of stress and anxiety harm people's health (DeLongis, Folkman, & Lazarus, 1988) and, unsurprisingly, also negatively influence students' academic well-being and performance (Richardson, Abraham, & Bond, 2012). Self-report measures of student stress fall into two categories: (a) measures of general stress and anxiety, and (b) domain-specific stress and anxiety.

### General Stress and Anxiety

SOTL researchers and professors may wish to assess students' general levels of stress to gauge whether they predict poorer academic performance, difficulty adjusting to college life, or other negative outcomes. The following discussion provides resources to assess these overarching levels of stress.

#### *College Students' General Stress*

The Social Readjustment Rating Scale (Holmes & Rahe, 1967) is a commonly used measure of life stressors but, for student populations we recommend the College Undergraduate Stress

Scale (CUSS; Renner & Mackin, 1998). The 51 items on the CUSS list major negative life stressors, such as being raped (or accused of rape), experiencing the death of a friend or family member, contracting an STD, and financial difficulties. The scale also includes less severe stressors, such as beginning a new academic semester and concerns over physical appearance. To calculate a total stress score, students add up the numerical stress values next to each item they experienced within the past year. For example, being raped carries a stress value of 100 and falling asleep in class carries a value of 40. The highest possible stress score is 3,623 and the sample from Renner and Mackin's (1998) initial validation study reported an average stress score of 1,247 (*SD* = 441). Although this scale has been underutilized in the SoTL literature, one paper establishes this scale as a useful tool for teaching aspects of research methodology and data collection (Thieman, Clary, Olson, Dauner, & Ring, 2009).

A discussion of general stress levels can coincide with course material from health psychology, biopsychosocial approaches to health, and the interplay between mental states and physical outcomes. The scale also provides opportunities to discuss positive sources of stress such as getting married, getting straight A's, or attending an athletic event. Students will likely enjoy taking this easy-to-score assessment and can consider how their unique stressors affect their college experience. For future SoTL research, this scale provides a relatively short measure to gather data on a broad range of stressors. SoTL researchers may question whether academic-related stressors (e.g., finals week) pose the same levels of harm to the academic experience than non-academic-related stressors (e.g., serious illness in a close friend or family member).

### *State-Trait Measures of Stress*

The CUSS approaches stress from a life events perspective, but some instructors may prefer to assess the affective or cognitive components of stress and anxiety. One of the most common measures for anxiety is the State-Trait Anxiety Inventory (STAI; Spielberger, 1983), a 20-item measure of general and transient, state anxiety. Although the STAI is considered one of the gold standards in measuring anxiety, one potential disadvantage is that its use is fee-based through the psychological assessment company Mind Garden Inc. Some alternatives to the STAI full form are the 6-item state anxiety scale (STAI-6; Marteau and Bekker, 1992) and the Perceived Stress Scale (PSS; Cohen, Kamarck, & Mermelstein, 1983). The STAI-6 lists six anxiety-related emotions such as “tense” and “worried.” Participants complete the scale by answering the extent to which they currently feel those emotions on a 4-point scale (0 = *not at all* to 4 = *very much*). The PSS consists of 14 items assessed on a 5-point scale (0 = *never* to 4 = *very often*). These items ask participants to rate the extent to which they felt or experienced anxiety-provoking stressors (e.g., "In the last month, how often have you found that you could not cope with all the things you had to do?"). Both the STAI-6 and PSS are internally reliable (STAI-6 Cronbach’s α = .82; PSS α = .85).

These state-trait assessments of stress give instructors an excellent way to illustrate the difference between personality tendencies and momentary, transient states. For more advanced discussion, state-trait assessments can illustrate the difference between moderators which tend to be personality variables and mediators which can manifest in state variables.

## Domain-Specific Stress and Anxiety

A number of measures are geared toward assessing domain-specific anxiety. For instructors seeking to enhance the specificity of their research questions, these scales offer an excellent solution. We discuss the domain-specific scales to measure test anxiety, math anxiety, computer anxiety, and social anxiety.

### *Test Anxiety*

For instructors examining test anxiety, the 21 true/false-item Test Anxiety Scale (Cronbach's alphas range between .68-.81; Sarason, 1984) and the 10-item Worry-Emotionality Questionnaire (Liebert & Morris, 1967) are viable options that both assess two factors presumed to underlie test anxiety: cognitive thoughts of worry, and affective or physiological emotionality. For those researchers needing a state measure of test anxiety (both factors), we recommend the 8-item State Test Anxiety scale (Hong & Karstensson, 2002).

Arguing that the cognitive component of test anxiety most strongly predicts deficits in performance, Cassady and Johnson (2002) developed a reliable (Cronbach's α = .86) 27-item scale to assess cognitive worry. The Cognitive Test Anxiety scale includes items to assess intrusive, ruminative thoughts during test-taking and engaging in social comparison or test-irrelevant thinking during test-taking. Participants respond using a 4-point scale to sample items such as, "During tests, I find myself thinking of the consequences of failing;" "When I take a test, my nervousness causes me to make careless errors;" and "During tests, the thought frequently occurs to me that I may not be too bright."

Instructors can find opportunities to illustrate the advantages and disadvantages of a moderate amount of anxiety by asking students to complete a test anxiety scale either before or after an exam. The instructor can then demonstrate whether test anxiety predicts exam grades. These scales can also open up discussion about test-taking strategies and how to improve test performance by focusing on anxiety-mitigation, metacognition, mindfulness, or other relaxation techniques.

### *Computer Anxiety*

As technology continues to advance, it may be difficult to imagine students' experiencing anxiety toward using computers. However, computer anxiety is very real, can influence students' attitudes toward taking computer-administered exams (Schult & McIntosh, 2004), and can negatively influence students' performance (Brosnan, 1998).

The Computer Anxiety and Learning Measure (CALM; McInerney, Marsh, & McInerney, 1999) is a 65-item measure consisting of four subscales: gaining initial computer scale (22 items), state anxiety (20 items), sense of control (12 items), and computing self-concept (11 items). The CALM is reliable (Cronbach's αs > .78) and the subscales allow researchers to administer all, some, or only one, and still gain insight regarding aspects of students' computer anxiety (see Schult & McIntosh, 2004 for a SoTL study in which only the state anxiety subscale was used).

One possible disadvantage to the CALM measure is that a few items may be outdated (e.g., questions measuring comfort using a mouse or printing documents) due to the prevalence of computers in most people's everyday lives. However, one other well-validated computer anxiety scale exists and may serve as an excellent alternative if the CALM does not meet an instructor's needs. The Computer Anxiety Scale (CAS; Lester, Yang, & James, 2005) contains six items and participants respond using a 6-point agreement anchor (strongly agree to strongly disagree). The six items load onto a single underlying factor of computer anxiety (Chronbach's *α*s > .74 across multiple samples). Items such as, "I feel confident and relaxed whiling working on a computer" and "I can usually manage to solve computer problems by myself" make up the scale. For instructors solely seeking to assess the affective components of computer anxiety, we recommend the state subscale of the CALM; but for researchers seeking to assess cognitive components, the other four subscales of the CALM or the CAS offer practical solutions.

### *Math Anxiety*

Psychological statistics professors often quip that teaching psychological statistics is one part teaching math and one part anxiety mitigation. It comes as no surprise that several scales exist to measure students' self-reported anxieties toward math. An early, yet still widely used, measure is the Math Anxiety Scale (Betz, 1978; Fennema & Sherman, 1976). Betz (1978) rewrote 10 items from the Mathematics Attitudes Scale (Fennema & Sherman, 1976) with the goal of assessing college students' math anxiety. Participants respond to these 10 items on a 5-point agreement scale (*strongly* disagree to *strongly agree*). The items offer good reliability (split-half coefficient of .92; Betz, 1978) and measure worry about solving math problems, feelings of tension during math exams or when thinking of difficult math problems, and loss of clear thinking when working with math.

Another widely used measure is the Math Anxiety Rating Scale (MARS-R; Plake & Parker, 1982). Although the original MARS-R contained 24 items, Hopko (2003) conducted a re-validation study and reduced the measure to 12 items. Participants respond to items using a 4-point scale from 0 (*no anxiety*) to 4 (high anxiety). Both the original and revised versions consist of two subscales: Learning Math Anxiety and Math Evaluation Anxiety. Both subscales have good reliability (Learning Math Anxiety, Cronbach's α = .87; Math Evaluation Anxiety, α = .85; Hopko, 2003).

By assessing students' trait or state feelings before engaging in an academic task, instructors can assess whether some of the positive well-being constructs buffer against the negative effects of test or computer anxiety. For instance, students with high math anxiety tend to perform more poorly on math exams, except when they possess high metacognitive skills (Legg & Locker, 2009). The high metacognitive skill allows these students to compensate for and overcome potentially debilitating math anxiety. Instructors can also use these scales to demonstrate peoples' varying levels of anxiety (e.g., some people may score high on math anxiety but low on computer anxiety).

## Conclusion

In sum, a variety of scales exist to measure the positive and negative aspects of trait and state well-being, and these assessments could serve as either predictors or dependent variables in research projects for SoTL scholars. Importantly, most of the scales are self-report and brief, and are therefore highly convenient for use within the classroom. In addition, although we reported some examples of instructors exploring well-being in the classroom, we also pointed out many new ways in which SoTL scores can contribute to this burgeoning literature. Lastly, not only can the scales themselves be informative to students' understanding of research constructs, but they can also help instructors understand and promote intangible emotional characteristics that might help students thrive.

# References

References marked with an asterisk indicate a study.

*Betz, N. E., (1978). Prevalence, distribution, and correlates of math anxiety in college students. *Journal of Counseling Psychology, 25,* 441-448. doi:10.1037/0022-0167.25.5.441

Borrello, A. (2005). *Subjective well-being and academic success among college students.* (Unpublished doctoral dissertation), Capella University, Minneapolis, Minnesota.

Brosnan, M. J. (1998). The impact of computer anxiety and self-efficacy upon performance. *Journal of Computer Assisted Learning, 14*(3)*,* 223-234. doi:10.1046/j.1365-2729.1998.143059.x

Busseri, M. A., & Sadava, S. W. (2010). A review of the tripartite structure of subjective well-being: Implications for conceptualization, operationalization, analysis, and synthesis. *Personality and Social Psychology Review, 15*(3), 290-314. doi:10.1177/1088868310391271

*Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2)*,* 270-295. doi:10.1006/ceps.2001.1094

*Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality of Life Research, 23*(10), 2809-2818. doi:10.1007/s11136-014-0726-4

*Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior, 24*(4), 385-396. doi:10.2307/2136404

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: self-affirmation and social psychological intervention. *Annual Review of Psychology*, *65*, 333–71. doi:10.1146/annurev-psych-010213-115137

DeLongis, A., Folkman, S., & Lazarus, R. S. (1988). The impact of daily stress on health and mood: Psychological social resources as mediators. *Journal of Personality and Social Psychology, 54*(3)*,* 486-495. doi:10.1037/0022-3514.54.3.486

*Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology, 47*(5), 1105-1117. doi:10.1037/0022-3514.47.5.1105

*Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*(1), 71-75. doi:10.1207/s15327752jpa4901_13

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*(2)*,* 276-302. doi:10.1037/0033-2909.125.2.276

*Fennema, E., & Sherman, J. A. (1976). Fennema-Sherman Mathematics attitudes scales: Instruments designed to measure attitudes toward the learning mathematics by males and females. *Journal for Research in Mathematics Education, 7*(5), 324-326.

Fredrickson, B.L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist, 56*(3)*,* 218–226. doi:10.1037/0003-066X.56.3.218

*Fredrickson, B. L., Tugade, M. M., Waugh, C. E., & Larkin, G. R. (2003). What good are positive emotions in crises? A prospective study of resilience and emotions following the terrorist attacks on the United States on September 11th, 2001. *Journal of Personality and Social Psychology, 84*(2)*,* 362-376. doi:10.1037/0022-3514.84.2.365

Gehlbach, H. (2015). Seven survey sins. *Journal of Early Adolescence*. doi:10.1177/0272431615578276.

Gruber, J., Kogan, A., Quoidbach, J., & Mauss, I. B. (2013). Happiness is best kept stable: Positive emotion variability is associated with poorer psychological health. *Emotion*, *13*(1), 1-6. doi:10.1037/a0030262

Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research, 11*(2)*,* 213-218. doi:10.1016/0022-3999(67)90010-4

Hong, E., & Karstensson, L. (2002). Antecedents of state test anxiety. *Contemporary Educational Psychology, 27*(2), 348-367. doi:10.1006/ceps.2001.1095

Hopko, D. R. (2003). Confirmatory factor analysis of the math anxiety rating scale-revised. *Educational and Psychological Measurement, 63*(2), 336-351. doi:10.1177/0013164402251041

Hussey, H. D., & Lehan, T. (2015).  A primer on scale development. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Kashdan, T. B., Biswas-Diener, R., & King, L. A. (2008). Reconsidering happiness: The costs of distinguishing between hedonics and eudaimonia. *The Journal of Positive Psychology*, *3*(4), 219-233. doi:0.1080/17439760802303044

Kashdan, T. B., & Yuen, M. (2007). Whether highly curious students thrive academically depends on perceptions about the school learning environment: A study of Hong Kong adolescents. *Motivation and Emotion, 31*(4), 260-270. doi:10.1007/s11031-007-9074-9

*Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior, 43*(2)*,* 207-222. doi:10.2307/3090197

Krypel, M. N., & Henderson-king, D. (2010). Stress, coping styles, and optimism: Are they related to meaning of education in students' lives? *Social Psychology of Education: An International Journal, 13*(3), 409-424. doi:10.1007/s11218-010-9132-0

*Kluemper, D. H., Little, L. M., & DeGroot, T. (2009). State or trait: Effects of state optimism on job-related outcomes. *Journal of Organizational Behavior, 30*(2), 209-231. doi:10.1002/job.591

Lamers, S. M. A., Westerhof, G. J., Bohlmeijer, E. T., ten Klooster, P. M., & Keyes, C. L. M. (2011). Evaluating the psychometric properties of the Mental Health Continuum-Short Form (MHC-SF). *Journal of Clinical Psychology, 67,* 99-110. doi:10.1002/jclp.20741

Legg, A. M., & Locker, L. (2009). Math performance and its relationship to math anxiety and metacognition. *North American Journal of Psychology, 11*(3)*,* 471-485.

Lepp, A., Barkley, J. E., & Karpinski, A. C. (2014). The relationship between cell phone use, academic performance, anxiety, and satisfaction with life in college students. *Computers in Human Behavior, 31*, 343-350. doi:10.1016/j.chb.2013.10.049

*Lester, D., Yang, B., & James, S. (2005). A short computer anxiety scale. *Perceptual and Motor Skills, 100*(3)*,* 964-968. doi:10.2466/pms.100.3c.964-968

*Liebert & Morris (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Physiological Reports, 20*(3)*,* 975-978. doi:10.2466/pr0.1967.20.3.975

Lyubomirsky, S., King, L., Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin, 131*(6), 803-855. doi:10.1037/0033-2909.131.6.803

Lyubomirsky, S., & Layous, K. (2013). How do simple positive activities increase well-being? *Current Directions in Psychological Science, 22*(1)*,* 57-62. doi:10.1177/0963721412469809

*Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 46*(2)*,* 137-155. doi:10.1023/A:1006824100041

*Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State-Trait Anxiety Inventory (STAI). *British Journal of Clinical Psychology, 31*(3)*,* 301-306. doi:10.1111/j.2044-8260.1992.tb00997.x

Maybury, K. K. (2013). The influence of a positive psychology course on student well-being. *Teaching of Psychology, 40*(1), 62-65. doi:10.1177/0098628312465868

*McInerney, V., Marsh, H. W., & McInerney, D. M. (1999). The designing of the Computer Anxiety and Learning Measure (CALM): Validation of scores on a multidimensional measure of anxiety and cognitions relating to adult learning of computing skills using structural equation modeling. *Education and Psychological Measurement, 59*(3)*,* 451-470. doi:10.1177/00131649921969974

O'Connor, B. P., Crawford, M. R., & Holder, M. D. (2014). An item response theory analysis of the subjective happiness scale. *Social Indicators Research*. doi:10.1007/s11205-014-0773-9.

Ouweneel, E., Le Blanc, P. M., Schaufeli, W. B. (2011). Flourishing students: A longitudinal study on positive emotions, personal resources, and study engagement. *The Journal of Positive Psychology, 6*(2), 142-143. doi:10.1080/17439760.2011.558847

Pavot, W., Diener, E., Colvin, C. R., & Sandvik, E. (1991). Further validation of the satisfaction with life scale: Evidence for the cross-method convergence of well-being measures. *Journal of Personality Assessment*, *57*(1), 149-161. doi:10.1207/s15327752jpa5701_17

*Plake, B. S., & Parker, C. S. (1982). The development and validation of a revised version of the Mathematics Anxiety Rating Scale. *Educational and Psychological Measurement, 42*(2)*,* 551-557. doi:10.1177/001316448204200218

Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General, 143*(6)*,* 2057-2066. doi:10.1037/a0038025

Rand, K. L., Martin, A. D., Shea, A. M. (2011). Hope, but not optimism, predicts academic performance of law students beyond previous academic achievement. *Journal of Research in Personality, 45*(6), 683-686. doi:10.1016/j.jrp.2011.08.004

*Renner, M. J., & Mackin, R. S. (1998). A life stress instrument for classroom use. *Teaching of Psychology, 25*(1)*,* 46-48. doi:10.1207/s15328023top2501_15

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2)*,* 353-387. doi:10.1037/a0026838

Ritchie, R. A., Meca, A., Madrazo, V. L., Schwartz, S. J., Hardy, S. A., Zamboanga, B. L., ... & Lee, R. M. (2013). Identity dimensions and related processes in emerging adulthood: Helpful or harmful? *Journal of Clinical Psychology*, *69*(4), 415-432. doi:10.1002/jclp.21960.

Robbins, A. S., Spence, J. T., & Clark, H. (1991). Psychological determinants of health and performance: The tangled web of desirable and undesirable characteristics. *Journal of Personality and Social Psychology, 61*, 755-765. doi:10.1002/jclp.21960

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55,* 68-78. doi:10.1037/0003-066X.55.1.68

*Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology, 57*(6), 1069-1081. doi:10.1037/0022-3514.57.6.1069

*Sarason, I. G., (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*(4)*,* 929-938. doi:10.1037/0022-3514.46.4.929

Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology, 4*(3)*,* 219–247. doi:10.1037/0278-6133.4.3.219

*Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology, 67*(6)*,* 1063-1078. doi:10.1037/0022-3514.67.6.1063

Schult, C. A., & McIntosh, J. L. (2004). Employing computer-administered exams in general psychology: Student anxiety and expectations. *Teaching of Psychology, 31*(3)*,* 209-211. doi:10.1207/s15328023top3103_7

Sheldon, K. M., & Hilpert, J. C. (2012). The balanced measure of psychological needs (BMPN) scale: An alternative domain general measure of need satisfaction. *Motivation and Emotion*, *36*(4), 439-451. doi:10.1007/s11031-012-9279-4

Sin, N. L., & Lyubomirsky, S. (2009). Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: A practice-friendly meta-analysis. *Journal of Clinical Psychology: In Session, 65*(5)*,* 467-487. doi:10.1002/jclp.20593

*Snyder, C. R., Harris, C., Anderson, J. R., Holleran, S. A., Irving, L. M., Sigmon, S. T., … & Harney, P. (1991). The will and the ways: Development and validation of an individual-differences measure of hope. *Journal of Personality and Social Psychology, 60,* 570-585. doi:10.1037/0022-3514.60.4.570

Snyder, C. R., Shorey, H. S., Cheavens, J., Pulvers, K. M., Adams III, V. H., & Wiklund, C. (2002). Hope and academic success in college. *Journal of Educational Psychology, 94,* 820-826. doi:10.1037/0022-0663.94.4.820

*Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory STAI (Form Y).* Palo Alto, CA: Consulting Psychologists Press.

*Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The Meaning in Life Questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology, 53,* 80-93. doi:10.1037/0022-0167.53.1.80

Thieman, T. J., Clary, E. G., Olson, A. M., Dauner, R. C., & Ring, E. E. (2009). Introducing students to psychological research: General psychology as a laboratory course. *Teaching of Psychology, 36*(3)*,* 160-168. doi:10.1080/00986280902959994

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*, 1447–1451. doi:10.1126/science.1198364

*Waterman, A. S., Schwartz, S. J., Zamboanga, B. L., Ravert, R. D., Williams, M. K., Agocha, V. B., … Donnellan, M. B. (2011). The Questionnaire for Eudaimonic Well-Being: Psychometric properties, demographic comparisons, and evidence of validity. *The Journal of Positive Psychology, 5,* 41-61. doi:10.1080/17439760903435208

*Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070. doi:10.1037/0022-3514.54.6.1063

# Chapter 12: Assessing Professor-Student Relationships Using Self-Report Scales

Jenna M. Meyerberg and Angela M. Legg

Pace University

Professor-student relationships are an important, and sometimes underestimated, part of college students' experiences. A positive relationship between professor and student can yield greater student motivation, learning, class attendance, effort, and overall satisfaction (Benson, Cohen, & Buskist, 2005; Wilson & Ryan, 2013). Professors can engage students through their behaviors and attitudes toward students. Positive behaviors can make students feel cared for and foster an interest in the class, motivating them to do better, resulting in the positive results mentioned above (Murray, 1997). The benefits of positive professor-student relationships can continue even after a student graduates. College graduates who report having had at least one professor who cared about them during college are 1.9 times more likely to feel engaged at work (Gallup, Inc., 2014). Unsurprisingly then, examining professor-student relationships through the perspective of scholarship on teaching and learning (SoTL) is a fruitful endeavor (Wilson, Wilson, & Legg, 2012).

Many factors influence the development of rapport and positive relationships between professors and students. Even positive first impressions developed prior to or before the first day of class can have far-reaching effects in terms of increased student motivation, retention, and grades (Legg & Wilson, 2009; Wilson & Wilson, 2007). Along with examining how positive professor-student relationships develop, scholars can also examine what types of outcomes arise due to these positive interactions. A. Richmond and colleagues, for example, measured students' evaluations of their professors' humor (see the Teacher Humor Scale below), perceived rapport (see the Professor-Student Rapport Scale below), student engagement, and a standard student rating of instruction for teacher effectiveness (A. Richmond, Berglund, Epelbaum, & Klein, 2015). These researchers demonstrated that professor-student relationships contribute a great deal to the perceptions of and experiences of college students. In fact, professor-student rapport alone accounted for 54% of the variance alone and the set of variables accounted for 59%.

Scholars typically assess professor-student relationships through the administration of self-report, survey measures given to students. Measures may focus on the relationship itself (Wilson, Ryan, & Pugh, 2010) or may assess other aspects of the relationship such as immediacy behaviors or use of humor (Frymier, Wanzer, & Wojtaszczyk, 2008; Keeley, Smith, & Buskist, 2006). Based on the extant literature within the professor-student relationship area, we created an illustration (see Figure 1) to demonstrate the connections between the three main relationship-related constructs discussed in this chapter: 1) immediacy behaviors, 2) rapport, and 3) the learning alliance (i.e., qualities of the professor-student relationship and student investment; Rogers, 2012). Using this literature, we theorize that each construct contributes to the next broader construct (Rogers, 2012; 2015; Wilson, Ryan, & Pugh, 2010). Thus, verbal and nonverbal immediacy behaviors are two aspects that build professor-student rapport and professor-student rapport forms a foundation for the learning alliance. Research also points to the relationship between these variables and desired student outcomes such as the positive relationship between instructors' expression of nonverbal immediacy behaviors and students' course grades (LeFebvre & Allen, 2014). At broader levels, rapport and the development of a learning alliance lead to positive outcomes such as student motivation, attitudes toward the course and professor, and affective and cognitive learning (Rogers, 2015; Wilson et al. 2011).

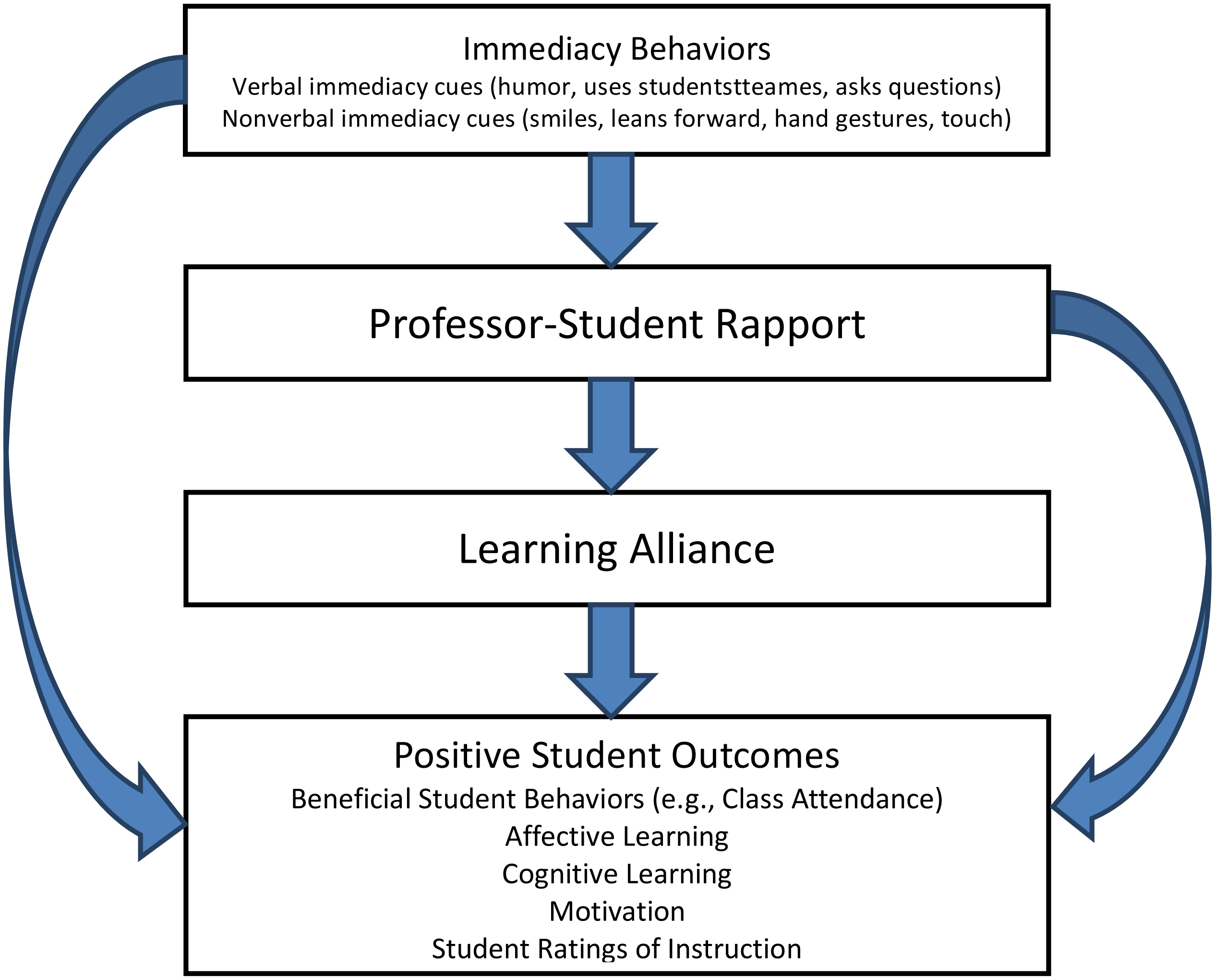
Immediacy Behaviors
Verbal immediacy cues (humor, uses studentstteames, asks questions)
Nonverbal immediacy cues (smiles, leans forward, hand gestures, touch)
Professor-Student Rapport
Learning Alliance
Positive Student Outcomes
Beneficial Student Behaviors (e.g., Class Attendance)
Affective Learning
Cognitive Learning
Motivation
Student Ratings of Instruction

*Figure 1.* Construct conceptualization of immediacy behaviors, professor-student rapport, learning alliance, and positive student outcomes.

Based on this framework, we begin our discussion with four scales that assess immediacy behaviors: the Immediacy Scale (Gorham, 1988), the Nonverbal Immediacy Scale (V. Richmond, McCroskey, & Johnson, 2003), the Teacher Humor Scale (Frymier et al., 2008), and the Teacher Behaviors Checklist (Keeley et al., 2006). We then consider the next largest construct, rapport, by discussing the Professor-Student Rapport Scale (Wilson et al., 2010). We end our scale analysis with the current conception of the largest construct, the bond, by describing the Learning Alliance Inventory (Rogers, 2012). Following review of each measure, exemplars are provided from the SoTL literature to demonstrate practical or empirical usage. The discussion ends with some unanswered questions regarding the assessment of professor-student relationships and call for future research in this area.

## Immediacy Scale

Immediacy is the nonverbal and verbal behaviors that communicate liking, caring, and approachability (Mehrabian, 1967; 1969). With regard to the classroom, immediacy refers to students' impressions of their instructors' availability or psychological distance. Early discussions of measuring immediacy came about in the late 1960s, with a focus on capturing verbal and nonverbal communications of closeness (Mehrabian, 1969). Mehrabian identified five primary immediacy behaviors: touch, physical distance between communicator and audience, leaning forward, making and maintaining eye contact, and physical orientation toward audience. While immediacy does not in and of itself represent rapport (immediacy is behavior-based whereas rapport relates more to cognitive perceptions), it is considered a catalyst and predictor of rapport and relationship-building (see above discussion of Figure 1).

Immediacy behaviors in the classroom received further exploration in the late 1980s (Gorham, 1988). Expanding on the five nonverbal behaviors outlined by Mehrabian (1969), Gorham created one of the first validated immediacy scales that included both nonverbal and verbal constructs. The 34-item scale contains 20 verbal items (e.g., "addresses students by name," "uses humor in class," and "asks questions that solicit viewpoints or opinions) and 14 nonverbal items (e.g., "gestures while talking to class," and "smiles at the class as a whole, not just individual students"). Students respond by rating the frequency with which each behavior occurs on a scale of 0 (*never occurs*) to 4 (*occurs very often*). This validation study demonstrated split-half reliabilities of .94 for the verbal items and .84 for the nonverbal items.

The original scale underwent some criticism when scholars argued that verbal behaviors represented teacher effectiveness, not immediacy (Robinson & Richmond, 1995). Inspired by this criticism, Wilson and Locker (2008) set out to empirically address this argument. Their validation provided evidence for discriminant validity between immediacy and teacher effectiveness. Further, their analysis of Gorham's (1988) original scale produced four distinct factors: 1) individual friendliness (8 items, α = .88), 2) flexibility during lecture (5 items, α = .80), 3) nonverbal immediacy (7 items, α = .76), and a single item assessing whether professors invite students to address them by their first name. Further, based on their analysis, Wilson and Locker (2008) recommended excluding the following three items: 1) "asks questions or encourages students to talk", 2) "refers to class as 'our' class or what 'we' are doing", and 3) "invites students to telephone or meet outside of class if they have a question or want to discuss something" from Gorham's (1988) original scale as these items did not add any predictive value to the scale. The most final version of this scale is Wilson and Locker (2008).

## Nonverbal Immediacy Scale

Whereas the updated immediacy scale (Gorham, 1980; Wilson & Locker, 2008) includes measures of both verbal and nonverbal psychological availability, the Nonverbal Immediacy Scale (NIS) focuses on nonverbal cues specifically (V. Richmond et al., 2003). This scale, which is not limited to professor-student relationships, contains a total of 26 items drawn from the previous literature (13 positively worded items and 13 negatively words items). An advantage to the NIS is the ability to distribute an other-reporting version (NIS-O) as well as a self-reporting (NIS-S) version. Although the items on the NIS-O and NIS-S measure the same constructs, their wording varies slightly in that the NIS-O frames items in terms of what "I" do and the NIS-S frames items in terms of what s/he (the professor) does. Participants rate the frequency of the behavior using a 5-point Likert-type scale from 0 (*never*) to 5 (*very often*). Sample items from the scale include, "I smile when I talk to people," "I use a monotone or dull voice while talking to people," and "I avoid eye contact while talking to people." Negative behaviors such as using a monotone voice and avoiding eye contact are reverse scored. This scale yielded high internal reliability estimates for the NIS-S and NIS-O of .90-.93 across several targets (self, teacher, supervisor, and a romantic date).

LeFebvre and Allen (2014) provided evidence for criterion validity of the Nonverbal Immediacy Scale. In their study, students enrolled in lecture/laboratory or self-contained courses completed the NIS, a measure of affective learning, an instruction evaluation measure, and allowed their grades to be tracked. Students who perceived greater nonverbal immediacy from their teaching assistants also received higher grades in the courses, reported more affective learning (e.g., liking of the teaching assistant, wanting to enroll in another course with that teaching assistant), and provided more positive instruction evaluations.

Aside from assessing students' perceptions of professors' nonverbal immediacy behaviors, students may also benefit from using this scale when discussing nonverbal communication, person-perception accuracy, and universal versus cultural understandings of emotional expression. Because the NIS (V. Richmond et al., 2003) has a self-assessment version and an other-assessment version, students can rate their own nonverbal immediacy behaviors and ask a close friend and a new acquaintance to also rate them. Through this activity, students can learn about measurement error, self- vs. other-rating methods, and convergent validity.

### Teacher Humor Scale

Humor is one component of verbal immediacy that may lead to increases in rapport when used in a context-appropriate and respectful manner. Research shows that the use of humor by the professor corresponds to teachers' immediacy behaviors in the classroom, and has an impact on learning outcomes (Gorham & Christophel, 1990). High ratings in professor humor orientation, or the predisposition for a professor to engage in humorous communication, are associated with greater positive affect toward the professor as well as increased perceived learning. There is also a positive correlation between perceived professor humor and nonverbal immediacy behaviors and responsiveness to students (Wanzer & Frymier, 1999).

An important part of successfully using humor, especially within a classroom setting, is the need to understand how the audience (e.g., the students) interprets humor. Torok, McMorris, and Lin (2004) conducted a study exploring students' perceptions of professor humor, specifically hoping to find the reasoning behind the positive relationship between professor humor and student engagement. Seventy three percent of students questioned stated that they felt very positively about their professors' use of humor, with 59% strongly believing that use of humor encourages a sense of community within the classroom. Regarding impact on ability to learn, 80% of students felt that the use of humor helped them master a concept. As Gladding (1995) noted, though, not all humor is positive. The types of humor most often endorsed by students as being welcome are, "funny stories, funny comments, jokes, and professional humor," (Toporek et al., 2004, pp. 16). Humor that students reported as the least enjoyable were aggressive or hostile humor, and humor of an ethnic or sexual nature.

Although the immediacy scales discussed previously (Gorham, 1988; Wilson & Locker, 2008) both include items about professors' use of humor, the Teacher Humor Scale (THS; Frymier et al., 2008) is an important tool for researchers seeking to assess this immediacy behavior more specifically. Further, because humor comes in many forms, the THS provides researchers with a tool to assess which types of humor positively influence professor-student relationships and which types should be avoided.

The Teacher Humor Scale (Frymier et al., 2008) can be used to measure different types of humor, for example, course-related and course-unrelated humor, self-disparaging humor, and unplanned humor. The THS is a 41-item scale that measures students' perceptions of the appropriateness or inappropriateness of a professor's use of humor. Students respond on a 5-point Likert-type scale ranging from 1 (*very inappropriate*) to 5 (*very appropriate*). A factor analysis of this scale yielded a five factor solution reflecting five distinct forms of humor for 25 of the original items: 1) Other Disparaging Humor (9 items, α = .93), 2) Related Humor (7 items, α = .85), 3) Unrelated Humor (3 items, α = .85), 4) Offensive Humor (3 items, α = .84), and 5) Self-Disparaging Humor (3 items, α = .80). Other Disparaging Humor includes items such as, "Makes humorous comments about a student's personal life or personal interests." An example of the Related Humor construct is, "Uses funny props to illustrate a concept or as an example." Unrelated Humor includes items such as, "Uses critical, cynical, or sarcastic humor about general topics (not related to the course)." Offensive Humor examples include, "Uses vulgar language or nonverbal behaviors in a humorous way." Finally, Self-Disparaging Humor includes, "Makes fun of themself [sic] when they make mistakes in class."

## Teacher Behaviors Checklist

One of the criticisms of early immediacy scales is that they merely measured teacher effectiveness, not immediacy (Robinson & Richmond, 1995). The Teacher Behaviors Checklist (TBC; Buskist, Sikorski, Buckley, & Saville, 2002) was created in order to assess qualities of professors who are highly esteemed in the eyes of their students and includes behaviors that also can communicate immediacy (see Kirk, Busler, Keeley & Buskist, 2015) for additional discussion of this measure). To create a scale that could serve as an evaluative instrument, the original 28-item list of behaviors was created such that students could rate the extent to which their professors engage in the checklist behaviors on a 5-point Likert-type scale from A (*frequently exhibits these behaviors*) to E (*never exhibits these behaviors*) (Keeley et al., 2006). Sample items include, "Enthusiastic About Teaching and About Topic (Smiles during class, prepares interesting class activities, uses gestures and expressions of emotion to emphasize important points, and arrives on time for class)," "Rapport (Makes class laugh through jokes and funny stories, initiates and maintains class discussions, knows student names, and interacts with students before and after class)," and "Respectful (Does not humiliate or embarrass students in class, is polite to students, does not interrupt students while they are talking, and does not talk down to students)." Factor analysis revealed one factor encompassing professional competency and communication skills (11 items, α = .90) and another factor more indicative of rapport which the authors refer to as the caring and supportive subscale (13 items, α = .93). Although only 24 items loaded onto the two factors, the authors recommend administering the full 28-item scale and calculating three scores, a total score, a caring and supportive score, and a professional competency and communication skills score.

Although most SoTL research examines students' perceptions of their actual professors, the TBC and many of the other scales we describe in this chapter can be used for hypothetical or imagined professors as well. For example, students completed the TBC and several other questionnaires while thinking about their ideal professor, one whom may not even exist (Komarraju, 2013). Her study provided evidence that different students view the ideal professor in different ways. Students with low self-efficacy and extrinsic motivations placed more importance on having caring professors than students who reported high self-efficacy and intrinsic motivations. This study further highlights the flexibility of the TBC as a scale that can facilitate research on both the predictors and outcomes of professor-student relationships.

## Professor-Student Rapport Scale

As described previously and illustrated in Figure 1, immediacy behaviors contribute to perceptions of professor-student rapport. Up until this point, our coverage predominantly has focused on scales that assess these smaller constructs that predict rapport and positive professor-student relationships. However, rapport represents much more than just behaviors and psychological availability. The Professor-Student Rapport Scale (PSRS; Wilson et al., 2010) was developed to address the potentially limiting factor of addressing only immediacy (see Figure 1 for our conceptualization of these constructs). "Rapport" is best understood as the agreement or harmony between two people (Wilson et al., 2010). Rapport has historically been measured within the therapist-client relationship, with no measures specifically addressing rapport in an academic setting, making the PSRS unique. Further, the PSRS is distinguishable from measures of immediacy by providing a larger scope of behaviors; if rapport is the positive relationship between professor and student, then immediacy behaviors are one way of achieving this bond (Wilson et al., 2010).

The 34-item scale assesses the professor-student relationship on a scale of 1 (*strongly disagree)* to 5 (*strongly agree*). The scale has excellent internal reliability (α = .96). Example items include, "My professor is understanding," "My professor's body language says, 'Don't bother me,'" and "My professor is aware of the amount of effort I am putting into this class." Further, the scale positively correlated with professor friendliness, flexibility, and nonverbal behaviors The PSRS also predicted students' attitudes toward their course and professor as well as their motivation, with higher scores on the PSRS predicting positive student attitudes and self-identified levels of motivation (Wilson et al., 2010).

The PSRS underwent further validation as the researchers sought to replicate the internal consistency demonstrated in their initial validation paper, establish test-retest reliability, and provide further convergent validation (Ryan, Wilson, & Pugh, 2011). As expected, the scale maintained a high level of internal reliability with a Cronbach's alpha of .89 and adequate test-retest reliability (*r* = .72). Convergent validity for this scale was assessed using The Working Alliance Inventory (Horvath & Greenberg, 1989), perceived social support, and verbal aggressiveness. As expected, the PSRS positively correlated with perceived professor-student relationships and perceived social support but negatively correlated with verbal aggressiveness.

Although the original 34-item measure (Wilson et al., 2010) will work for many researchers, some may require a shorter measure of rapport. Further investigation led to a shorter 15-item measure with nine items assessing "perceptions of teacher" (α = .92) and six items assessing "student engagement" (α = .84; Wilson & Ryan, 2013). Most notably, the six-item "student engagement" subscale emerged as the strongest predictor of perceived teacher effectiveness, attitude toward the teacher, student motivation, attitude toward the course, and perceived learning. As a final criterion validity indicator, student engagement significantly predicted students' course grades.

The PSRS, especially the six-item version, is a short, easy to administer scale. We recommend incorporating this scale into classroom discussions about impression formation, liking and relationship formation, and the halo effect. The PSRS is a good tool to illustrate how people can use heuristics and stereotypes when forming impressions of others. For example, participants who viewed older and younger photographs of a male or female professor gave lower (i.e., more negative) ratings to the older professors (Wilson, Beyer, & Monteiro, 2014). Students also tended to give lower ratings to the older female professor. Using this study, professors can discuss the PSRS, impression formation, and the way in which people use uncontrollable attributes (e.g., age) to assess others. This example can open up a discussion of the various ways people create rapport (e.g., by using immediate behaviors or humor or by improving their physical attractiveness).

## Learning Alliance Inventory

Just as the PSRS derived from measures in the clinical domain, the Learning Alliance Inventory (LAI; Rogers, 2012) also received inspiration from the clinical concept of the working alliance. This 18-item inventory uses a 7-point Likert-type response scale in which participants rate the extent to which the student endorses each statement from 1 (*not at all*) to 7 (*very much so*). Example items include, “My teacher and I work well together,” “My teacher welcomes all student input and feedback,” and “I enjoy doing the required tasks for this course.” The LAI contains three subscales, 1) collaborative bond (6 items, α = .91), 2) teacher competency (6 items, α = .93), and 3) student investment (6 items, α = .95). The scale’s initial validation demonstrated adequate test-retest reliability with reliabilities for each of the subscales that are in line with other similar scales (*r*s = .63 - .73) (e.g., Ryan et al., 2011). Additionally, all three subscales share a small effect in predicting course grades (*r*s = .19 - .25). Rogers (2015) provided further validation of his scale and specifically compared the LAI to the PSRS (Wilson et al., 2010) and the NIS (V. Richmond et al., 2003). The LAI did positively correlate with both scales, as expected, though the LAI shared a stronger relationship with the PSRS than with the NIS. Further, the collaborative bond subscale significantly contributed to the students' self-reported learning and actual course grade even after controlling for the NIS and PSRS.

Figure 1 provides a construct conceptualization of the relationships between immediacy behaviors, rapport, perceptions of a learning alliance, and positive student outcomes. Of note, several mediational relationships appear in the figure. For example, immediacy behaviors have direct effects on learning but also have indirect effects through their connections to rapport and the learning alliance (Rogers, 2015). Although Rogers provided an excellent validation of the Learning Alliance Inventory, its use in the SoTL literature is nascent. We recommend its incorporation in future research that examines rapport, professor-student relationships, and the effects of positive relationships on students' and professors' outcomes (e.g., grades for students, burnout for professors).

## Future Directions

The study of professor-student relationships has come a long way since Mehrabian began operationalizing immediacy in the 60s. Hundreds of studies exist on the relationships between immediacy behaviors, instructors' use of humor, professor-student rapport, and perceptions of a learning alliance. Further, each one of these constructs predict positive student outcomes. Despite the vast attention given to the valid and reliable measurement of professor-student relationships, several important future directions exist.

Given the increase in online classes email, learning management systems, and social media, students now frequently interact with professors over electronic modes. One disadvantage of the scales we described is that some of the items may not translate well in an online course (e.g., nonverbal immediacy items measuring eye contact and physical distance). At the present time, a computer-mediated-communication scale that assesses professor-student relationships does not exist, despite evidence that professor behaviors can hinder or help rapport building through online classes (Arbaugh, 2001) or through the use of electronic communication (Legg & Wilson, 2009). Future scales would benefit from including items that address online classes and electronic communication, for example: “Instructor responds within a reasonable amount of time to my electronic requests,” “My instructor shows s/he cares about me in the way s/he responds to my emails,” or “I can tell my instructor cares about students by the way s/he designed our Learning Management System (e.g., Blackboard).”

The vast majority of research on professor-student relationships examines the relationship formed between undergraduate students and their professors. However, these findings may not generalize to the relationship formed and experienced by graduate students and their advisors. For example, the same amount of warmth and caring that may promote learning for undergraduate students who lack intrinsic motivation (Komarraju, 2013) may be seen as coddling or lacking in rigor within the context of graduate education. Compared to undergraduates' relationships with their professors, graduate students and their advisors may spend more time together, both in academic and casual settings. This added time may facilitate building learning alliances and rapport but may also pose some additional challenges that could help or hinder the graduate students' success. In one rare study focusing on a graduate population, graduate students listed interest/support and characteristics such as sense of humor and empathetic as some of their top qualities in a good graduate mentor (Cronan-Hillix, Gensheimer, Cronan-Hillix, & Davidson, 1986). This study, however, did not compare graduate students' perceptions with those of undergraduates. We recommend that future research assess the similarities and differences between grad students and undergrads. In addition, future research can apply the existing validated measures to the graduate student-professor dyad or develop new measures to assess this unique relationship.

## Conclusion

A large amount of research now points to the critical influence of positive professor-student relationships (A. Richmond et al., 2015; Rogers, 2015; Wilson & Ryan, 2013). Students learn best when they have caring and competent mentors to facilitate their learning, spark their motivation, and provide emotional and cognitive support. It is our hope that this chapter provides a resource and foundation for new SoTL researchers who wish to measure the predictors and outcomes of positive professor-student relationships.

# References

References marked with an asterisk indicate a scale.

Arbaugh, J. B. (2001). How instructor immediacy behaviors affect student satisfaction and learning in web-based courses. *Business Communication Quarterly, 64*(4), 42-54. doi:10.1177/108056990106400405

Benson, A. T., Cohen L. A., & Buskist, W. (2005). Rapport: Its relation to student attitudes and behaviors toward teachers and classes. *Teaching of Psychology, 32*(4), 237–239. doi:10.1080/00986283.2010.510976

*Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer* (pp. 27–39). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Christophel, D. M. (1990). The relationships among teacher immediacy behaviors, student motivation, and learning. *Communication Education, 39*(4), 323–340. doi:10.1080/03634529009378813

Christophel, D. M., & Gorhamn, J. (1995). A test-retest analysis of student motivation, teacher immediacy, and perceived sources of motivation and demotivation in college classes. *Communication Education, 44*(4), 292–306. doi:10.1080/03634529509379020

Cronan-Hillix, T., Gensheimer, L. K., Cronan-Hillix, W. A., & Davidson, W. S. (1986). Students' views of mentors in psychology graduate training. *Teaching of Psychology, 13*(3), 123-127. doi:10.1207/s15328023top1303_5

*Frymier, A. B., Wanzer, M. B., & Wojtaszcyk, A. M. (2008). Assessing students' perceptions of inappropriate and appropriate teacher humor. *Communication Education, 57*(2), 266–288. doi:10.1080/03634520701687183

Gallup, Inc. (June 1, 2014). *Great jobs, great lives: The 2014 Gallup-Purdue Index Inaugural national report*. Retrieved from: http://www.gallup.com/services/176768/2014-gallup-purdue-index-report.aspx

Gladding, S. T. (1995). Humor in counseling: Using a natural resource. *Journal of Humanistic Education & Development, 34*(1), 3 – 13. doi:10.1002/j.2164-4683.1995.tb00106.x

*Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education, 37*(1), 40–53. doi:10.1080/03634529009378813

Gorham, J., & Christophel, D. M. (1990). The relationship of teachers' use of humor in the classroom to immediacy and student learning. *Communication Education, 39*(1), 46 – 62. doi:10.1080/0363452900937878

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology, 36*(2), 223-233. doi:10.1037/0022-0167.36.2.223

*Keeley, J., Smith, D., & Buskist, W. (2006). The teacher behaviors checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*(2), 84–91. doi:10.1207/s15328023top33021

Kirk, C., Busler, J., Keeley, J. & Buskist, (2015). Effective tools for assessing characteristics of excellent teaching: The teacher behaviors checklist. In R. S. Jhangiani, J. D. Troisi , B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning.*

Komarraju, M. (2013). Ideal teacher behaviors: Student motivation and self-efficacy predict preferences. *Teaching of Psychology, 40*(2), 104-110. doi:10.1177/0098628312475029

LeFebvre, L., & Allen, M. (2014). Teaching immediacy and student learning: An examination of lecture/laboratory and self-contained course sections. *Journal of Scholarship of Teaching and Learning, 14*(2), 29-45. doi:10.14434/josotl.v14i2.4002

Legg, A. M., & Wilson, J. H. (2009). E-mail from professor enhances student motivation and attitudes. *Teaching of Psychology, 36*(3), 205-211. doi:10.1080/00986280902960034

Mehrabian, A. (1967). Attitudes inferred from nonimmediacy of verbal communication. *Journal of Verbal Learning and Verbal Behavior, 6*(2), 294-295. doi:10.1016/S0022-5371(67)80113-0

Mehrabian, A. (1969). Some referents and measures of nonverbal behavior. *Behavior Research Methods and Instrumentation, 1*(6), 203–207. doi:10.3758/BF03208096

Murray, H. (1997). Effective teaching behaviors in the college classroom. In R. Perry & J. Smart (Eds.), *Effective teaching in higher education: Research and practice* (171– 204). New York: Agathon.

Richmond, A. S., Berglund, M. B., Epelbaum, V. B., & Klein, E. M. (2015). A + (b1) professor–student rapport + (b2) humor + (b3) student engagement = (Y) student ratings of instructors. *Teaching of Psychology, 42*(1), 119–125. doi:10.1177/0098628315569924

*Richmond, V. P., McCroskey, J. C., & Johnson, A. D. (2003). Development of the nonverbal immediacy scale (NIS): Measures of self- and other-perceived nonverbal immediacy. *Communication Quarterly, 51*(4), 504-517. doi:10.1080/01463370309370170

Robinson, R. Y., & Richmond, V. P. (1995). Validity of the verbal immediacy scale. *Communication Research Reports, 12*(1), 80-84. doi:10.1080/01463370309370170

*Rogers, D. T. (2012). The learning alliance inventory: Instrument development and initial validation. *International Journal for the Scholarship of Teaching and Learning, 6*(1), 1-16.

Rogers, D. T. (2015). Further validation of the learning alliance inventory: The roles of working alliance, rapport, and immediacy in student learning. *Teaching of Psychology, 42*(1), 19–25. doi:10.1177/0098628314562673

Ryan, R. G., Wilson, J. H., & Pugh, J. L. (2011). Psychometric characteristics of the professor-student rapport scale. *Teaching of Psychology, 38*(3), 135–141. doi:10.1177/0098628311411894

Torok, S. E., McMorris, R. F., & Lin, W. (2004). Is humor an appreciated teaching tool? Perceptions of professors' teaching styles and use of humor. *College Teaching, 52*(1), 14–20. doi:10.3200/CTCH.52.1.14-20

Wanzer, M. B., & Frymier, A. B. (1999). The relationship between student perceptions of instructor humor and students' reports of learning. *Communication Education, 48*(1), 48-62. doi:10.1080/03634529909379152

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger professors. *College Teaching, 6*(1), 20-24. doi:10.1080/87567555.2013.825574

*Wilson, J. H., & Locker Jr., L. (2008). Immediacy scale represents four factors: Nonverbal and verbal components predict student outcomes. *Journal of Classroom Interactions, 42*(2), 4-10.

*Wilson, J. H., Ryan, R. G., & Pugh, J. L. (2010). Professor-student rapport scale predicts student outcomes. *Teaching of Psychology, 37*(4), 246-251. doi:10.1080/00986283.2010.510976

Wilson, J. H., & Ryan, R. G. (2013). Professor-student rapport scale: Six items predict student outcomes. *Teaching of Psychology, 40*(2), 130–133. doi:10.1177/0098628312475033

Wilson, J. H., & Wilson, S. B. (2007). The first day of class affects student motivation: An experimental study. *Teaching of Psychology, 34*(4), 226–230. doi:10.1080/00986280701700151

Wilson, J. H., Wilson, S. B., & Legg, A. M. (2012). Building rapport in the classroom and student outcomes. In B. M. Schwartz & R. A. R. Gurung (Eds.), *Evidence-based teaching for higher education* (pp. 23-38). Washington, DC: American Psychological Association.

# Chapter 13: Effective Tools for Assessing Characteristics of Excellent Teaching: The Teacher Behaviors Checklist as Exemplar

Claire Kirk[1], Jessica Busler[1], Jared Keeley[2], and William Buskist[1]

[1]Auburn University, [2]Mississippi State University

Learning the intricate craft of teaching requires that teachers undergo frequent and multiple-sourced assessment to improve their effectiveness. Fortunately, tools and techniques for evaluating instruction are numerous. Beneficial and comprehensive feedback is best derived from implementing several different assessment methods across multiple informants, which may include students, colleagues, and self-reflection. Gathering information from these sources regarding one's teaching is important because each contributes a unique perspective to understanding one's teaching abilities (Keeley, 2012). This chapter will briefly review some general strategies regarding the assessment of teaching characteristics such as feedback from students and peers, as well as actual teaching materials, followed by an in-depth examination of one measure: the Teacher Behaviors Checklist (Buskist, Sikorski, Buckley, & Saville, 2002; Keeley, Smith, & Buskist, 2006).

## Measuring Teaching Quality—Multimodal Assessment

The most commonly used teaching assessors are naturally the individuals being taught: students. They experience and observe any particular teacher quite frequently, and can be an excellent source of information. However, gathering students' feedback on one's teaching is controversial due to their lack of objectivity. Additionally, students are often unable to accurately assess the amount of information they are learning, or how well that knowledge will be retained over time.  Relying solely on students to provide insight about one's teaching is also controversial due to the lack of objectivity among students. For example, the way in which students perceive their progress in a course may vary as a function of their unique individual experiences with the course material and the instructor. Factors such as grading leniency and difficulty of the material also may potentially affect student ratings (Ellis, Burke, Lomire, & McCormack 2004; Marsh & Roche, 2000). Regardless, student perspectives of teaching and learning are valuable in that they may be compared to institutional and teacher-developed measures of learning (Keeley, 2012). To increase the validity of student evaluations of teaching, several methods may be used in conjunction, including mini-assessments, rating scales, graded assignments and tests, and student focus groups.

### Mini-Assessments

There are a variety of simple and quick assessments that students can complete in class to help teachers gather basic information on student learning (Angelo & Cross, 1993). One example is the "minute paper," which is a quickly written response to a topic or question of the teacher's choice. The minute paper may focus on a topic covered in class, or may be used as a student evaluation of teaching. Students may write directly about what the teacher is doing well and what might be improved. Once the time limit (frequently 1-2 minutes) has expired, the teacher asks students to turn in their papers anonymously, and later reads through the responses looking for themes regarding, for example, what students are or are not learning well or what improvements might be made in the course to improve student learning.

### Graded Assignments and Tests

Students' graded assignments and tests are readily available yet often overlooked as evaluation tools of instruction. Grades provide an objective albeit an indirect measure of teaching quality as well as student learning. Grades may be a good indicator of how well students understand certain material; teachers can use this information to improve the way in which subsequent instruction is structured and presented. Any graded assignment (e.g., tests, quizzes, papers, activities) could potentially serve this purpose, assuming the teacher is able to directly tie the material in the graded assignment to a learning objective for the course (e.g., a test question to identify the statistical concepts in an advertisement as a measure of statistical literacy). However, keep in mind that graded assignments and tests may not be as easily interpretable as other measures of teaching quality because many factors contribute to an individual's performance (e.g., motivation, ability, study skills, and educational history).

### Student Focus Groups

The purpose of student focus groups is to obtain more detailed information than might be possible if the entire class was surveyed. To avoid bias in student feedback, it is advantageous for an outside consultant (e.g., staff from the institution's teaching and learning center, instructor from another department) to gather information from a subset of students in a course. Instructors may work with the consultant beforehand to determine the type of information to be gathered from the student groups.

### Student Evaluations of Teaching

The most common form of teaching evaluation is the student evaluation of teaching (SET), which is typically a set of written fixed-answer questions that is most beneficial when well developed and empirically supported. Numerous rating scales with unknown psychometric properties are accessible; however, there are also several empirically-supported instruments available. These instruments are reliable, valid, and consistent over time, and include the Student Evaluation of Educational Quality (SEEQ; Marsh, 1982), Barnes et al.'s (2008) measure, and the Teacher Behaviors Checklist (TBC; Buskist et al., 2002; Keeley et al., 2006). In addition to these rating scales, the Office of Educational Assessment at the University of Washington (2005) has developed a system consisting of several evaluation forms to measure teaching effectiveness. The Office of Educational Assessment at the University of Washington developed this detailed, empirically-supported evaluation system through a series of methodical interviews with faculty, administrators, and through student focus groups.

Despite the valuable feedback students might provide, they do not have the experience and expertise that professional teaching colleagues may contribute to the evaluation process. Peers may also have the ability to assist in troubleshooting classroom and teaching issues as well as offer guidance based on their experiences. Fellow teachers may provide a more objective teaching evaluation due to the removal of various factors such as grade leniency and difficulty of material (Keeley, 2012). Similar to student evaluations of instruction, peer evaluations are available in several forms including in vivo, peer review, and teaching portfolios.

### Peer Evaluation of Teaching

Peer evaluation of teaching, sometimes called peer review of teaching or peer consultation, provides teachers with highly specific information based on a sample of actual teaching behavior. This form of teaching evaluation involves a qualified peer (i.e., a person who is knowledgeable in pedagogy or who has otherwise been trained in the intricacies of peer review of teaching) observing an instructor teach a class session in order to gather observational information that is then later used as the basis for offering constructive comments regarding teaching content, style, and interaction with students. These facets of teaching, among others, are difficult, if not impossible, to capture by other evaluative methods. In order to obtain the most useful information from peer review of teaching, Ismail, Buskist, and Groccia (2012; see also Buskist, Ismail, & Groccia, 2013) recommend a five-step, thorough practice that includes a pre-observation meeting with the teacher, classroom observation, student focus groups, a written report prepared by the observer, and post-observation meeting with the teacher.

Peer review allows for a more detailed and comprehensive analysis of teaching behavior than other sources, and is held by some pedagogical researchers as the highest quality evaluative measure for the analysis of teaching effectiveness (Ismail et al., 2012). Benefits of the peer review process include empirical support for the effective improvement of teaching, and the opportunity for teachers to improve their teaching if conducted mid-semester (as opposed to the more typical end-of-the-semester SET). Peer review offers both observers and observees the opportunity to learn new teaching techniques and participate in collegial discussions regarding effective teaching practices (Ismail et al., 2012; Buskist et al., 2013).

### Teaching Portfolios

Teaching portfolios may take many forms and sometimes can provide a more in-depth sample of an instructor’s teaching than peer review. Portfolios are advantageous in that they are particularly effective for self-reflection (Seldin, 2004). A portfolio can also be disseminated widely, thereby benefiting numerous teachers in addition to the instructor who created it. Instructors may choose whether their portfolio covers their entire teaching career or simply a single course. Edgerton, Hutchings, and Quinlan (1991) suggested that the richest and most beneficial teaching portfolios combine primary teaching documents such as syllabi, tests, and presentations with corresponding personal reflective essays.

As an exemplar of the development of teaching portfolios, Xavier University of Louisiana has long implemented what it calls the Course Portfolio Working Group (CPWG), which encourages teaching improvement across colleges, departments, and content areas. Instructors who participate in the group focus on important topics such as student learning outcomes, teaching methods and practices, and assessment (Schafer, Hammer, & Berntsen, 2012). At the end of each school year, CPWG participants submit a completed course portfolio. The last session of Xavier's CPWG is devoted to reviewing each other's work and providing supportive feedback and suggestions on how to improve the teaching process in order to best benefit students. Course portfolios can be a tool with which to breathe new life into a course by consulting different colleagues and thinking critically about positive and feasible changes (Schafer et al., 2012).

### Syllabus and Teaching Philosophy Review

Unfortunately, peers are not always readily available to review one's teaching practices, and work groups akin to CPWG at Xavier University do not exist at every institution. Similarly, sometimes students do not respond to evaluation requests or there is not enough class time to allot for teaching evaluations. However, the practice of self-reflection is possible at any time throughout the year, as it only requires the course instructor's participation. Teachers, like students and peers, have a unique perspective on their teaching practices and how improvements may be made. Therefore, it is worthwhile to frequently reflect over one's teaching through a course syllabus and materials review as well as a teaching philosophy review.

Other aspects of a course, besides simply time spent teaching, may contribute to the overall effectiveness of one's teaching. A systematic review of the course syllabus may provide a wealth of information regarding class goals and direction, which often changes throughout the semester. In addition to reviewing the course syllabus, a review of one's teaching philosophy can be beneficial in ensuring congruence between one's personal philosophy of teaching and current course related activities. If course activities and goals do not align with one's teaching philosophy, changing specific elements within a course may be warranted.

## Development of the Teacher Behaviors Checklist

We now turn to the specific case of the TBC as an illustrative example of how to develop, investigate, and utilize an effective measure of teaching quality. The TBC is a SET that can be used in several ways to benefit one's teaching. Unlike most investigators who develop SETs, we did not set out to develop an evaluative instrument of teaching. Instead, our original aim was to conduct an exploratory investigation of the key behavioral attributes of excellent teachers. Our idea was that if we could identify such attributes, then perhaps we could teach these behaviors to others, particularly new faculty and graduate students who aspire to the professoriate. Our review of the vast college and university teaching literature at this time (circa 1998) revealed list after list of global teacher traits that researchers linked to outstanding teaching, for example, being approachable, caring, enthusiastic, interesting, and knowledgeable (e.g., Baiocco & DeWaters, 1998; Eble, 1984; Feldman, 1976; Lowman, 1995). Unfortunately, such lists do not lend themselves well to teaching others to teach—after all, what does it mean to be approachable or enthusiastic or knowledgeable? How do teachers actually demonstrate such traits?  Thus began our search for concrete and demonstrable behaviors that comprise master teaching.

## The Original TBC Study

Our approach to accomplishing this task involved exploring a range of personality qualities and their corresponding behaviors (Buskist et al., 2002). Additionally, we compared faculty and student perspectives on which of these qualities/behaviors are most important to excellent teaching. Phase 1 of our research asked undergraduates to list at least three qualities they judged to be reflective of master teaching at the college and university level. This sample produced a list of 47 characteristics. We then presented this list to a different group of undergraduates whom we instructed to "list or otherwise indicate up to three *specific* behaviors that reflect these qualities and characteristics."

We next analyzed students' behavioral descriptors for commonalities. In many instances they found the descriptors students used to characterize the 47 qualities showed substantial overlap, which resulted in collapsing the number of those categories to 28.

In Phase 2 of the original TBC study, another set of undergraduates and a sample of Auburn University faculty members selected the top 10 qualities/behaviors they judged to be key to master teaching at the college and university level. Students and faculty agreed on 6 of the top 10 qualities/behaviors (although in different order): (a) realistic expectations, (b) knowledgeable, (c) approachable/personable, (d) respectful, (e) creative/interesting, and (f) enthusiasm. With respect to the four remaining items on which students and faculty did not agree, there was an interesting, and as it turns out, generalizable difference between faculty and student rankings. Faculty tended to emphasize specific elements related to teaching technique (i.e., effective communication, prepared, current, and promoting critical thinking), whereas students emphasized aspects of the student and teacher relationship (i.e., understanding, happy/positive/humorous, encouraging, flexible). Indeed, recent work has found that perceived teacher-student rapport is one of the most important predictors of student SETs (Richmond, Berglund, Epelbaum, & Klein, 2015). Thus, teachers and students appear to share several similar views on behaviors reflective of master teaching but at the same time show important differences in their perspectives on key elements of excellent teaching.

## Factor Analysis of the TBC

Now that we had a scale, we needed to determine whether it was a valid and reliable instrument. We conducted a factor analysis to examine the basic factor structure of the instrument as well as measure its construct validity and internal reliability (Keeley et al., 2006).

### *Conversion to SET*

To collect psychometric data on the TBC, we converted the instrument to an evaluative inventory by adding a set of instructions and a 5-point Likert-type rating of the frequency of exhibiting each quality ranging from 1 (*never*) to 5 (*frequent*). The instructions asked students to rate their teacher on the extent to which they believed that their professor possessed each the 28 teacher qualities and their attendant behaviors. Our sample of students completed the TBC as well as the standard Auburn University end-of-the-semester eight-item teaching evaluation. Items on the standard evaluation addressed teacher qualities including helpfulness, organization and preparedness for the course, ability to motivate students and stimulate their thinking, clarity of teaching, and whether the professor spoke audibly. Having students complete both evaluations allowed us to have a standard of comparison for how the TBC related to the Auburn University evaluation.

### *Factor Analysis Results*

We submitted students' ratings to a factor analysis, which produced two subscales: (a) professional competency (11 items: approachable/personable, authoritative, confident, effective communicator, good listener, happy/positive/humorous, knowledgeable, prepared, punctuality/manages time, respectful, and technologically competent) and (b) caring and supportive behaviors (13 items: accessible, encourages and cares for students, enthusiastic, flexible/open-minded, humble, promotes class discussion, intellect stimulating, provides constructive feedback, rapport, realistic expectations and grading, sensitive/persistent, strives to be a better teacher, and understanding load on to the caring and supportive factor).

Our data derived from TBC student evaluations of four different instructors. We used two one-way ANOVAs to compare these teachers in order to assess whether these subscales discriminated among professors. For each subscale, we found significant differences among professors that correlated well with students' evaluations of these professors on the standard Auburn University evaluation.

We found internal consistency to be .95 for the total of all items. The professional competency subscale had a reliability coefficient of .90 and the caring and supportive subscale .93. We also examined the test-retest reliability of the scale using a new set of data from another group of different instructors and found that the total scale had a coefficient of .70 using midterm and end-of-term comparisons. The two subscales were also strongly reliable with .68 for the caring and supportive subscale and .72 for the professional competency subscale.

Thus, the TBC is a psychometrically sound and effective instrument for evaluating teaching quality and in particular, for assessing teaching excellence. The strong psychometric properties of the TBC along with its clear behavioral anchors allow teachers and others to diagnose and remediate specific problems that may characterize one's teaching.

## The TBC as a Research Tool

Soon after we published our initial article on the TBC (Buskist et al., 2002), we and others began examining the scale's applicability for the study of excellence in teaching across different institutional and cultural environments. At a liberal arts college and a community college, faculty and students tended to agree on the top teaching qualities (Schaeffer, Epting, Zinn, & Buskist, 2003; Wann, 2001). We found that students and faculty at both institutions rated six qualities similarly (realistic expectations, knowledgeable, approachable, respectful, creative/interesting, and enthusiastic). When comparing only the faculty from each institution, we found agreement on seven of the top 10 qualities (the same six qualities as the combined student/faculty rating with critical thinking as the seventh quality). In looking at student ratings only, students from both institutions ranked all of the same qualities as being in their top 10. The top 10 rated TBC qualities for students consisted of the same six as the combined student/faculty ratings along with happy/positive/humorous, encouraging, flexible, and understanding. This finding further supports the notion that students do indeed care significantly about the student-teacher relationship.

In a more recent study, we compared these sets of faculty with national award-winning faculty on the top 10 TBC qualities (Keeley, Ismail, & Buskist, in press). Interestingly, eight of the top 10 qualities selected by national award-winning faculty fell within the top 15 of both research institution faculty and community college faculty. Those qualities were: enthusiastic about teaching and topic, strives to be a better teacher, creative and interesting, knowledgeable about subject matter, approachable/personable, effective communicator, respectful, and encourages/cares for students. Notably, however, the two remaining qualities in the top 10 for national-award winning faculty, preparedness and rapport, were ranked as being in the 20th position or worse for both research institution and community college faculty. This finding suggests that excellent teachers (as operationalized by individuals who have won a national teaching award) place more emphasis on being thoroughly prepared for class as well as making an effort to create a caring and supportive classroom atmosphere for their students.

Thus far, our findings comparing data at different types of institutions provide evidence for the generalizability of the TBC as a measure of teaching excellence. Several cross-cultural studies extend the general nature of the TBC even further (e.g., Jõemaa, 2013; Keeley, Christopher, & Buskist, 2012; Liu, Keeley, & Buskist, 2015; Vulcano, 2007). For example, Vulcano (2007) surveyed two samples of Canadian undergraduates on their view of a “perfect instructor.” Students identified as many descriptors as they wished, which Vulcano then categorized into 26 sets of qualities and behaviors. The top 10 categories were (a) knowledgeable; (b) interesting and creative lectures; (c) approachable; (d) enthusiastic about teaching; (e) fair and realistic expectations; (f) humorous; (g) effective communicator; (h) flexible and open-minded; (i) encourages student participation; and (j) encourages and cares for students. Of the 26 categories devised, 24 of them were the same or similar to TBC items, which offers some in international support for general categories of excellent teaching, at least in terms of North America.

Keeley et al. (2012) recruited students at a small liberal arts school in Japan—Miyazaki International College—and from a small liberal arts school in the U.S.—Albion College—and had all participants complete the 28-item TBC by rating the extent to which a “master teacher” displays each quality and its attendant behaviors. American and Japanese students agreed on 7 of the top 10 teacher qualities: knowledgeable, confident, approachable/personal, enthusiastic, effective communicator, prepared, and good listener. The three discordant qualities for American students were accessible, respectful, and intellectually stimulating. These qualities were ranked 21st, 20th and 25th, respectively, for the Japanese students. For the Japanese students, the qualities of being creative and interesting, strives to be a better teacher, and humble rounded out their top 10 teacher qualities. American students rated these qualities as 18th, 22nd, and 24th, respectively.

In a similar study, participants at a large university in Eastern China rated three TBC qualities the same as Japanese and U.S. students: prepared, sensitive and persistent, and understanding (Liu, Keeley, & Buskist, 2015). We also observed some interesting differences among the three sets of students. One the one hand, Chinese students placed less emphasis on their teachers being approachable, confident, enthusiastic, knowledgeable, an effective communicator, and a good listener than either the Japanese or U.S. students. On the other hand, Chinese students placed more emphasis on only one quality, technologically competent, than both the Japanese and U.S. students. Chinese students seemed less interested in the interpersonal factors of teaching than Japanese students. Chinese students also ranked qualities such as accessible, flexible, punctual, respectful, establishes daily and academic term goals, presents current information, promotes critical thinking, and provides constructive feedback lower than the U.S. students.

In comparing our student rankings of the top 10 TBC qualities of master teaching results from the three U.S. samples, Canada, Japan, and China, only one quality, knowledge, made its way into the top 10 in each of those countries in every sample. However, in terms of rank order, six qualities, knowledgeable, approachable/personable, realistic expectations, creative and interesting, enthusiastic, and effective communicator, were ranked in the top 10 in at least five of the six samples. The implication of these findings is that students from different world regions may value teacher qualities differently. Thus, teachers using the TBC must recognize that higher or lower values on a particular item must be interpreted in the normative context of student preferences for that region.

Examining differences across academic disciplines, Liu, Keeley, and Buskist (in press) found that Chinese students majoring in psychology, education, and chemical engineering regarded five items as top qualities of a master teachers: respectful, knowledgeable, confident, strives to be a better teacher, and realistic expectations. Chemical engineering students placed greater emphasis on the TBC items of prepared, punctuality/manages class time, and daily and academic goals than their psychology or education counterparts. Education students rated the approachable/personable quality as being more important than psychology students. These findings suggest that Chinese students have a generally recognized set of qualities they associate with master teachers regardless of discipline, but nevertheless, students across academic disciplines may differ modestly in which qualities they link to excellence in teaching.

The cross-cultural work using the TBC discussed thus far has focused on students. Ismail (2014) compared two groups of faculty—U.S.-educated faculty and foreign-educated faculty (at the baccalaureate level)—teaching at U.S. institutions. Foreign educated-faculty and U.S.-educated faculty were in agreement on 9 of their top 10 TBC qualities: knowledgeable, enthusiastic, creative and interesting, promotes critical thinking, effective communicator, approachable, encouraging, manages time well, and accessible. Comparing these data with our research institution (Buskist et al., 2002) and community college faculty data (Schaeffer et al., 2003) reveals that five qualities (knowledgeable, enthusiastic, creative/interesting, promotes critical thinking, and approachable) were ranked in the top 10 across the four samples.

In summary, several TBC qualities emerged as being in the top 10 as rated by faculty in at least three of our faculty samples: knowledgeable, enthusiastic, creative/interesting, effective communicator, approachable, promotes critical thinking, encouraging.  We see a similar pattern for students in five of six student samples we have studied resulting in the following top qualities: knowledgeable, enthusiastic, creative/interesting, effective communicator, approachable, and realistic expectations.  The student and faculty lists of top qualities have five qualities in common: knowledgeable, enthusiastic, creative/interesting, effective communicator, and approachable. Such consistent agreement suggests that there may exist a universal set of qualities that comprise excellent teaching.

## The TBC as a SET and Professional Development Tool

Although research into excellent teaching qualities using the TBC is valuable in its own right, the TBC also has substantial practical utility as a student evaluation of teaching (SET). As noted earlier, Keeley et al. (2006) adapted the TBC by incorporating a 5-point rating scale for use as a SET. The measurement structure of the TBC is such that it provides a global estimate of quality teaching (the Total scale), which can be split into subscales of (a) Caring and Supportive behaviors and (b) Professional Competency and Communication Skills. Student ratings using the TBC are reliable (Keeley et al., 2006; Landrum & Stowell, 2013), and have been shown to meaningfully differentiate the quality of instruction (Keeley, English, Irons, & Henslee, 2013; Keeley, Furr, & Buskist, 2009).

The TBC is a behaviorally-based scale: each characteristic of teaching included in the measure encompasses several specific behaviors. These behavioral anchors undergird the utility of the TBC as a tool for teaching improvement. For example, if a teacher scores poorly on a particular characteristic, he or she may refer to the behaviors comprised by that characteristic for suggestions as how to implement positive changes to his or her teaching. Other teaching evaluation instruments typically only include qualitative descriptors of teaching, and can be difficult to translate into suggestions for making behavioral change in teaching (Keeley et al., 2006).

Feeley (2002) found that if students have a particularly strong positive or negative view of a teacher, they are likely to rate that teacher more positively or negatively overall, based on their opinion of only a single aspect of the teacher. Rating teachers on an overall “feeling” instead of providing objective, behaviorally-anchored feedback, makes it difficult for teachers to know precisely which specific aspect(s) of their teaching to improve. Several factors in addition to rating too broadly have been found to decrease the accuracy of TBC scores (Keeley et al., 2013). Students tend to rate teachers more highly in smaller courses, and non-instructional variables such as professors’ personality style can skew ratings in either direction (Clayson & Sheffet, 2006). Similarly, one area of low scores can decrease all of the others, creating an inaccurate rating profile (i.e., a negative halo effect). It is important for teachers to be aware of these rating biases and interpret scores appropriately.

One way to increase the accuracy of student TBC ratings is to explain to students the importance of accurate and conscientious rating. All too often, teachers rush through SETs at the end of the semester, without much explanation or guidance. Bernardin (1978) found that students provided more accurate ratings after being trained and educated about rating errors. Teachers can also compare their scores to a normative basis of similar courses at similar institutions to obtain more information regarding the meaning of their scores (Keeley et al., 2013).

The TBC is primarily intended for use in formative assessment—its focus is on improving one's teaching. However, it is equally useful for summative assessment, or the kind of evaluative assessment done for promotion, tenure, quality control, and merit raises. As an example, the TBC has been used in the teaching training program for graduate students in the Auburn University Psychology Department for many years. As part of their doctoral training, graduate students enroll in a teaching practicum course while simultaneously serving as a teaching assistant. As part of that experience, students use the TBC to evaluate their performance at the middle and end of the semester. They also provide mock lectures in the didactic portion of the experience, which are evaluated by an experienced faculty member using the TBC as a rating form. The behavioral anchors provide both the student and the faculty rater with direct suggestions for how to improve areas of weakness.

## Final Thoughts

The original development of the TBC took place more than a decade ago. Although it has generated much research and has been widely used as a SET, it is possible that changes, especially in technology and corresponding classroom practices, have occurred in the past 10 years that would change the behavioral anchors of the TBC. As such, we are currently undergoing a revitalization of the TBC (tentatively entitled TBC 2.0) that will update the list of qualities and their attendant behaviors.

In summary, it is crucial for all teachers—young and old, new and advanced, to utilize a variety of available methods of evaluation in order to improve the quality of their teaching. The development and use of the TBC provides an exemplar of how to create measures of quality teaching and evaluate their utility. Unfortunately, such a process has only been rarely undertaken, and most measures of teaching quality lack empirical support. However, we hope, and indeed challenge, all teachers to adopt a rigorous and systematic approach to the measurement of the quality of their teaching.

## References

References marked with an asterisk indicate a scale.

*Angelo, T. A., & Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). San Francisco, CA: Jossey-Bass.

Baiocco. S. H., & DeWaters, J. N. (1998). *Successful college teaching: Problem-solving strategies of distinguished professors.* Boston, MA: Allyn & Bacon.

*Barnes, D., Engelland, B., Matherine, C., Martin, W., Orgeron, C., Ring, J, Smith, G., & Williams, Z. (2008). Developing a psychometrically sound measure of collegiate teaching proficiency. *College Student Journal*, *42*, 199-213.

*Bernardin, H. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63,* 301-308. doi:10.1037/0021-9010.63.3.301

Buskist, W., Ismail, E., & Groccia, J. E. (2013). A practical model for conducting helpful peer review of teaching. In J. Sachs & M. Parsell (Eds.), *Peer review of learning and teaching in higher education: International perspectives* (pp. 33-52). Berlin, Germany: Springer.

*Buskist, W., Sikorski, J., Buckley, T., & Saville, B. K. (2002). Elements of master teaching. In S. F. Davis & W. Buskist (Eds.), *The teaching of psychology: Essays in honor of Wilbert J. McKeachie and Charles L. Brewer.* (pp. 27-39). Mahwah, NJ: Erlbaum.

Clayson, D. E., & Sheffet, M. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education, 28,* 149-160. doi:10.1177/0273475306288402

Eble, K. E. (1984). *The craft of teaching.* San Francisco, CA: Jossey-Bass.

Edgerton, R., Hutchings, P., & Quinlan, K. (1991). *The teaching portfolio: Capturing scholarship in teaching*. Washington, DC: American Association for Higher Education.

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2004). Student grades and average ratings of instructional quality: The need for adjustment. *Journal of Educational Research*, *9*, 35-41. doi:10.1080/00220670309596626

Feeley, T. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education, 51*(3), 225-236. doi:10.1080/03634520216519

Feldman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education, 5,* 243-288.

*Ismail, E. A. (2014). Foreign and US-educated faculty members' views on what constitutes excellent teaching. Unpublished Doctoral Dissertation, Auburn University, Auburn AL.

Ismail, E. A., Buskist, W., & Groccia, J. E. (2012). Peer review of teaching. In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 79-91). Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/evals2012/index.php

*Jõemaa, K. (2013). Student perceptions of master teachers in Estonian universities. Unpublished Master's Thesis, University of Tartu.

*Keeley, J., Christopher, A. N., & Buskist, W. (2012). Emerging evidence for excellent teaching across borders. In J. E. Groccia, M. Al-Sudairy, & W. Buskist (Eds.), *Handbook of college and university teaching: Global perspectives* (pp. 374-390). Thousand Oaks, CA: Sage.

*Keeley, J. (2012). Course and instructor evaluation. In W. Buskist & V. A. Benassi (Eds.). *Effective college and university teaching: Strategies and tactics for the new professoriate* (pp. 173-180). Thousand Oaks, CA: Sage.

*Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement, 73*, 440-457. doi:10.1177/0013164412475300

*Keeley, J., Furr, R. M., & Buskist, W. (2009). Differentiating psychology students' perceptions of teachers using the teacher behaviors checklist. *Teaching of Psychology*, *37*, 16-20. doi:10.1080/00986280903426282

*Keeley, J., Ismail, E. A., & Buskist, W. (in press). Excellent teachers' perspectives on excellent teaching. *Teaching of Psychology*.

*Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33,* 84-91. doi:10.1207/s15328023top3302_1

*Landrum, R. E., & Stowell, J. R. (2013). The reliability of student ratings of master teacher behaviors. *Teaching of Psychology*, *40*, 300-303. doi:10.1177/0098628313501043

*Lowman, J. (1995). *Mastering the techniques of teaching* (2nd ed.). San Francisco, CA: Jossey-Bass.

*Liu, S., Keeley, J., & Buskist, W. (2015). Chinese college students' perceptions of characteristics of excellent teachers. *Teaching of Psychology*, *42*, 83-86. doi:10.1177/0098628314562684

*Liu, S., Keeley, J., & Buskist, W. (in press). Chinese college students' perceptions of excellent teachers across three disciplines: Psychology, chemical engineering, and education. *Teaching of Psychology*.

*Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77–95. doi:10.1111/j.2044-8279.1982.tb0205.x

*Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload of students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92*, 202-228. doi:10.1037/0022-0663.92.1.202

Richmond, A. S., Berglund, M. B., Epelbaum, V. B., & Klein, E. M. (2015). a + ($b_1$) professor-student rapport + ($b_2$) humor + ($b_3$) student engagement = ($\hat{Y}$) student ratings of instructors. *Teaching of Psychology*, *42*, 119-125. doi:10.1177/0098628315569924

*Schaeffer, G., Epting, K., Zinn, T., & Buskist, W. (2003). Student and faculty perceptions of effective teaching. A successful replication. *Teaching of Psychology, 30,* 133-136.

Schafer, P., Hammer, E. Y., & Berntsen, J. (2012). Using course portfolios to assess and improve teaching. *Effective evaluation of teaching:* In M. E. Kite (Ed.), *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 71-78). Retrieved from the Society for the Teaching of Psychology web site: http://teachpsych.org/ebooks/evals2012/index.php

Seldin, P. (2004). The teaching portfolio: A practical guide to improved performance and promotion/tenure decisions (3rd ed.). San Francisco, CA: Jossey-Bass.

Vulcano, B. A. (2007). Extending the generality of the qualities and behaviors constituting effective teaching. *Teaching of Psychology, 34,* 114-117. doi:10.1080/00986280701293198

*Wann, P.D. (2001, January). *Faculty and study perceptions of behaviors of effective college teachers.* Poster presented at the National Institute for the Teaching of Psychology, St. Petersburg Beach, FL.

# Chapter 14: SoTL Scales: The Case of Missing Links

Aaron S. Richmond

Metropolitan State University of Denver

## SoTL is NOT Cryptozoology: A Brief SoTL History

Throughout this wonderful compendium of Scholarship of Teaching and Learning (SoTL) scales, many have described specific SoTL scales, how to use them, and how to create them. For instance, scales to measure student learning and/or self-efficacy (see Marek, Williamson, & Taglialatela, 2015), critical thinking skills (see Landrum & McCarthy, 2015), student interests and perceptions (see Zabel and Heger, 2015), service learning (see Simons, 2015), metacognitive awareness and skills (see Ku, 2015), student happiness, stress, and anxiety (see Layous, Nelson, & Legg, 2015), professor-student relationships (see Meyerberg & Legg, 2015), and professor efficacy (see Kirk, Busler, Keeley, & Buskist, 2015) have all been described. However, this is not where SoTL began. More likely, SoTL has been around for as long as teachers have been in existence—at least informally. That is, if you have ever observed great teachers, they often reflect and assess their teaching and make needed modifications and adjustments based on their conclusions. Informally, we also see SoTL as early as the 1920s. Robert Maynard, the incoming president of the University of Chicago, famously stated in his inaugural address,

> [A Ph.D. candidate who plans to be a teacher]…must be in touch with the most recent and most successful movements in undergraduate education, of which he now learns officially little or nothing. How should he learn about them? Not in my opinion by doing practice teaching upon the helpless undergraduate. Rather he should learn about them through seeing experiments carried on in undergraduate work by the members of the department in which he is studying for the degree…. (Scholarship of Teaching and learning: History, n.d.)

Yet formally, SoTL began with the onset of Boyer's (1990) Scholarship Reconsidered: Priorities of the Professoriate and continued by Shulman's (1993) work on Teaching as Community Property: Putting an End to Pedagogical Solitude. Between these two scholars, SoTL began to take root as a form of research intent on improving teaching in higher education. Since this time, scholars have conducted 1000s of SoTL studies in 100s of academic fields. As an artifact of this productivity—naturally—SoTL scales have been developed, tested, and implemented.

Considering the explosion of SoTL and subsequent SoTL scales (i.e., the long overdue need for this e-book), what then is needed or missing? That is, what are we SoTL scholars not measuring that we should be measuring? What are some issues with preexisting SoTL scales? Or in other words, how can we improve upon existing SoTL scales? It is my hope, throughout this chapter to answer these questions by identifying and illuminating these missing links (i.e., debunk cryptozoology!). Specifically, I will discuss the need for new and improved metacognitive and learning strategy measures, the need for SoTL scales that assess syllabi, comment on the issue that many SoTL self-report scales lack matching behavioral measures, and the need for a SoTL scale assessing model teaching characteristics.

## Nessie the Mythical Loch Ness Monster Rivals the Mythical SoTL Syllabus Scale

As mythical and mysterious as Nessie the Loch Ness Monster, research on syllabus construction and best practices has shown equal mystery and elusiveness. However, the syllabus has received some SoTL attention, as of late, as a key element to best practices in higher education instruction (e.g., Boysen, Richmond, & Gurung, 2015; Richmond et al., 2014; Slattery & Carlson, 2005). For example, the syllabus can have an immense (positive or negative) impact on how students perceive teaching effectiveness (Richmond, Becknell, Slattery, Morgan, & Mitchell, 2015; Saville, Zinn, Brown, & Marchuk, 2010). As this body of research grows, the need to assess syllabi, both reliably and validly, becomes more and more important. To date, the only SoTL scale to evaluate syllabi was developed by Cullen and Harris (2009). Cullen and Harris created a rubric to assess the degree to which a syllabus is considered to be learning-centered as opposed to teacher-centered. In the rubric, they describe three main factors (e.g., community, power and control, and evaluation/assessment). Within each main factor there are several subfactors. See Table 1 for a complete list and description. The scale is measured on a categorical level from 1 (*more teacher-centered*) to 4 (*more learner-centered*) rated by the instructor, not by students. For example, if a syllabus was learner-centered it would have a learning rationale that had a "rational provided for assignments, activities, methods, policies, and procedures; tied to learning outcomes" (p. 123).

Table 1
Cullen and Harris (2009) Rubric Assessing Learner-Centered Syllabi

| Factors and Sub Factors | | |
|---|---|---|
| *Community* | *Power and Control* | *Evaluation/Assessment* |
| Accessibility of Teacher | Teacher's Role | Grades |
| Learning Rationale | Student's Role | Feedback Mechanisms |
| Collaboration | Outside Resources | Evaluation |
| | Syllabus Tone | Learning Outcomes |
| | Syllabus Focus | Revision/Redoing |

Whereas, if the syllabus were teacher-centered it would have a learning rationale that had "no rationale provided for assignments or activities" (p. 123). Unfortunately, there are a number of issues with this SoTL measure. First, the level of measurement is categorical. That is, the rubric is on a 1-4 rubric/scale that only describes categories or degrees of level of learner-centeredness. Because of this level of measurement it makes it almost impossible to understand a factor-structure to the scale, and assess reliability and validity. Second, although a few studies have used the rubric (e.g., Slattery et al., 2014), there is virtually no further evidence that may suggest how reliable or valid the scale is. Third, this rubric only takes the pedagogical perspective of student-centered instruction. There are several other effective forms of pedagogies that should also be assessed as valuable tools for syllabus construction (e.g., Inter-teaching, Just-in-time teaching, etc.).

Based on my review of the Cullen and Harris (2009) SoTL rubric and the lack of any other SoTL scales that assess syllabi, I suggest two major directions in this area of SoTL. First, SoTL researchers should modify Cullen and Harris' rubric to at least have an interval or better scale. That is, convert the 1-4 categories into Likert-type questions. For instance, the factor of Evaluation/Assessment and the subfactor of feedback mechanisms, scale question would read, "The syllabus describes summative and formative evaluations including written and oral presentations, group work, self-evaluation and peer evaluation" anchored in *1* (*strongly agree*) to *5* (*strongly disagree*) or could have anchors of frequency, such as *1* (*always*) to *6* (*never*). This would allow SoTL researchers to conduct factor analyses, test-retest reliability, split-half reliability, and convergent, construct, content, and predictive validity studies of this SoTL scale. Second, other forms of SoTL scales that assess the efficacy of syllabi are needed. It has been argued that exemplar syllabi should serve as a contract to students, a permanent record, a cognitive map, a learning tool, and a communication device (Matejka & Kurke, 1994; Parkes & Harris, 2002). As such, SoTL researchers should devise scales that assess the degree to which syllabi contain these elements. For example, researchers could use these scales to assess differences between college and university teachers' pedagogy. Or, researchers could use these scales to assess the efficacy of the syllabus design and how it may affect student learning. In the end, let's demystify the syllabus (aka Nesse the Lochness Monster) because not only is syllabus research under-studied in SoTL, there is great room and need to develop and validate SoTL scales which attempt to assess the efficacy of syllabi.

## Prove that Bigfoot is Real! Self-Report SoTL Scales Need Matching Behavioral Measures

Do you know someone who believes in Bigfoot? How do they know? Did they see the elusive creature? Or did they say, "My cousin once saw Bigfoot" or "My neighbor Billy lost three goats to Bigfoot"? Unlike the TV show MonsterQuest by the beloved History Channel, it is likely that most people who believe in Bigfoot do so because of self-report or the report of others. Herein lies the rub. Somewhat like cryptozoology, many SoTL scales rely heavily on self-report and not enough on actual behavior. As psychologists, we know all to well the pitfalls of self-report scales. That is, issues of honesty, introspection, social desirability, understanding and comprehension, response bias, response set, and on and on. As Baumeister, Vohs, and Funder (2007) so poignantly said,

> the eclipse of behavior…in which direct observation of behavior has been increasingly supplanted by introspective self reports, hypothetical scenarios, and questionnaire ratings. We advocate a renewed commitment to including direct observation of behavior whenever possible and in at least a healthy minority of research projects. (p. 396)

However, this does not mean that self-report SoTL scales are worthless. Rather, I'm here to suggest that there are three primary methodological solutions.

First, when using self-report SoTL scales, it is important to have matching or complementing behavioral measures that support and have consistent results with the self-report scales. Let's illustrate this issue through a common measure of metacognitive awareness, the Metacognitive Awareness Inventory (MAI; Schraw & Dennison, 1994). The MAI has been used and cited in over 1000 studies and is a 52-item inventory that attempts to measure the metacognitive components of knowledge of cognition and regulation of cognition (Schraw & Dennison, 1994). For example, one of the items used to assess the knowledge of cognition subscale states, "I know when each strategy I use will be most effective" (p. 473). The MAI is scored on a 5-point Likert scale with the anchors of *1* (*true of me*) to *5* (*not true of me*). The problem here is the MAI reports what the participant believes to be true, not what he or she actually does (i.e., behavior).

Instead, I suggest that when using such SoTL scales or developing new SoTL scales, that researchers should also use complementing behavioral measures to complement and validate said self-report measures. In this example, the complement to the MAI would be to collect the behavioral measure of actual metacognitive behavior (e.g., calibration). Calibration is a measure of metacognition that measures of the difference between one's judgment of performance and one's actual performance (see Schraw, Kuch, & Guitierrez, 2013; Schraw, Kuch, Guitierrez, & Richmond, 2014). To measure calibration you ask individuals to answer a question on a assessment, you then ask individuals to answer whether they believed they got the answer correct or not. Next, you record whether they got the question correct or incorrect. Then you mathematically calculate the difference between what they said they got correct or incorrect and what they actually did. The resulting product -1.0 (always underconfident) to +1.0 (always overconfident) is the degree to which they were accurate at knowing what they know or knowing what they do not know. A score closer to 0 indicates high accuracy. (For more information on scoring see Schraw et al., 2013; 2014). By combining both the MAI and calibration, SoTL researchers will not only have a highly reliable measure of metacognition and the effects that a treatment may have on it, but they will also have a very internally and externally valid measure of metacognition.

Landrum and Stowell (2013) provide a great example of how to match self-report measures with actual behavior. The purpose of their study was to validate the Teacher Behaviors Checklist (TBC; Keeley, Smith & Buskist, 2006) by matching corresponding self-reported behaviors to observed teaching behaviors. Specifically, Landrum and Stowell had over 700 students watch several 5-minute video vignettes that were designed to demonstrate master teacher behaviors purported by the TBC (e.g., respectful, enthusiastic, approachable, etc.). They then asked the students to rate each professor on the TBC. Landrum and Stowell found that when students viewed the same vignette (e.g., demonstrating respect to students), that the students were often very consistent in their rating of the TPC (i.e., agreement ranged from 68% - 91%). Therefore, as demonstrated by Landrum and Stowell, when conducting future SoTL studies and scales, researchers should include both self-report and complementary behavioral measures.

A second solution to the problems with self-report scales is to use scales that are either behavioral checklists or scenario-based scales. As discussed in multiple chapters of this e-book, one of the prime examples of a SoTL scale, which measures behavior, is the Teacher Behaviors Checklist by Keeley, Smith, and Buskist (2006). The Teacher Behaviors Checklist (TBC) is a 28-item Likert Type scale where students assess the behaviors (e.g., effective communicator, preparedness, knowledgeable, enthusiastic, flexible/open-minded, etc.) of teachers. The TBC is rated on a 5-point Likert scale from *1* (*never*) to *5* (*frequent*) on how often the professor exhibits the given teaching behavior. For instance, for the behavior of 'provides constructive feedback,' the question would state, “Writes comments on returned work, answers students’ questions, and gives advice on test taking” (Keeley et al., 2006, p. 85). In the case of SoTL scales such as the TBC (or check out The Teacher Immediacy Scale by Gorham, 1988) these measures are, albeit indirect records of specific behaviors, but allow researchers to approximate specific behavioral outcomes.

Third, to mitigate some of the pitfalls with self-report scales is to use SoTL scales that are scenario based. For instance, Berry, West, and Denehey (1989) developed a self-report SoTL scale to measure self-efficacy for memory tasks called the Memory for Self-Efficacy Questionnaire (MSEQ). The MSEQ has been widely used and retains acceptable validity and reliability (Berry et al., 1989; West, Thorn & Bagwell, 2003). The measure is comprised of 40 confidence ratings on 10 memory scenarios. Within each scenario the first question describes the simplest memory task for the scenario then the 4th question describes the most difficult memory task. Individuals are asked to give their confidence rating (0-100%) on their ability to successfully accomplish the memory task. For example, in the MSEQ individuals are given a scenario about their ability to remember a grocery list. The first question (which is the easiest memory task) would state, “If I heard it twice, I could remember 2 items from a friend’s grocery list of the 12 items, without taking any list with me to the store” (Berry et al., 1989, p. 713). Whereas the most difficult memory task would state, “If I heard it twice, I could remember 12 items from a friend’s grocery list of 12 items, without taking any list with me to the store” (p. 713). Confidence ratings are then totaled for each memory scenario (e.g., grocery list, phone numbers, pictures, location, words, etc.). As you can see, SoTL scales such as the MSEQ are self-report, however they are rooted in past performance and actual real-life examples of what students or teachers may experience. The benefit of these types of measures is that they attempt to obtain external validity by putting the respondents in real-life scenarios.

In SoTL research, self-reports are inevitable and do serve a purpose. However, we are not cryptozoologists, we are SoTL scholars. So in order to debunk the mythical creatures such as Bigfoot, it is important that we complement self-report measures with behavioral measures and/or select and create SoTL scales that are rooted in actual behavior.

## Measuring Yeti’s Metacognition and Use of Learning Strategies

Do you think Yeti, aka the abdominal snowman, thinks about thinking or employs effective learning strategies? Likely—not! But if he did, how would we know? For that matter, how do we know our students’ metacognition and use of learning strategies? We could use the previously described MAI by Schraw and Dennison (1994), but it is 52-items long and a little theoretically dated (e.g., metacognitive awareness is more than knowledge and regulation of cognition). If you are combining this with other measures for a study, you run the risk of getting participant fatigue and consequential response sets (e.g., answering 5 to all questions or playing connect the dots). If you want something shorter you could use the Need for Cognition Scale (NCS) by Cacioppo, Petty, Feinstein, and Jarvis (1996). The NCS is a personality measure of metacognition intended to measure the extent to which an individual might need to engage in a cognitive activity (Cacioppo et al., 1996). The NCS consists of 18 questions rated on a 5-point Likert scale and has demonstrated to be quite reliable. However, many metacognitive researchers would suggest that the NCS does not measure metacognition, rather it measures a personality trait. Or you can use behavioral measures such as calibration (although a little tricky to calculate), or judgments of learning or feelings of knowing that are measured on confidence ratings (e.g., 0-100% confident). But this doesn't measure holistic beliefs about metacognition (i.e., all aspects of metacognition). You may also use the MSEQ, as previously described, but it takes a long time to complete and is hyper-specific to memory and self-efficacy—only one component of metacognition. What truly is needed, is a brief current SoTL measure of metacognition. For example, a researcher could possibly condense the MAI (Schraw & Dennison, 1994) and combine it with behavioral measures such as the MSEQ (Berry et al., 1989) and/or personality measures such as the NCS (Cacioppo et al., 1996). Future SoTL researchers should take up this endeavor to allow SoTL a quick, reliable, and valid metacognition measure that can be easily administered.

What if the Yeti was a teacher and you wanted to measure whether or not he was using metacognition in his teaching. How would you measure this? There is a new SoTL area that refers to this process as meta-teaching (Chen, 2013). Chen describes meta-teaching as,

> Like meta-cognition and meta-learning, meta-teaching, as 'teaching about teaching', can serve to design, examine and reflect on teaching. From practice-orientation, it defines what teaching activity is and what it is for, under which theoretical framework it is being carried out, and what experience and rules can be applied to it. Meanwhile, meta-teaching can assist teachers in discovering drawbacks in the teaching system and solving problems. This demonstrates that meta-teaching contains such functions such as understanding teaching, changing teaching and reflecting on teaching. (p. S64)

Furthermore, Spring (1985) argued that effective college and university teachers should use the meta-teaching strategies of proper lesson planning and goal setting, reflecting critically on appropriate use of effective instructional strategies to achieve the instructional goals, both formally and informally monitor student learning, and constantly evaluate the efficacy of chosen instructional strategies. To date, there are no SoTL scales that attempt to measure meta-teaching. As such, SoTL researchers and psychometricians should focus on building a meta-teaching inventory (both self-report and behavioral measure), which assess how often teachers engage in these practices.

What about measuring how the Yeti (aka students) uses learning strategies? The hallmark and stalwart measure for learning strategies is the Motivated Strategies for Learning Questionnaire (MSLQ) by Pintrich, Smith, Garcia, and McKeachie (1991). The MSLQ is a well-documented measure of motivation and metacognition, with consistently high reliability and validity. See Table 2 for a list of scales and sub scales. The MSLQ is comprised of 15 separate measures with 81 questions on a 7-point Likert scale. The MSLQ has been used in many SoTL studies and can be broken into subscales. Or you could use a SoTL scale developed by Gurung, Weidert, and

Jeske (2010) called the Study Behaviors Checklist (SBC). The SBC is a 35-item assessment with one general factor (i.e., study behaviors) that measures things like notetaking, highlighting, cramming, and practice testing. Gurung and colleagues report acceptable reliability statistics for the SBC, however, it has not been used in many studies and is undetermined how valid the instrument is.

Table 2
Pintrich et al. (1991) MSLQ Scales and Subscales

| Scale | Subscale |
|---|---|
| Motivation Scales | |
| *Value Components* | 1. Intrinsic Goal Orientation |
| | 2. Extrinsic Goal Orientation |
| | 3. Task Value |
| *Expectancy Components* | 4. Control Beliefs |
| | 5. Self-Efficacy for Learning and Performance |
| *Affective Components* | 6. Test Anxiety |
| Learning Strategies | |
| *Cognitive and Metacognitive Strategies* | 7. Rehearsal |
| | 8. Elaboration |
| | 9. Organization |
| | 10. Critical Thinking |
| | 11. Metacognitive Self0regulation |
| *Resource Management Strategies* | 12. Time and Study Environment |
| | 13. Effort Regulation |
| | 14. Peer Learning |
| | 15. Help Seeking |

Both the SBC and MSLQ are fine measures, but they are missing key elements of current learning strategy research. That is, in a lauded review by Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013), they describe some of the more current popular, effective and ineffective learning strategies and techniques studied today. These include (in order from high effectiveness to low effectiveness) practice testing, distributed practice, interleaved practice, elaborative interrogation, self-explanation, keyword mnemonic, visual imagery, rereading, summarization, and highlighting. If you look at the MSLQ and SBC, many of these strategies are missing. Additionally, consistent with my suggestion above, a new SoTL scale that assesses learning strategies should include behavioral measures as well. For instance, when asking about the use of self-explanation strategy, SoTL researchers should also ask, “How many minutes in a week do you use this strategy?” In sum, instead of the MSLQ and SBC, what is needed is an updated SoTL scale that looks at the use of current learning strategies and techniques students use or don’t use in our classrooms with complementing behavioral measures.

## Sasquatch and the Importance of Model Teaching Characteristics

Imagine Sasquatch does not roam the forest aimlessly as some surmise and instead is a great teacher who exhibits model teaching characteristics. How would we know she is a great teacher? In other words, how do we measure model teaching characteristics? In a recent Society of Teaching of Psychology presidential taskforce, Richmond and colleagues (2014) set out to define and delineate model teaching characteristics. In their research they determined that model teachers exhibit six model characteristics with 19 separate criteria. See Table 3 for a list of these characteristics.

Table 3
Richmond et al.'s (2014) Model Teaching Characteristics

| Model Teaching Characteristics | Model Teaching Criteria |
|---|---|
| Training | 1. Subject Knowledge |
| | 2. Pedagogical Knowledge |
| | 3. Continuing Education in Pedagogical Knowledge |
| Instructional Methods | 4. Pedagogy |
| | 5. Teaching Skills |
| Assessment Processes | 6. Student Learning Goals and Objectives |
| | 7. Assessment of Student Learning Outcomes |
| | 8. Reflection on Assessment |
| | 9. Scholarship of Teaching and Learning |
| | 10. Evaluation Directness |
| | 11. Evaluation Utility |
| Syllabi | 12. Course Transparency |
| | 13. Course Planning |
| Content | 14. Scientific Literacy |
| | 15. Psychology Knowledge Base and Application |
| | 16. Liberal Arts Skills |
| | 17. Values in Psychology |
| Students Evaluations of Teaching | 18. Student Feedback |
| | 19. Reflection on Student Feedback |

To further this line of research, Boysen, Richmond, and Gurung (2015) empirically investigated these model characteristics and their respective criteria. In so doing, they developed a 52-item SoTL scale designed to measure model teaching competencies. The 52-items were measured on a dichotomous (e.g., yes/no) scale. For example, to measure pedagogy, they asked, "Class observation records support effective use of effective instructional methods" (p. 51). Over 200 psychology teachers from across the country at all types of institutions of higher education (e.g., community colleges, research-focused universities, private colleges, etc.) participated in the study. Boysen and colleagues found that baseline data for this self-report scale correlated strongly with the TBC and the Big-5 Inventory of personality (Gosling, Rentfrow, & Swann, 2003). Additionally, they found that the scale had strong intercorrelations among the model teaching characteristics and criteria.

However, there were some issues with this scale. First, the 52-item inventory is long and cumbersome. Second, the nature of a dichotomous scale makes it difficult to assess the factor structure of the scale and conduct internal reliability analyses. As such, Boysen and colleagues did not conduct a factor analyses. Accordingly, SoTL researchers should consider changing the scale to a Likert-type scale, to truly investigate the efficacy of this measure. Third, although there is initial evidence of reliability and some validity, further research is needed. That is, SoTL researchers should investigate this SoTL scale with other measures of teaching known measures of teaching effectiveness (e.g., Approaches to Teaching Inventory by Trigwell & Prosser, 2004; or the Teaching Goals Inventory by Angelo & Cross, 1993). Fourth, as mentioned previously in this chapter, there needs to be behavioral measures which complement this scale (e.g., classroom observation records). In the end, model teaching characteristics are important to understand and as this is a new area of SoTL research, there is much room for improvement and exploration.

## A Call to Action! Debunking Cryptozoology and Focusing on the Missing Links of SoTL Measures

Whether it is the Yeti, Sasquatch, Bigfoot, or la Chupacabra (I just like saying that), the aim of SoTL research and SoTL scales is to demystify such urban myths and focus on the science of learning and teaching. Accordingly, how we measure SoTL is of utmost importance and drives at the heart of credibility of our field. Therefore, I summarily and humbly suggest that:

1. There needs to be more SoTL scale development that intentionally targets the efficacy of syllabi.
2. When using preexisting SoTL scales always include complementary behavioral measures. More so, when developing SoTL scales, consider including behavioral measures either in the form of checklists or scenario-based.
3. There needs to be a SoTL scale developed to measure metacognition and current use of effective learning strategies. Don't forget to include the behavioral component to these new measures.
4. There needs to be a SoTL scale that assesses meta-teaching skills. This is an untapped SoTL area ripe for the picking.
5. There needs to be an accurate, reliable, and valid measure of model teaching competencies.

Please, please, consider these suggestions as a call to action. If you choose to answer my call to action, I highly encourage you to refer to the outstanding chapters in this e-book to assist you in this process. For instances, refer to Regan Gurung's chapter (2015) on the best practices in SoTL scale use. Gurung suggests that in SoTL we should be "measuring the usual suspects" such as self-efficacy, metacognition (see there is safety in numbers), motivation, study behaviors (YES!), etc. Or, when writing items for a SoTL scale, use clear, concise, unambiguous, and gender neutral or culturally sensitive language. Also, take heed to Georgeanna Wilson-Doenges' (2015) chapter in which she explicates the state of SoTL scale validation. Here, Wilson-Doenges explains how to properly assess reliability and validity within the context and particularly nuanced nature of SoTL research and she explains how to develop SoTL scales in light of these issues. Please also read and consider Andrew Christopher's (2015) chapter on how to select the right SoTL scale. This chapter will provide some ways in which you can avoid common traps or pitfalls that SoTL scales have (e.g., growing divide between SoTL and the learning sciences).

Christopher (2015) also suggests that when creating SoTL scales, they should be specific to a set of behaviors, yet simple.

In the end, if you consider the great advice provided by these scholars, and take up my call to action, I know that you will create some outstanding SoTL scales that will advance our beloved scholarship of teaching and learning.

# References

References marked with an asterisk indicate a scale.

Angelo, T. A., & Cross, P. K. (1993). *Classroom assessment techniques: A handbook for college teachers.* San Francisco, CA: Jossey-Bass.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?. *Perspectives on Psychological Science*, *2*(4), 396-403. doi:10.1111/j.1745-6916.2007.00051.x

*Berry, J. M., West, R. L. & Dennehey, D. M. (1989). Reliability and validity of the Memory Self-Efficacy Questionnaire. *Developmental Psychology, 25*(5), 701-713. doi:10.1037/0012-1649.25.5.701

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate. Carnegie Foundation for the Advancement of Teaching.* Princeton, N.J.: Princeton University Press.

Boysen, G. A., Richmond, A. S., & Gurung, R. A. R. (2015). Model teaching criteria for psychology: Initial documentation of teachers' self-reported competency. *Scholarship of Teaching and Learning in Psychology, 1,* 48-59. doi: http://dx.doi.org/10.1037/stl0000023

*Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in feed for cognition. *Psychological Bulletin, 119*(2), 197-253.

Chen, X. (2013). Meta-teaching: Meaning and strategy. *Africa Education Review, 10*(1), S63-S74. doi:10.1080/18146627.2013.855431

Christopher, A. N. (2015). Selecting the right scale: An editor's perspective. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

*Cullen, R., & Harris, M. (2009). Assessing learner-centredness through course syllabi. *Assessment & Evaluation in Higher Education, 34*(1), 115-125.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

*Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education*, *37*(1), 40-53. doi:10.1080/03634529009378786

*Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528. doi:10.1016/S0092-6566(03)00046-1

Gurung, R. A. R. (2015). Best practices in scale use in SoTL. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

*Gurung, R. A., Weidert, J., & Jeske, A. (2010). Focusing on how students study. *Journal of the Scholarship of Teaching and Learning*, *10*(1), 28-35.

*Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33,* 84-91. doi:10.1207/s15328023top3302_1

Kirk, C., Busler, J., Keeley, J., & Buskist, B. (2015). Effective tools for assessing characteristics of excellent teaching: The teacher behaviors checklist as exemplar. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Ku, K. L. Y. (2015). Measuring individual differences in epistemological beliefs. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Landrum, R. E., & McCarthy, M. A. (2015). Measuring critical thinking skills. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Landrum, R. E., & Stowell, J. R. (2013). The reliability of student ratings of master teacher behaviors. *Teaching of Psychology, 40,* 300-303. doi:10.1177/0098628313501043

Layous, K., Nelson, S. K., & Legg, A. M. (2015). Measuring well-being in the scholarship of teaching and learning. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Marek, P., Williamson, A., & Taglialatela, L. (2015). Measuring learning and self-efficacy. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Matejka, K., & Kurke, L. B. (1994). Designing a great syllabus. *College Teaching, 42*(3), 115-117. doi:10.1080/87567555.1994.9926838

Meyerberg, J. M., & Legg, A. M. (2015). Assessing professor-student relationships using self-report scales. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Parkes, J., & Harris, M. B. (2002). The purposes of a syllabus. *College Teaching, 50*(2), 55-61. doi:10.1080/87567550209595875

*Pintrich, P. R, Smith, D. A., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Ann Arbor, Ml: National Center for Research to Improve Postsecondary Teaching and Learning.

Richmond, A. S., Becknell, J., Slattery, J., Morgan, R., & Mitchell, N. (2015, August). Students' perceptions of a student-centered syllabus: An experimental analysis. Poster presented the annual meeting of the *American Psychological Association*, Toronto, Canada.

Richmond, A. S., Boysen, G. A., Gurung, R. A. R., Tazeau, Y. N., Meyers, S. A., & Sciutto, M. J. (2014). Aspirational model teaching criteria for psychology. *Teaching of Psychology, 41,* 281-295, doi:10.1177/0098628314549699

Saville, B. K., Zinn, T. E., Brown, A. R., & Marchuk, K. A. (2010). Syllabus detail and students' perceptions of teacher effectiveness. *Teaching of Psychology, 37,* 186-189. doi:10.1080/00986283.2010.488523

*Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*, 460-475.

*Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, *24*, 48-57. doi:http://dx.doi.org/10.1016/j.learninstruc.2012.08.007

Schraw, G., Kuch, F., Gutierrez, A., & Richmond, A. S. (2014). Exploring a three-level model of calibration accuracy. *Journal of Educational Psychology, 106*(4), 1192-1202. doi:10.1037/a0036653

Scholarship of Teaching and Learning: History (n.d.). Retrieved from http://academics.georgiasouthern.edu/sotlgsu/history/

Shulman, L. S. (1993). Teaching as community property: Putting an end to pedagogical solitude. *Change: The Magazine of Higher Learning*, *25*(6), 6-7.

Simons, L. (2015). Measuring service-learning and civic engagement. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Slattery, J. M., & Carlson, J. F. (2005). Preparing an effective syllabus: Current best practices. *College Teaching, 53*, 159-164. doi:10.3200/CTCH.53.4.159-164

Slattery, J. M., Haney, M., Richmond, A. S., Venzke, B. Morgan, R. K., & Mitchell, N. (2014, August). *Project syllabus: Student responses to syllabi.* A symposium presented at the annual meeting of the American Psychological Association, Washington, D.C.

Spring, H. T. (1985). Teacher decision making: A metacognitive approach. *The Reading Teacher,* 290-295.

*Trigwell, K., & Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educational Psychology Review*, *16*(4), 409-424.

West, R. L., Thorn, R. M., & Bagwell, D. K. (2003). Memory performance and beliefs as a function of goal setting and aging. *Psychology and aging*, *18*(1), 111.

Wilson-Doenges, G. (2015). The state of scale validation in SoTL research in psychology. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.

Zabel, K. L., & Heger, A (2015). Student engagement toward coursework: Measures, considerations, and future directions. In R. S. Jhangiani, J. D. Troisi, B. Fleck, A. M. Legg, & H. D. Hussey (Eds.), *A compendium of scales for use in the scholarship of teaching and learning*.