Pragmatic Measurement for Education Science: A Method-Substance Synergy of Validation and Motivation

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfilment

Of the Requirements for the Degree of

Doctor of Philosophy

By

Jeff J. Kosovich, B.A., M.A.

May 2017

Abstract

Education researchers often require quick and efficient assessments of various student characteristics (e.g., motivation) to use in classroom settings. Unfortunately, guidelines for addressing practical measurement obstacles, such as scale length, are ambiguous at best and non-existent at worst. Many measures lack sufficient evidence that the conclusions they produce are merited, and short measures have received particular criticism from measurement experts. The result is a tension between technical and pragmatic constraints when conducting measurement in field research. This three-paper dissertation is aimed at identifying and addressing these tensions in one area of motivation research. Paper 1 provides the substantive frame for the overall dissertation. The goal was to understand short-term student motivation change in a classroom setting. Paper 2 provides a typical approach to assessing a scale's quality and viability for use in the field. The goal was to use traditional psychometric approaches to evaluate a brief measure of motivation. Finally, Paper 3 presents a pragmatic approach to determining validity evidence (i.e., pragmatic measurement) by considering the underlying uses and restrictions of collecting data. The goal was to evaluate the pragmatic approach as a framework for measure users to identify the relevant validity evidence needed based on the potential uses and interpretations of a measure. Together, these papers highlight the nature and benefit of advancing methodological goals by pursuing substantive goals. The current research is a methodological-substantive synergy (i.e., work that advances a substantive domain, such as motivation, while developing and utilizing state-of-the-art methodology) aimed alleviating technical and practical tensions.

The Education Leadership, Foundations and Policy
Curry School of Education
University of Virginia
Charlottesville, VA

APPROVAL OF THE DISSERTATION

This dissertation, Pragmatic Measurement for Education Science: A Method-Substance Synergy of Validation and Motivation, has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

____
Chris S. Hulleman

____
Sara Rimm-Kaufman

____
Robert C. Pianta

____
Karen Schmidt

____________________Date

TABLE OF CONTENTS

Dedication

To my mom, grandma, and grandpa for being rock stars, for giving me the best opportunities they could, and for all the sacrifices they made for me.

To my family, for all of their love and support.

To Cortney, mostly for putting up with me. Also for being awesome.

To all of my other friends, also for putting up with me and being awesome.

To all the other kids whose potential love of learning was stifled by our society's obsession with achievement and financial gain. Instead of curiosity and excitement we got complacency, anxiety, and fear. Our society did us a disservice, and I hope my work can be a small step in fixing that.

Acknowledgements

Thank you to Chris Hulleman for six years of mentoring, friendship, and all of the work that led to this dissertation. Thank you for helping me to accomplish one of the most important goals I've adopted in my life up to this point.

Thank you to Sara Rimm-Kaufman, Bob Pianta, and Karen Schmidt for agreeing to be on my committee and helping to take my last big step.

Thank you to Kenn Barron for the continued advice, mentoring, and support.

Thank you to Amanda Durik and Meredith McGinley for jump-starting this whole research journey.

Thank you to Jessica Flake and Rory Lazowski for commiserating, for collaborating, and for being great role models.

Thank you to Julie Phelps and Maryke Lee for taking a chance and working with us on a really phenomenal researcher-practitioner partnership.

Thank you to Karen Givvin and Rachel Beattie giving me the chance for such a unique graduate school experience.

LIST OF TABLES

Page

Paper 1

Paper 2

Paper 3

LIST OF FIGURES AND APPENDICES

Page

Study 1

Study 2

Study 3

Conceptual Link:

Pragmatic Measurement for Education Science: A Method-Substance Synergy of Validation and Motivation

Jeff J. Kosovich

University of Virginia

Conceptual Link: Pragmatic Measurement for Education Science: A Method-Substance Synergy of Validation and Motivation

Education researchers often require quick and efficient assessments of various student characteristics (e.g., motivation) to use in classroom settings. Unfortunately, guidelines for addressing measurement obstacles in practice, such as scale length, are ambiguous at best and non-existent at worst (e.g., Csikszentmihalyi & Larson, 2014; Deno, 1985; Stanton, Sinar, Balzer, & Smith, 2002; Yeager, Walton, & Cohen, 2013). Many measures may be suspect because of a general lack of evidence for the quality of the data they produce (Paulhus & Vazire, 2007), and short measures have received particular criticism from measurement experts (e.g., Widaman, Little, Preacher, & Sawalani, 2011). In addition, there is a growing call to action to improve the quality of validity evidence for measurement in theory, research, and practice (Flake, Pek, & Hehman, in press; Graham, 2015). Despite a robust body of research on validation (AERA, APA, & NCME, 2014; Cronbach & Meehl, 1955; Kane, 2013; Messick, 1989), best practices in measurement are not widely disseminated due to a shortage of methodologists and insufficient methodological training (Aiken, West, & Millsap, 2008). The result is a tension between technical and pragmatic constraints when conducting measurement in field research (Yeager, Bryk, Muhich, Hausman, & Morales, 2013). By developing measures that validly and reliably capture constructs of interest in classrooms, we can address the tensions between the technical and pragmatic aspects of measurement. Thus, the current dissertation uses an adaptive approach to measurement for educational research on student motivation with the goal of alleviating the technical and pragmatic tensions.

The context in which individuals are measured and the intended use and interpretation of the measure plays a role in determining what evidence is appropriate for a particular situation (AERA et al., 2014; Kane, 2013). For example, compare the intended uses and contexts of the Graduate Record Exam (GRE) to a daily in-class quiz. The goal of the GRE is to assess students' aptitude for graduate level education at a very precise level. The goal of the in-class quiz is to obtain a rough idea of students' knowledge about a topic in an efficient manner. For both situations, the purpose of the measure and the contexts for measurement play a role in determining how the measure is designed. Because the central goal for the GRE is precision, the test contains many questions and covers a broad scope of material. In contrast, the main goal of the quiz is to inform instruction in real time, and therefore it has fewer items on a few key points. In fact, it is unlikely that a teacher would expect to (or even be able to) administer a GRE-level assessment as a daily in-class quiz. Thus, failing to consider the assessment environment can lead to a poor response rate (Dillman, Smyth, & Christian, 2014) or poor data quality (Wise & Smith, 2011) and undermine the measurement goals. The solution is an adaptable approach to measurement that explicitly considers both the technical qualities and the practical obstacles.

In the current research we present *pragmatic measurement*, by adapting an argument-based validation framework (e.g., Kane, 2013). The pragmatic measurement framework considers intended use, substantive theory, and contextual obstacles (e.g., limited time) to identify necessary evidence for a measure of student motivation. We characterize this endeavor as a *methodological-substantive synergy*. Methodological-substantive synergies are described as the use of state-of-the-art methods, such as measure modern validation, to overcome substantive issues, such as understanding how student motivation changes over time (Marsh & Hau, 2007). Although the three papers address motivation research specifically, the dissertation as a whole furthers serves as a methodological starting point for the pragmatic measurement framework.

Paper 1 of this dissertation (Kosovich, Flake, & Hulleman, 2017) provides a substantive frame for the research of interest. This paper focuses on the advancement of theory, and sets up the tensions between technical and practical measurement within the substantive field of motivation research. We use a brief measure of two motivation constructs (*expectancies* for success and task *value*) to examine the simultaneous growth of student motivation during a single semester of introductory psychology in college. We suggest that future research could provide more insight on this issue if motivation were measured more frequently. However, we also note that more frequent measurement is likely to be cumbersome in a classroom setting. The result is a need for the development of valid and reliable measures that minimize classroom disruption. **Paper 2** (Kosovich, Hulleman, Barron, & Getty, 2015) uses a typical validation approach (i.e., factor analysis) to examine the Expectancy-Value-Cost Scale. We assessed three qualities of the scale: (a) if the scale can be deployed efficiently, (b) if scale scores can be interpreted at the observed level without advanced statistical modeling (i.e., latent variable modeling), and (c) if the scale can accurately show differences across groups (e.g., gender) and over time. This paper represents a common approach to validation that prioritizes technical concerns (i.e., the ability to conduct factor analyses). However, practical concerns are no less important. For example, analyses in Paper 2 suggest some of items may be redundant—a quality that may lead students to feel their time is being wasted and reduce response rates or quality. Thus, a validation approach that prioritizes both technical and practical concerns is needed.

Paper 3 introduces the pragmatic measurement perspective as an approach to alleviating the technical-practical tensions in measurement. Given the logistical constraints of frequent measurement in classroom settings, we examine validity evidence for further shortening the expectancy-value-cost scale. More specifically, we select validation procedures that fit the scale being assessed, rather than change the scale so that particular validation procedures can be used. We use an argument-based validity approach to identify the intended uses of expectancy-value-cost measures and the assumptions that underlie those uses. By identifying what assumptions are made about the scale, we can compile validity evidence that those assumptions are being met.

In combination, these three papers define the conceptual space around short-term motivation assessment in classrooms and examine approaches for optimizing a scale for use in that setting. The culmination is a method-substance synergy that balances best-practices of scale validation with applied research demands so that substantive knowledge can be developed.

Paper 1: Short-term Motivation Trajectories: A Parallel Process Model of Expectancy-Value

The expectancy-value-cost framework of achievement motivation (Barron & Hulleman, 2015; Eccles et al., 1983) is a prominent theory of student motivation in educational research that examines motivation development across the academic career. As its name suggests, there are three components that most proximally determine students' achievement and achievement-related choices. Expectancy is an individual's perception that they can succeed at a task. Value is the perceived enjoyment, importance, and/or usefulness an individual ascribes to a task. Cost is perceived barriers (psychological or otherwise) to succeeding at a task. All three constructs are related to important outcomes such as domain interest or course performance (Eccles et al., 1983; Hulleman, Durik, Schweigert, & Harackiewicz, 2008; Perez, Cromley, & Kaplan, 2014).

Despite motivation's importance, our knowledge of expectancy-value-cost trajectories is limited. The literature shows that student motivation in reading and math declines from year to year (Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). However, there are few published studies on short-term motivation change, and those that contain at least two time points of motivation data are not focused on examining short term change (e.g., Bong, 2001). Given one course can be the difference between obtaining a degree or not, particularly in U.S. two-year colleges (Goudas & Boylan, 2012), understanding short-term motivational dynamics is an important endeavor for researchers and non-researchers. This point is highlighted by motivational interventions showing efficacy to improve student outcomes through both value and expectancy (e.g., Hulleman & Harackiewicz, 2009; Hulleman, Kosovich, Barron, & Daniel, 2016). Thus, in Paper 1, *Short-term Motivation Trajectories: A Parallel Process Model of Expectancy-Value* (Kosovich et al., in press), we examine short-term change in expectancy and value over the course of a single semester in introductory college psychology. Data were collected three times using brief scales of student's expectancy and value motivation.

Several finding emerged from this study. First, expectancy and value showed decreasing trajectories over the course of the semester, mirroring the trends found in long-term change models. Whereas expectancy trajectories demonstrated substantial variability among students, value trajectories were relatively uniform in their direction and intensity. Despite variability among individual trajectories, value change was strongly related to change in expectancy (*r* = .80). Second, expectancy change predicted continuing interest, but neither expectancy change nor value change was related to end-of-semester exam performance. This study extends our understanding of motivational trajectories in the short term, showing that even within a semester motivation appears to decline over time. However, the data used cannot answer more fine-grained questions regarding the motivation dynamics. Are there points in the semester where motivation sharply increases or decreases? How does cost factor into these trajectories?

This paper sets the foundation for this particular program of research by exploring short-term motivational trajectories in the classroom. It highlights the need for more in-depth study of the constructs of interest, particularly through the use of more frequent measurement. In considering the next phase of the substantive research agenda (i.e., a more detailed assessment of motivation over time), we identify validation concerns regarding the measurement tools being used. In regard to frequent data collection, it is possible that overburdening students with too many questions may invalidate the data gained (Galesic & Bosnjak, 2009; Wise & Smith, 2011), and introduce concern of greater non-response bias (Dillman et al., 2014). In regard to including the additional measure of cost, the construct has only recently received attention in the literature and most of the previously published scales have far more items than would be practical for a frequent data collection design. These two concerns coupled with the general lack of published validity evidence for expectancy-value-cost scales necessitate more validation work to identify viable measures for continuing the substantive motivation inquiry.

Paper 2: A Practical Measure of Student Motivation: Establishing Validity Evidence for the Expectancy-Value-Cost Scale in Middle School

Paper 2 examines the psychometric structure of a brief measure of student motivation, the Expectancy-Value-Cost Scale. We use a common methodological approach (i.e., factor analysis) to ascertain whether or not the scale can be used for specific research goals such as comparing groups of students and tracking motivation over time. The Expectancy-Value-Cost Scale is an attempt to have brief subscales of expectancy (3 items), value (3 items), and cost (4 items) that are psychometrically sound and theoretically consistent. Prior validation work on expectancy-value-cost instruments is sparse in the literature. For example, most expectancy-value-cost studies cite one or two papers (Eccles & Wigfield, 1995; Wigfield, Harold, & Blumenfeld, 1993) regarding the scale validity, if they report anything at all. In addition, cost has been largely neglected in the research literature until recently (e.g., Barron & Hulleman, 2015; Flake, Barron, Hulleman, McCoach, & Welsh, 2015; Wigfield, Rosenzweig, & Eccles, in press). Cost is a particularly important aspect to consider as there is debate about whether it is a part of the value construct, or separate.

Many steps are involved in the validation process of a new scale (AERA et al., 2014; Benson, 1998; Schmeiser & Welch, 2006), and the initial validation work on the Expectancy-Value-Cost Scale is presented elsewhere (Hulleman et al., in prep). Paper 2 focuses on the structural validity of the Expectancy-Value-Cost Scale, particularly whether or not it is able to effectively measure differences across time, across academic domains, and across gender. In addition, we tested the practical question of whether the observed scores could yield the same conclusions as latent scores obtained through advanced statistical methods. We tested both of these questions using confirmatory factor analysis (CFA). Overall, the paper demonstrates that expectancy, value, and cost are separate factors, and that in most of the circumstances examined these measures can accurately capture differences across gender, domain, and time without requiring significant time during a class (M = 4.00 minutes, SD = 0.97 minutes).

In spite of these positive validation results, the Expectancy-Value-Cost Scale may elicit unexpected consequences among students who are more sensitive to time constraints (e.g., college students). For example, model diagnostics suggested that some items may be redundant. This concern was further confirmed in independent work with community college students using interviews, qualitative inquiry, and Item Response Theory (Kosovich, unpublished data). Given that the goal is to eventually measure student motivation frequently throughout a semester, the presence of redundant items suggest that shorter scales are possible and that students may be more resistant to repeated participation. However, further reducing the length of the Expectancy-Value-Cost subscales introduces some concern from the commonly held views of measurement when fewer items are used, such as reduced reliability and construct representation. Thus, Paper 3 is a validation study for briefer measures of expectancy-value-cost to balance technical and practical demands (i.e., pragmatic measurement).

Paper 3: Pragmatic Measurement: Using Argument-Based Validation in Applied Education Sciences

Scales that are too long or too short can be a source of concern in regard to their validity for use or interpretation, respectively (e.g., Csikszentmihalyi & Larson, 2014; Deno, 1985; Stanton et al., 2002; Yeager, Walton, et al., 2013). From the technical side, experts question the quality of short scales for many reasons (e.g., Stanton et al., 2002; Wanous & Hudy, 2001; Widaman et al., 2011). For example, calculating reliability becomes more difficult as the number of items decreases (e.g., Traub & Rowley, 1992), and single-item reliability requires advanced statistical procedures. In contrast, longer assessments can undermine measurement quality due to respondent motivation (as seen in achievement testing, e.g., Sundre & Kitsantas, 2004) and by being disruptive to educational practice (Yeager, Bryk, et al., 2013). Thus, a minimalist Expectancy-Value-Cost Scale requires both substantive and methodological work. Modern validation theory (e.g., Kane, 2013) recommends argument-based validation focused on the intended uses and interpretations of a measure. Paper 3 is designed to explore methods for compiling validity evidence for short (1-item) subscales, and to use those methods for validation of Expectancy-Value-Cost Scale items.

We identified four common uses of expectancy-value-cost measures, their underlying assumptions, and corresponding validity evidence. We cross-validated the four uses in a second dataset that was drawn from the same population. The first use was measuring expectancy, value, and cost in a classroom setting. We argue that measures which reproduce established relationships demonstrate reliability. To validate this use, we examined reliability in the form of inter-construct correlations and prediction, as well as item-total correlations. Overall, evidence strongly supported the expectancy and value subscales, whereas the cost scales showed somewhat lower reliability and potential multidimensionality. Importantly, we were able to reduce scale length by 10 items to 3 items (one per scale) while only experiencing a decrease in variance explained of 4%.

These analyses generally supported the first use and also formed a foundation of validity evidence for the latter uses.

The second use focused on examining known-group differences, particularly for female students' expectancies. We argue that measures which reproduce established group differences show further validity evidence. To validate this use, we examined female students' expectancy and pass rates in relation to males' expectancy and pass rates. As would be suggested by prior research, women reported lower expectancy despite displaying higher achievement than men. This analysis provided additional validity evidence for the measures and groundwork for the third use.

The third use was to monitor motivation change over time. We argue that measures which reproduce established trends over time are likely sensitive to natural changes in the environment to a similar degree as prior measures. To validate this use, we computed latent growth models to assess the trajectories of students' motivation. As seen in prior research, expectancy and value decreased over time whereas cost increased over time. The single-item measures were generally able to capture individual construct trajectories, but the relationships between those trajectories were less consistent. We suggest that such latent correlations may be a boundary for the usage of pragmatic measures. However, we also hypothesized that the inconsistency may again be due to multidimensionality. Overall, evidence suggested that the scales could be used monitoring change and we ultimately pursued evidence for the fourth use.

The fourth use was to monitor intervention processes. We argued that measures highly-aligned with the intervention should be sensitive to intervention effects in the classroom. To validate this use we examined the effects of a value intervention on the value scale., Supplemental analyses suggested that the intervention was not effective. As a result, it was not clear if the value scale was able to detect changes caused by the intervention. However, the data shortened scales did seem to corroborate the trends shown by supporting motivation measures. Though inconclusive for the fourth use specifically, the results of the intervention analyses support the need to use several measures to assess critical outcomes.

Summary

This three-paper dissertation is method-substance synergy aimed at developing methods for overcoming measurement obstacles for educational researchers. Paper 1 provides the substantive frame for the overall dissertation. The goal was to understand short-term student motivation change in a classroom setting. Paper 2 uses a common approach to assessing a scale's quality and feasibility for use in the field. The goal was to assess test the psychometric quality of the expectancy-value-cost scale for comparing groups and tracking change over time. Finally, Paper 3 presents a pragmatic approach to determining validity evidence (i.e., pragmatic measurement) by considering the underlying uses and restrictions of collecting data. The goal is to evaluate the pragmatic approach for measure users to identify potential uses and interpretations of a measure and to identify relevant evidence needed. Together, these papers highlight the nature and benefit of advancing methodological goals by pursuing substantive goals.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno’s (1990) survey of PhD programs in North America. *The American Psychologist*, *63*(1), 32–50. http://doi.org/10.1037/0003-066X.63.1.32

Barron, K. E., & Hulleman, C. S. (2015). Expectancy-Value-Cost Model of Motivation. In *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 8, pp. 503–509). Elsevier. http://doi.org/10.1016/B978-0-08-097086-8.26099-6

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, *17*(1), 10–17. http://doi.org/10.1111/j.1745-3992.1998.tb00616.x

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task value, and achievement goals. *Journal of Educational Psychology*, *93*(1), 23–34. http://doi.org/10.1037/0022-0663.93.1.23

Cronbach, L. J., & Meehl, P. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology* (pp. 35–54). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-017-9088-8_3

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Eccles, J. S., & Wigfield, A. (1995). In the Mind of the Actor: The Structure of Adolescents' Achievement Task Values and Expectancy-Related Beliefs. *Personality and Social Psychology Bulletin*, *21*, 215–225. http://doi.org/10.1177/0146167295213003

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology*, *41*, 232–244. http://doi.org/10.1016/j.cedpsych.2015.03.002

Flake, J. K., Pek, J., & Hehman, E. (in press). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*.

Galesic, M., & Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, *73*(2), 349–360. http://doi.org/10.1093/poq/nfp031

Goudas, A. M., & Boylan, H. R. (2012). Addressing flawed research in developmental education. *Journal of Developmental Education*, *36*(1), 2-4-13.

Graham, S. (2015). Inaugural editorial for the Journal of Educational Psychology. *Journal of Educational Psychology*, *107*(1), 1–2. http://doi.org/10.1037/edu0000007

Hulleman, C. S., Durik, A. M., Schweigert, S. B., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, *100*(2), 398–416. http://doi.org/10.1037/0022-0663.100.2.398

Hulleman, C. S., Getty, S., Barron, K. E., Lazowski, R. A., Ruzek, E., & Stuhlsatz, M. (n.d.). Validating a rapid measure of expectancy-value-cost motivation in high school science.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science (New York, N.Y.)*, *326*(5958), 1410–1412. http://doi.org/10.1126/science.1177067

Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2016). Making Connections: Replicating and Extending the Utility Value Intervention in the Classroom. *Journal of Educational Psychology*. http://doi.org/10.1037/edu0000146

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children’s self-competence and values: gender and domain differences across grades one through twelve. *Child Development*, *73*(2), 509–527. http://doi.org/10.1111/1467-8624.00421

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

Kosovich, J. J., Flake, J. K., & Hulleman, C. S. (n.d.). Understanding Short-term Motivation Trajectories: A Parallel Process Model of Expectancy-Value.

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2015). A Practical Measure of Student Motivation: Establishing Validity Evidence for the Expectancy-Value-Cost Scale in Middle School. *The Journal of Early Adolescence*. http://doi.org/10.1177/0272431614556890

Marsh, H. W., & Hau, K. T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology*, *32*(1), 151–170. http://doi.org/10.1016/j.cedpsych.2006.10.008

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11. http://doi.org/10.3102/0013189X018002005

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York: Guilford Press.

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, *106*(1), 315–329. http://doi.org/10.1037/a0034027

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167–194. http://doi.org/10.1111/j.1744-6570.2002.tb00108.x

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-talking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6–26. http://doi.org/10.1016/S0361-476X(02)00063-2

Traub, R., & Rowley, G. (1992). Understanding Reliability. *Instructional Topics in Educational Measurement*. http://doi.org/10.1111/j.1745-3992.1991.tb00183.x

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In *Secondary data analysis: An introduction for psychologists*. (pp. 39–61). http://doi.org/10.1037/12350-003

Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and Gender Differences in Children ' s Self- and Task Perceptions during Elementary School. *Child Development*, *64*(3), 830–847.

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. *High-Stakes Testing in Education: Science and Practice in K–12 Settings*, 139–153. http://doi.org/10.1037/12330-009

Yeager, D. S., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). Practical measurement. *Palo Alto, CA: Carnegie Foundation for the Advancement of Teaching*, *78712*. JOUR. Retrieved from http://www.carnegiefoundation.org/wp-content/uploads/2014/09/Practical_Measurement_Yeager-Bryk1.pdf

Yeager, D. S., Walton, G., & Cohen, G. L. (2013). Addressing Achievement Gaps with Psychological Interventions. *Phi Delta Kappan*, *94*(5), 62–65. http://doi.org/10.1177/003172171309400514

Paper 1: Short-term Motivation Trajectories: A Parallel Process Model of Expectancy-Value

Jeff J. Kosovich*

University of Virginia

Jessica K. Flake*

York University

Chris S. Hulleman

University of Virginia

Author Note

*Authors contributed equally to the work, corresponding author is listed first.

Abstract

Motivation plays a critical role in human behavior and is particularly important during college, where a single class can make or break an academic career. The longitudinal research on expectancies for success and utility value primarily focuses on prediction or change over many years, rather than change over a short period of time. However, a single class in college can often be the difference between getting a degree or not. To better understand how motivation progresses in the short-term, we examined changes in expectancy and utility value simultaneously during a single college class. Both constructs declined during the class and showed significant variability across individuals. In addition, change in expectancy was strongly correlated with change in utility value, and the expectancy slope estimates were significant predictors of continuing interest. We discuss the need for a better understanding of short-term dynamic relationships between expectancies, utility value, and outcomes.

Highlights

- Empirical, published research on short-term expectancy-value change is rare.
- Short-term expectancy-value change mirrors long-term (downward) trends.
- Expectancy trajectories are strongly correlated with utility value trajectories.
- Expectancy trajectories, but not utility value trajectories, predict continuing interest.

Keywords: Academic Achievement; Educational Psychology; Longitudinal Methodology; Motivation; Structural Equation Modeling; Dual-process Models

1. Short-term Motivation Trajectories: A Parallel Process Model of Expectancy-Value

Short-term events can have lasting consequences in education. In college, a single class can be the difference between getting a degree or not (Goudas & Boylan, 2012). For example, gatekeeper courses (i.e., foundational courses that are necessary for completion of a degree, Atanda, 1999) can single-handedly stall educational progress rather than enhance it. Negative experiences may prevent students from graduating at all (Bailey, Jeong, & Cho, 2010; Silva & White, 2013). Alternatively, positive experiences may lead to greater likelihood of pursuing a particular domain (Harackiewicz, Barron, Tauer, & Elliot, 2002) or career (Pike & Dunne, 2011). Given the positive effects that degree attainment can have on life outcomes (Bureau of Labor Statistics, 2015), it is necessary to understand the short-term dynamic processes that lead to academic success or failure. In the current study, we approach these dynamic processes from a motivational perspective by building on prior educational research.

Motivation is a critical predictor of academic achievement and engagement, particularly in higher education (Lazowski & Hulleman, 2016; Robbins et al., 2004). As a result, it is paramount that we have an understanding of how motivation changes over short periods of time, such as over the duration of a single class. Expectancy-value models of motivation have been influential for understanding motivation in educational settings (e.g., Eccles et al., 1983). However, the bulk of expectancy-value motivation research that incorporates longitudinal designs within the educational context focuses on long-term change with yearly time points (Wigfield & Eccles, 2002). Existing research indicates that expectancies and values decline steadily over the academic career (Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Musu-Gillette, Wigfield, Harring, & Eccles, 2015). However, it is less clear how that change occurs and what factors influence it in the interim. The majority of research also examines the relation between expectancy and value at a single time point, rather than how the constructs change in tandem with one another. Finally, research indicates that expectancy-value motivation is influenced by prior achievement from year to year (Eccles et al., 1983), yet it is unclear how motivation change relates to proximal outcomes of an interim goal (e.g., passing a single class). The current paper extends the literature by testing a process model of expectancy and value that incorporates a short-term timescale and allows for a more complete picture of the dynamics between motivation and achievement. In the current study, we constrain our discussion to the utility value component of as several studies demonstrate that it can be leveraged to improve academic outcomes such as GPA and interest (Brown, Smith, Thoman, Allen, & Muragishi, 2015; Hulleman, Godes, Hendricks, & Harackiewicz, 2010; Hulleman & Harackiewicz, 2009).

2. A Brief Overview of Expectancy-value Models

Expectancy-value models of motivation were developed in social psychology (Atkinson, 1964; Lewin, Dembo, Festinger, & Sears, 1944) and business management fields (Vroom, 1964) and later adapted to education (Eccles et al., 1983). According to Eccles' (1983) expectancy-value model of achievement motivation, the most proximal determinants of achievement behavior are expectancies for success and subjective task values, including utility value. Expectancy refers to individuals' confidence in their ability to succeed at a task, and comprises effort and ability beliefs (Eccles et al., 1983). Generally, findings indicate that expectancy is most strongly related to academic achievement, including grades, course taking, activity participation, academic standing, and later expectancy (e.g., Durik, Vida, & Eccles, 2006; Eccles, Vida, & Barber, 2004; Hulleman, Kosovich, Barron, & Daniel, 2016; Simpkins, Davis-Kean, & Eccles, 2006). Value refers to the importance, usefulness, or enjoyment an individual associates with success on a task. It is comprised of four components (Eccles & Wigfield, 1995): attainment value (the importance of the task to one's self), intrinsic value (the interest or enjoyment gained from the task), utility value (the usefulness of a task to one's goals), and cost (sacrifices or negative emotions related to the task).

Expectancies and value are positively correlated (Wigfield & Eccles, 2000), with the magnitude of the relation ranging from small (Finney & Schraw, 2003) to large (Papaioannou, Marsh, & Theodorakis, 2004) depending on temporal proximity, domain overlap, and group characteristics (e.g., Kosovich, Hulleman, Barron, & Getty, 2015). Expectancy, value, the context, and the student experience inform each other (Perez, Cromley, & Kaplan, 2014) and students' performance (Flake, Barron, Hulleman, McCoach, & Welsh, 2015) as time progresses.

2.1. The Importance of Interest in the Expectancy-Value Model

The expectancy-value model (Eccles et al., 1983) accounts for achievement behaviors and choices, and also describes many factors that lead to lasting motivation in the form of interest. Interest refers to a psychological state that can develop into an enduring predisposition to re-engage with a particular task, topic, or content (Hidi & Renninger, 2006). Interest is an important aspect of the achievement experience that is influenced by components of the expectancy-value model (Eccles et al., 1983; Hidi & Renninger, 2006). To understand how expectancies and utility value relate to interest, it is helpful to compare and contrast utility value and interest. For example, Hidi and Renninger's (2006) four-phase model of interest development notes that interest is more likely to be maintained (i.e., Phase 2) if the topic or task is perceived as meaningful or useful. At the same time, expectancy pervades every level of interest development and can theoretically enhance or undermine interest (Eccles, Fredricks, & Epstein, 2015). Thus, interest should be positively correlated with expectancy and utility value trajectories. Individuals with higher expectancy-value slopes should also report higher interest.

Interest is also a powerful predictor of individuals' persistence and career choices. For example, Harackiewicz, Barron, Tauer, and Elliot (2002) found that interest in psychology during freshman year is a stronger predictor of college major six years later than GPA or prior performance. Qualitative research on scientists' recollections of their career paths also invokes interest as a major drive of decisions (Pike & Dunne, 2011). Given the propensity for expectancy-value constructs to lead to greater continuing interest in a particular domain (Hulleman et al., 2010; Hulleman & Harackiewicz, 2009), interest is an important construct to consider in relation to expectancy and utility value growth. However, the lack of information on short-term expectancy and utility value change makes it difficult to know how the two constructs relate to subsequent interest.

2.2. Motivation Development and Conceptualizing Trajectories

The seminal work on expectancy-value motivation in education focused on motivation development. Research in this area has since emphasized development with longitudinal methodologies. Early research demonstrated many of the factors that contribute to students' expectancy and value at a given point in time. For example, Eccles and colleagues (Eccles et al., 1983) used path models to examine predictors of expectancy and value as well as their ability to predict intentions for further course-taking. Later research (Simpkins et al., 2006) also examined reciprocal relationships between expectancy, value, and grades. In both early and later studies, researchers also examined the change in motivation over the academic career (e.g., Eccles et al., 1983; Jacobs et al., 2002). Generally speaking, expectancy-value motivation and similar constructs declines over the long-term (Durik et al., 2006; Eccles et al., 1983; Marchand & Gutierrez, 2016; Perez et al., 2014; Ren, 2000); whether this downward trajectory is a consistent decline or not in the short-term is unclear.

Though there is a history of longitudinal work in this area, the existing research may not extend to short time intervals. One problem is the tendency towards predicting expectancy-value constructs or using expectancy-value constructs to predict outcomes with single time point or pre-post designs (Durik et al., 2006; Eccles et al., 1983; Gillet, Berjot, Vallerand, & Amoura, 2012; Jacobs et al., 2002; Simpkins et al., 2006; Ullrich-French & Cox, 2014). The limitation of prediction designs is that the models focus on differences between individuals (i.e., inter-individual change), rather than construct change within individuals (i.e., intra-individual change). Even studies that do calculate pre-post differences between two time points do not necessarily describe the constructs' change, as some researchers argue that true change requires at least three time points to be effectively modeled (e.g., McArdle, 2009). Of the existing research that does include three or more time points, the focus has been on long-term change over several years, rather than storm-term change (Bandura & Schunk, 1981; Chularut & DeBacker, 2004; Hanus & Fox, 2015; Hidi, Berndorff, & Ainley, 2002; Luzzo, Hasper, Albert, Bibby, &

Martinelli, 1999; Moely, McFarland, Miron, Mercer, & Ilustre, 2002). These obstacles highlight the gap between existing longitudinal research and the current study. Although existing research findings do not necessarily extend from the long-term to the short-term, the underlying methods can be adopted relatively easily.

There are many methods available to understand motivation processes in the short-term, and different methods can answer different research questions. The number of methods available for researchers to assess change also increases with the number of data points collected. Two common approaches to assess change with such data are repeated measures ANOVA (e.g., Field, 2013) and latent growth modeling (e.g., Duncan, Duncan, & Strycker, 2013). Both approaches can produce estimates of the intra-individual change by capturing differences over time in terms or rates of change. However, latent growth modeling is being used more frequently and offers numerous advantages as a more flexible framework. This flexibility can accommodate different functional forms of change (e.g., cubic) and be easily expanded to include proximal and distal outcomes.

Frameworks such as ANOVA or path analysis conceptualize change as a series of mean-difference tests or as a series of chronologically ordered predictive models. With latent growth modeling the available data are used to construct a trajectory of the construct(s). The estimation of growth parameters such as the starting point (i.e., intercept) and how the construct changes over time (i.e., slope) allows researchers to extrapolate the constructs' progression (Chou, Bentler, & Pentz, 1998; Raudenbush & Bryk, 2002). Such models convey a more developmental picture of the constructs under consideration. These growth models also allow researchers to consider inter-individual change by estimating the variance of individual trajectories. As such, growth modeling can illuminate how experiences or characteristics are related to construct development (Raudenbush & Bryk, 2002) rather than test if prior experiences (e.g., prior performance, parental attitudes) explain variation in expectancy-value at a single point in time,.

Although studies that utilize growth curve modeling to understand short-term changes in expectancy and value are rare, there are other applications of growth curves that display the potential of the models to answer new and different kinds of research questions. For example, several studies have found that over the long term, expectancy and value tend to decrease on average, but the rate of change varies across people (Chouinard & Roy, 2008; Fredricks & Eccles, 2002; Jacobs et al., 2002). Further, Fredricks and Eccles (2002) incorporated predictors into their models to explain the variability in trajectories and found that when parents' have positive perceptions of their children's ability, the decline of students' ability beliefs is less extreme. In another example, Archambault, Eccles, and Vida (2010) found that different trajectory patterns characterized different groups of students by combining growth curve modeling with mixture modeling (Muthen & Muthen, 2000).

In the current study, we utilize *parallel process models* within the latent growth curve framework, which allow for capturing growth in multiple constructs simultaneously (Byrne, 2012; Muthén & Muthén, 2012). The added benefit of parallel process models (also referred to as simultaneous growth models) is that researchers can assess the covariance between intercepts and slopes within and across constructs. For example, researchers could test if expectancy and utility value change is positively related, or if a person's baseline expectancy is related to how their utility value changes. Given that expectancy and value are only moderately correlated at any time point, how one construct fluctuates during a semester may be related or unrelated to how the other fluctuates. Only a few studies examine expectancy-value change through the growth-modeling perspective (Archambault et al., 2010; Chouinard & Roy, 2008; Fredricks & Eccles, 2002; Jacobs et al., 2002) and we know of no previous studies explicitly investigating if expectancy change corresponded with utility value change in the short-term.

2.3. The Current Study

The current study was designed to extend research on short-term expectancy and utility value change in an introductory psychology classroom. Although previous literature on expectancy-value change may shed some light on what we would expect to see in the short-term, we recognize that introductory and gatekeeper courses are likely to be novel to students. As a result, many of the existing theoretical explanations used to justify motivational trends may not apply. For example, observed changes over time may be larger in the current study because students are learning about the material for the first time and have fewer prior experiences with the subject. College students are also more likely to be concerned with degree completion as well as job prospects than younger students. In recognition of these gaps in the research literature, we seek to answer three research questions:

(1) What are initial levels of expectancy and utility value in a college class, how do they change over the course of the class, and is there significant variability in these trajectories across participants? Based Eccles’ expectancy-value model as well as prior empirical research, we hypothesize intra-individual expectancy and utility value will start out relatively high on average and decline over time. On an inter-individual level, we hypothesize that there will be substantial variability in growth among individual students—that is, some students will do well in the course and learn that it is related to their goals and interests, whereas other students will do more poorly and decline in motivation.

(2) If there is variability in how expectancy and utility value change within a semester, how is change in in the two constructs are related? We hypothesize that expectancy and utility value change will show similar patterns of change. Due to the high-stakes nature of introductory college courses (e.g., they are often required for graduation, they may be gatekeepers to specific majors), we expect a moderate relationship between the two constructs as students either develop a greater appreciation for the domain or choose to pursue other domains in which they've experienced greater success.

(3) What are the relationships between changes in expectancy and utility value and student outcomes (e.g., performance and interest) during the course of a semester? Exam scores are likely a major driver of shifts in student motivation at the college level because poor performance is not likely to encourage continued pursuit of the topic. Thus, we would expect exam scores to predict expectancy trajectories and utility value trajectories as students gain familiarity with the material. Similarly, we would expect expectancy and utility value trajectories to be related to whether or not students intend to pursue the topic in future course (i.e., continuing interest). Though we had limited access to demographic information in these data, we also included gender as a covariate, which is consistent with prior research.

3. Method

3.1. Participants & Procedures

Participants (N = 389; 51% female) were enrolled in an introductory psychology course at a large, mid-western public university. Ethnicity and age information was not collected for individual students; however, the college population was 74.2% White, 5.4% Asian American, 2.9% African American, 3.3% Latino, 8.9% international, and 5.3% other[1]. The class was primarily first year students (M_{age} = 18.1 years old). All students in the class were eligible to participate.

Students provided responses for expectancy and utility value measures at three time points within a 15-week semester: week 3 (Time 1), week 9 (Time 2), and week 14 (Time 3). These time points were spaced to be far enough apart to capture change and to cohere with the course and exam schedule. Questionnaires were filled out by students in class and their responses were entered by researchers.

3.2. Missing Data

Missing data were present during each wave of collection (M = 14%). Wave one was missing approximately 24% of registered students, wave two was missing approximately 39% of students, and wave three was missing approximately 40% of students. Typically, students who were missing one measure were also missing the other measure within any give wave of data collection (e.g., they were not in class the day of the data collection). The correlation between expectancy missingness and utility value missingness was high at wave one (*r* = .97, n = 389), wave two (*r* = 1.00, n = 389), and wave three (*r* = .98, n = 389) suggesting that differences in missing data were uncommon

[1] Demographic information was obtained from the institutions' offices of institutional research and registrar.

within each wave. However, all exam scores were correlated with missingness in each wave. For example, exam one was correlated with missing utility value at wave one (*r* = -.23; n = 358), wave two (*r* = -.33, n = 358), and wave three (*r* = -.40, n = 358). Similarly, measures of motivation were highly correlated with each other across time. We used full information maximum likelihood (FIML) to account for missing data, and exam scores and gender were used in the unconditional model as auxiliary variables (Enders, 2010). The presence of exam scores and repeated measures can reduce bias in the estimates due to missingness; however it is impossible to ensure that bias is completely removed.

3.3. Measures

3.3.1. Expectancy and utility value.

Items for this study were adapted from Eccles and colleagues (Eccles et al., 1983) and worded to capture students' levels of general expectancy (5 items; = .88; e.g., "I expect to do well in this class") and utility value in the class (3 items; = .87; "What I am learning in this class is relevant to my life."). Students indicated how true they felt the items to be on a seven point scale ranging from 1 ("Not at all True") to 7 ("Very True") scale. Previous work with these scales has indicated that their scores have adequate psychometric properties (Eccles & Wigfield, 1995). Each subscale was averaged to create a final composite score and reliability coefficients for each wave are presented in Table 1.

3.3.2. Continuing interest.

Students' self-reported continuing interest was adapted from Hulleman and colleagues (Hulleman et al., 2010). Continuing interest was collected during waves one and three. The 4-item measure (α = .89; e.g., “I am interested in majoring in psychology.”) used the same response scale as the expectancy and utility value items. Individual item responses were averaged to create a final composite score.

3.3.3. Exam scores.

Students completed their first, second, third, and fourth exams in class during the fourth, eighth, twelfth, and sixteenth weeks of class, respectively. The scores were weighted equally in the course. Each exam was instructor-designed over several semesters, was multiple-choice, and was machine-scored with a maximum score of 60. Student scores were generated by summing the number of correct responses. Item-level data were unavailable for exam scores, meaning that it was not possible to calculate reliability indices for exam scores. However, the fact that exam scores here strongly and positively correlated over time, and that exam scores were positively correlated with measures of expectancies at each time point at expected magnitudes provides secondary evidence of score reliability (AERA, APA, & NCME, 2014). Furthermore, all exams were a part of an accredited university course that contributed to students’ permanent academic GPA. As a result, these are meaningful representations of achievement in this college class.

3.4. Analytic Procedures

We used latent growth curve analyses in a structural equation modeling (SEM) framework, executed in Mplus version 7.2. SEM provided a flexible analytical framework for modeling students’ change over time on two constructs simultaneously. We were also able to model the relationships between change and other variables of interest. Models were estimated using full information maximum likelihood estimation (FIML) to make use of all available data.

4. Results

4.1. Descriptive Statistics

Table 1 contains descriptive statistics and correlations. Generally speaking, both expectancy and utility value displayed decreasing trajectories over time. For example, expectancy began relatively high and decreased over time: M_{T1} = 5.46, SD = 0.83; M_{T2} = 4.92, SD = 1.16; M_{T3}= 4.68, SD = 1.34. Utility value followed a similar trajectory: M_{T1} = 4.97, SD = 1.07; M_{T2} = 4.83, SD = 1.12; M_{T3}= 4.76, SD = 1.26. Expectancy and utility value were correlated with each other at all three time points (r_{T1} = .32, r_{T2} = .42, r_{T3} = .44). In general, expectancy and exam scores became more correlated over time; a similar pattern was present between utility value over time and interest.

4.2. Model A: Unconditional Growth Model

4.2.1 Model A Specification

We began by building an unconditional parallel process model (see Figure 1), which models two separate constructs simultaneously (Byrne, 2012; Muthén & Muthén, 2012). In addition to the discussed benefits of a typical latent growth model—parallel process models provide estimates of growth-parameter covariances. Model A captured the initial levels of expectancy and utility value at the beginning of the semester (i.e., the intercepts), the change over time (i.e., the slopes), their variances, and the relationships between the intercepts and the slopes[2]. Time was coded by data collection wave[3], such

[2] Model A has one constraint, the residual variance of expectancy at Time 3 was fixed to zero—in initial runs this variance was estimated to be negative and non-significant, causing non-positive definitive residual covariance matrix. Fixing this residual variance to zero remedied the issue with the residual covariance matrix.

that the intercept represents the first wave (third week of the semester) and the slope represents the expected change as one wave passes. Model A provided a baseline comparison for later models in terms of growth parameters and model fit. Statistically, expectancy and utility value are modeled as two separate, straight lines that represent motivation across multiple time points. Conceptually, however, it is more accurate to think of expectancy and utility value as two intertwined constructs that gradually change together over time. Model A answers the first research question, *What are initial levels of expectancy and utility value in a college class, how do they change over the course of the class, and is there significant variability in these trajectories among participants?* And research question 2, *How is change in expectancy related to change in utility value?*

4.2.2. Model A Results

The primary estimates of interest in Model A were the intercept and slope means, variances, and correlations between the intercepts and slopes. The growth parameter estimates provided are unstandardized and the correlations provided are standardized for ease of interpretation (see Table 2). The expectancy and utility value intercepts were statistically significantly correlated, r = .52, p < .01, meaning that expectancy and utility value were strongly related at Time 1. Both expectancy slope, -0.31, p < .01, 99% CI [-.40, -.22] and the utility value slope -0.09, p < .01, 99% CI [-.17, -.003] were negative and statistically significant. In other words, expectancy-value motivation decreased over the course of the semester on average. The expectancy slope variability was also

[3] Data waves were coded as 0 (week 3), 1 (week 9) and 1.83 (week 14). We chose 1.83 rather than 2 because it more accurately accounts for the fact the time between the first two waves was 6 weeks long and the second two waves was 5 weeks (i.e., 83%) long. We chose this coding scheme for two reasons: (1) the time between waves was rather equally spaced out, and (2) coding by smaller units (days or weeks) produced estimates in small units, which caused estimation problems for near zero variance estimates.

statistically significant, SD= .51, *p* < .01, meaning that individual students displayed varying expectancy trajectories. The 95% plausible values range describes the range of potential expectancy slopes in the population and was from -1.47 to 0.79, with some trajectories being negative and some being positive. In contrast, the utility value slope variability was not statistically significant, SD = .28, *p* = .31[4]; the decline in utility value was consistent across students. Relatedly the 95% plausible values range was from -0.69 to 0.49. Interestingly, the slopes for expectancy and utility value were strongly and significantly correlated, *r* = .83, *p* < .01 (see Figure 2), meaning that individuals with positive expectancy slopes also likely had positive utility value slopes and individuals with negative expectancy slopes likely also had negative utility value slopes. Importantly, fit indices suggested that Model A accurately represented the data $\chi2$ (8) = 24.70, *p* < .01; RMSEA = .08; CFI = .98, TLI = .95.

4.3. Model B: Conditional Growth Model with Outcomes and Gender

4.3.1. Model B Specification.

Model B was developed to test the parallel growth of expectancy and utility value in the context of major factors and experiences in the learning environment (see Figure 3). To do so, we introduced three common factors that play a role in students' motivational development—gender, interest, and achievement. Because gender differences have historically been found in motivation and educational outcomes, we included it as a predictor in the model. As coursework is largely focused on achievement outcomes in the form of exams, we included four exams from the semester in the

[4] The results of Model A were conflicting because the variance slope, while non-significant statistically, improved model fit and was correlated with other estimates in the model. As a result, we elected to retain the utility value slope variance despite its lack of statistical significance.

model. The exams were administered between expectancy-value data collection and so provided an opportunity to study the dynamic relationship between expectancy-value growth and performance feedback. Whereas the first three exams were modeled concurrently with the growth process, the fourth exam was modeled as an outcome. Because the expectancy-value model posits that expectancies are partially determined by prior achievement, we modeled the first three exam scores as covariates. Theoretically, exam performance should lead to changes in expectancy. However, the correlational nature of the study precludes any causal inference. We suggest that path coefficients represent controlling of motivational trajectories and points of student achievement, rather than causal effects of one on the other.

Because the exams were on discrete content, we included them as an autoregressive component of the model rather than as an additional growth trajectory. The autoregressive approach is more appropriate than a growth model in this case because the exams would not effectively measure performance growth. Scoring lower on Exam 2 than on Exam 1 would not indicate that one’s psychology knowledge decreased. However, we would expect students who perform better than other students on Exam 1 to also perform better than other students on Exam 2. This difference highlights a contrast between autoregressive and growth modeling paradigms. The most accurate interpretation of the slope estimates for expectancy and utility value in this model are as conditional estimates (e.g., conditional after controlling for other factors in the model such as exam scores).

Finally, students' continuing interest was included as an additional important educational outcome near the end of the semester—baseline interest was also included as a covariate. Model B answers the third, *What are the dynamic relationships between motivation change and our primary outcomes (interest and exam performance) during the course of a semester?*

4.3.2. Model B Results.

Figure 4 displays information regarding significant paths from Model B (for full model specifications, see Table S4 in the supplemental materials). The primary relationships of interest in Model B were the dynamics between motivation change and achievement, the effects of gender on motivation growth (binary variable where men = 1 and women = 0), and the effects of motivation growth on interest and fourth exam scores. First, we found that there was a gender effect on initial expectancy, *b* = - 0.22, SE = 0.04, suggesting that women in the psychology class were likely to report lower expectancy than men. Gender had no other statistically significant relationships with variables in the model.

Expectancy intercepts were predictive of Exam 1 scores, *b* = .35, SE = 0.07, suggesting that students' confidence in their ability to succeed in a class early in the semester is indicative of their early course performance—however, their early expectancy showed no direct effects on later exam scores. Utility value intercepts were unrelated to performance at any point.

Exam 1 performance directly predicted later exam performance but was not related to expectancy change, *b* = 0.01, SE = 0.01, or utility value change *b* < -0.01, SE = 0.01. In contrast, Exam 2, *b* = 0.02, SE = 0.01, and Exam 3, *b* = 0.02, SE = 0.01, were related to expectancy change. Although these coefficients appear small, the R^2 value for expectancy slopes is quite large, showing that 52% of the variance is explained. Given that only Exam 2 and Exam 3 are significantly related to the expectancy slopes, this suggests that they share the majority of explained variance in the expectancy slopes. This means that individuals who did well on exams were more likely to have less rapidly declining motivational trajectories, and conversely, those who did poorly on these exams were likely to have more rapidly declining trajectories. Interestingly, utility value slopes related to Exam 3 scores, *b* = 0.01, SE = 0.01, but were unrelated to any other achievement measures in the study.

Finally, we examined the effects of the motivation growth on our primary outcomes, long-term interest and fourth exam performance. Interest was predicted by utility value intercepts, *b* = 0.35, SE = 0.17, and expectancy slopes, *b* = 1.03, SE = 0.48, however, it was unrelated to utility value slopes, *b* = 0.30, SE = 0.76. These results suggest that students' expectancy trajectories are indicative of their late-semester interest. Fourth exam was only predicted by other exam scores, suggesting that any relationships to motivation, if they exist, are either accounted for by prior achievement, or are indirect.

5.1. Discussion

The goals of the current research were to understand short-term changes in college students' motivation towards an introductory psychology class. Our results indicate that motivation meaningfully changes during the semester, and is related to and predictive of important educational factors. We first discuss our general findings about change and then the theoretical implications of the results, particularly as they relate to the dynamics of motivation.

5.2. Expectancy and Utility Value Change

The results of the current study correspond with trends found in the limited number of short-term (Moely et al., 2002; Perez et al., 2014) and long-term studies on motivation change (Jacobs et al., 2002). We found that both expectancy and utility value were relatively high at the beginning of the semester and both declined over time. We observed greater variance in expectancy slopes than utility value slopes. Despite the relative uniformity of utility value change over time, the correlation between expectancy and utility value slopes was statistically significant and large. Given that change in expectancy is predictive of interest, it is plausible that expectancy is the major interest leverage point in this natural education setting. The casual relationship between expectancy and utility value change is ambiguous in these data, as both were collected at the same time. Investigating a causal relationship between expectancy change and utility value change is an important area for future research; as such a link would greatly inform education and intervention practice. We hypothesize that expectancy and utility value operate through recursive processes and ultimately lead to increased or decreased performance, but because these data are correlational, future research his needed to investigate our hypothesis.

The lack of variation if utility value slopes was puzzling and yields at least two potential hypotheses that merit further investigation. Information in the classroom that can augment expectancies is abundant and typically explicit across diverse learning environments, as students get grades from exams and assignments. That feedback is critically linked to students expectancies across time (Bandura & Schunk, 1981). Utility value, however, may be less explicitly addressed in the classroom. This could result in students' entering the classroom with certain level of utility value and maintaining those attitudes over time. Another hypothesis is that even when utility value is addressed in the classroom, it does not necessarily cause students' to change their attitudes. For example, a review of research on relating math to real life suggested that faculty and students often have different interpretations of what it means to make coursework relevant (Carraher & Schliemann, 2002). Because we did not measure the amount of utility value messaging sent by teachers or received by students, we can only speculate as to the reasons behind the stability of utility value. Replicating and further exploring this finding could be the subject of future research.

5.3. Dynamics of Motivation Change and Theoretical Implications

This study was meant to examine motivation dynamics in achievement at a more fine-grained level than most prior research. In considering these motivational dynamics, two major threads emerged from our results: (1) the positive relationship between initial utility value, expectancy change, and continuing interest and (2) the positive relationship between expectancy change and exam scores.

The current study supports prior research demonstrating a strong link between interest, utility value (Shechter, Durik, Miyamoto, & Harackiewicz, 2011) and expectancy (Hidi & Renninger, 2006). First, we found a positive link between initial utility value and end-of-semester interest even after controlling for baseline interest. This finding is consistent with Hidi and Renninger's (2006) model of interest development that suggests individuals who perceive meaningfulness will develop more enduring interest. Second, continuing interest was also predicted by expectancy slopes, suggesting that self-perceptions of ability increase along with the desire to re-engage with the domain. Individuals with more negative expectancy trajectories were also more likely to report lower continuing interest at the end of the semester whereas the degree of utility value change reported by students was unrelated to interest. This is also consistent with Hidi and Renninger’s (2006) model of interest development that suggests interest continually requires external support at all stages. In the case of the current study, positive expectancy growth may correspond with positive external support whereas negative expectancy change may reduce interest.

These two findings suggest differentiated, independent relationships with interest (although the lack of utility value slope variability may be masking a relationship).The results of this study are consistent with theoretical discussions linking expectancy-value models with interest (Eccles et al., 2015; Wigfield & Cambria, 2010) as well as Hidi and Renninger’s theorized dynamics of interest development. These sorts of dynamic processes can only be captured when constructs are modeled together, and the results highlight the benefits of parallel process and flexible latent variable models. An important extension beyond this work, however, is that these processes are happening dynamically over a very short period of time, a few months.

The second point to consider is that exam performance as a whole explained about half of the variation in expectancy change (albeit with small unique effects), but expectancy change does not predict later exam performance. It may be that the correlations between expectancies at a given time point and fourth exams are simply proxies for recent achievement. In other words, measuring proximal achievement may reduce the predictive relationship expectancies on achievement outcomes. What does this mean for researchers? Including proximal achievement data may be a key to understanding expectancy effects, but expectancy is a strong proxy for prior achievement. Theoretically, the results are also interesting. The fact that exam scores predict expectancy change seems to follow logic and past research—individuals who perform worse should lose confidence that they will succeed. However, that expectancy slopes are unrelated to fourth exam scores contrasts the idea that expectancies actually influence performance. Given this fact, further research is needed on the early achievement motivation dynamics during a semester, particularly in terms of the causal direction (if any) of these relationships. The first couple of weeks of a semester may be some of the most critical in terms of motivation and subsequent achievement.

5.4. Limitations and Future Directions

There are notable limitations that constrain the conclusions drawn from this study and direct our discussion. First, as we do not have access to the content of the exam, it is possible that these findings do not extend to all psychology courses. Even if they are representative of college psychology in general, it is unlikely that they are generalizable to other academic domains. As described by Jacobs and colleagues (Jacobs et al., 2002), motivational trajectories depend on the domain in which they are contextualized—there is no reason to expect otherwise in the current study. Future research would benefit from conducting such a study with the same students in multiple courses and contexts. Not only would such a study illuminate differences in motivation trajectories in different domains, it would allow researchers to examine common motivation change that may reach across domains.

Second, motivational decline appears to be generally consistent across ages, domains, and contexts in the prior literature as well as the current study. In spite of that mean-level uniformity, there is evidence of individual variation in motivational trajectories (Jacobs et al., 2002), with some students in the current study showing upwards rather than downwards trajectories. The current study only includes one measure of individual characteristics (i.e., gender) and no measures of students' contextual experiences. In the case of gender, men appeared to display higher initial expectancy than women—however it is difficult to ascertain whether or not such a difference is to be expected. For example, psychology tends to have more women in undergraduate course whereas men are heavily overrepresented in the field. In either case, students' individual group membership could result in higher expectancies that the other group. At the same time, psychology is less familiar at the introductory level than other fields such as mathematics or language arts and there may no differences in motivation at all. Thus, any pattern of results could feasibly be hypothesized. In order to better understand this gender discrepancy, as well as why individuals' motivation declines in this particular setting, future research would benefit from assessing students from many classrooms at once. Such an undertaking could test which teaching practices or institution-level policies might explain variation in trajectories, as well as other ways students might differentiate themselves within a class.

Third, these findings are restricted to the growth of expectancy and utility value, rather than all types of value. Thus, the generalizability of these findings only applies to expectancy and utility value. The other components of Eccles' model (intrinsic value, attainment value, and cost) are empirically and theoretically distinct constructs (Eccles et al., 1983; Flake, Barron, Hulleman, McCoach, & Welsh, 2015; Gaspard et al., 2015), and may follow different trajectories during the course of an individual semester. For example, prior research (Hulleman & Harackiewicz, 2009) has demonstrated that utility value is uniquely related to expectancies. It may be that case that expectancy and value change in our study was more highly correlated than it would be otherwise because of our focus on utility value. Future research would benefit by investigating the relationships between the change estimates of different motivation components. Particularly, it is unknown how change in other types of value and change in cost change is related to expectancy or interest change. This would provide a more detailed statistical model that incorporates the dynamic processes theorized by motivational researchers (Eccles et al., 1983; Hidi & Renninger, 2006).

Fourth, the generalizability of this study is limited to relatively high achieving students. These data were collected in a research intensive university with a selective admissions process. Further, these students are relatively homogenous in their race, age, and other demographic factors. With a more diverse sample we may see a different pattern of variability of these constructs, and accordingly, a different set of predictive relationships. As such, studies of short-term change in different populations of students maybe particularly insightful for understanding the experience and outcomes of different types of students in higher education. Learning processes in gatekeeper courses may be particularly informed by motivation.

Finally, only three time points were present in the current study, meaning that we were unable to model non-linear growth or incorporate time-varying covariates. It is possible that with more time points we would observe a quadratic or cubic growth trajectory. Further we may be able to better understand performance influences motivation alongside time. Understanding the ebb and flow of student motivation during the short-term would facilitate the generation of more specific and detailed theoretical models, providing an avenue to test them. These advances expand our knowledge of the complex nature and dynamic processes of motivation, which would allow us to inform intervention research and education practice. Ultimately, we may be able to use that knowledge to catalyze recursive processes and set students on positive long term trajectories, particularly in gatekeeper courses, which are required for success in college. The current study showcases what we can learn by re-focusing some research questions on short-term motivation change. Furthermore, the limitations serve to strengthen the call for more research on this topic—the current study only scratches the surface.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology*, *102*(4), 804–816. http://doi.org/10.1037/a0021075

Atanda, R. (1999). Do gatekeeper courses expand educational options? *Educational Statistics Quarterly*, *1*(1), 33–38.

Atkinson, J. W. (1964). *An introduction to motivation*. Princeton, NJ: D. Van Nostrand Company, Inc.

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, *29*(2), 255–270. http://doi.org/10.1016/j.econedurev.2009.09.002

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology*, *41*(3), 586–598. http://doi.org/10.1037/0022-3514.41.3.586

Brown, E. R., Smith, J. L., Thoman, D. B., Allen, J. M., & Muragishi, G. (2015). From bench to bedside: A communal utility value intervention to enhance students' biomedical science motivation. *Journal of Educational Psychology*, *107*(4), 1116–1135. http://doi.org/10.1037/edu0000033

Bureau of Labor Statistics. (2015). Employment Projections. Retrieved March 9, 2015, from http://www.bls.gov/emp/ep_chart_001.htm

Byrne, B. M. (2012). *Structural equation modeling with Mplus : basic concepts, applications, and programming*. *Multivariate applications series*.

Carraher, D. W., & Schliemann, A. D. (2002). Chapter 8: Is everyday mathematics truly relevant to Mmathematics. *Journal for Research in Mathematics Education. Monograph*, *11*(May), 131. http://doi.org/10.2307/749968

Chou, C., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 247–266. http://doi.org/10.1080/10705519809540104

Chouinard, R., & Roy, N. (2008). Changes in high-school students' competence beliefs, utility value and achievement golas in mathematics. *British Journal of Educational Psychology*, *78*, 31–50. http://doi.org/10.1348/000709907X197993

Chularut, P., & DeBacker, T. K. (2004). The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language. *Contemporary Educational Psychology*, *29*(3), 248–263. http://doi.org/10.1016/j.cedpsych.2003.09.001

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Routledge Academic.

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology*, *98*(2), 382–393. http://doi.org/10.1037/0022-0663.98.2.382

Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Eccles, J. S., Fredricks, J. A., & Epstein, A. (2015). Understanding well-developed Interests and activity commitment. In K. A. Renninger & S. Hidi (Eds.), *The Power of Interest for Motivation and Engagement* (pp. 315–330). Routledge.

Eccles, J. S., Vida, M. N., & Barber, B. (2004). The relation of early adolescents' college plans and both academic ability and task-value beliefs to subsequent college enrollment. *Journal of Early Adolescence*, *24*(1), 63–77. http://doi.org/10.1177/0272431603260919

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*, 215–225. http://doi.org/10.1177/0146167295213003

Enders, C. K. (2010). Maximum Likelihood Estimation. *Applied Missing Data*.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*. http://doi.org/10.1016/S0361-476X(02)00015-2

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory.

Contemporary Educational Psychology, *41*, 232–244. http://doi.org/10.1016/j.cedpsych.2015.03.002

Fredricks, J. a., & Eccles, J. S. (2002). Children’s competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology*, *38*(4), 519–533. http://doi.org/10.1037//0012-1649.38.4.519

Gaspard, H., Dicke, A., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663–677. http://doi.org/10.1037/edu0000003

Gillet, N., Berjot, S., Vallerand, R. J., & Amoura, S. (2012). The role of autonomy support and motivation in the prediction of interest and dropout intentions in sport and education settings. *Basic and Applied Social Psychology, 34*(3), 278–286. http://doi.org/10.1080/01973533.2012.674754

Goudas, A. M., & Boylan, H. R. (2012). Addressing flawed research in developmental education. *Journal of Developmental Education*, *36*(1), 2-4-13.

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. *Computers & Education*, *80*, 152–161. http://doi.org/10.1016/j.compedu.2014.08.019

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation.

Journal of Educational Psychology, *94*(3), 562–575. http://doi.org/10.1037/0022-0663.94.3.562

Hidi, S., Berndorff, D., & Ainley, M. (2002). Children's argument writing, interest and self-efficacy: An intervention study. *Learning and Instruction*, *12*(4), 429–446. http://doi.org/10.1016/S0959-4752(01)00009-3

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. http://doi.org/10.1207/s15326985ep4102_4

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, *102*(4), 880–895. http://doi.org/10.1037/a0019506

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science (New York, N.Y.)*, *326*(5958), 1410–1412. http://doi.org/10.1126/science.1177067

Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2016). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*. http://doi.org/10.1037/edu0000146

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: gender and domain differences across grades one through twelve. *Child Development*, *73*(2), 509–527. http://doi.org/10.1111/1467-8624.00421

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2015). A practical measure of student motivation: Establishing validity evidence for the expectancy-value-cost scale in middle school. *The Journal of Early Adolescence*, *35*(5–6), 790–816. http://doi.org/10.1177/0272431614556890

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, *86*(2), 602–640. http://doi.org/10.3102/0034654315617832

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (Ed.), *Personality and the behavior disorders* (pp. 333–378). New York: Ronal Press. http://doi.org/10.1037/10319-006

Luzzo, D. A., Hasper, P., Albert, K. A., Bibby, M. A., & Martinelli, E. A. (1999). Effects of self-efficacy-enhancing interventions on the math/science self-efficacy and career interests, goals, and actions of career undecided college students. *Journal of Counseling Psychology*, *46*(2), 233–243. http://doi.org/10.1037/0022-0167.46.2.233

Marchand, G. C., & Gutierrez, A. P. (2016). Processes involving perceived instructional support, task value, and engagement in graduate education. *The Journal of Experimental Education*, 1–20. http://doi.org/10.1080/00220973.2015.1107522

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(1), 577–605. http://doi.org/10.1146/annurev.psych.60.110707.163612

Moely, B. E., McFarland, M., Miron, D., Mercer, S., & Ilustre, 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.

Musu-Gillette, L. E., Wigfield, A., Harring, J. R., & Eccles, J. S. (2015). Trajectories of change in students' self-concepts of ability and values in math and college major choice. *Educational Research and Evaluation*, *21*(4), 343–370. http://doi.org/10.1080/13803611.2015.1057161

Muthen, B., & Muthen, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, *24*(6), 882–891. http://doi.org/10.1111/j.1530-0277.2000.tb02070.x

Muthén, L., & Muthén, B. (2012). *Mplus user's guide (version 7.2)*. *Los Angeles: Author*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Mplus+user+guide#8

Papaioannou, A., Marsh, H. W., & Theodorakis, Y. (2004). A multilevel approach to motivational climate in physical education and sport settings: An individual or a group level construct? *Journal of Sport & Exercise Psychology*, *26*, 90–118.

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, *106*(1), 315–329. http://doi.org/10.1037/a0034027

Pike, A. G., & Dunne, M. (2011). Student reflections on choosing to study science post-16. *Cultural Studies of Science Education*, *6*(2), 485–500. http://doi.org/10.1007/s11422-010-9273-7

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (Second). SAGE Publications, Inc.

Ren, W.-H. (2000). Library instruction and college student self-efficacy in electronic information searching. *The Journal of Academic Librarianship*, *26*(5), 323–328. http://doi.org/10.1016/S0099-1333(00)00138-5

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychological Bulletin*, *130*(2), 261–288. http://doi.org/10.1037/0033-2909.130.2.261

Shechter, O. G., Durik, A. M., Miyamoto, Y., & Harackiewicz, J. M. (2011). The role of utility value in achievement behavior: the importance of culture. *Personality and Social Psychology Bulletin*, *37*(3), 303–317. http://doi.org/10.1177/0146167210396380

Silva, E., & White, T. (2013). Pathways to improvement: Using psychological strategies to help college students master developmental math. *Carnegie Foundation for the Advancement of Teaching*. Retrieved from http://www.carnegiefoundation.org/sites/default/files/pathways_to_improvement.pdf

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, *42*(1), 70–83. http://doi.org/10.1037/0012-1649.42.1.70

Ullrich-French, S., & Cox, A. E. (2014). Normative and intraindividual changes in physical education motivation across the transition to middle school: A multilevel growth analysis. *Sport, Exercise, and Performance Psychology*, *3*(2), 132–147. http://doi.org/10.1037/spy0000005

Vroom, V. H. (1964). *Work and motivation*. *Classic readings in organizational behavior*. New York: John Wiley & Sons, Inc.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1–35. http://doi.org/10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (2000). Expectancy – value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. http://doi.org/10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In *Development of Achievement Motivation* (pp. 91–120). http://doi.org/10.1016/B978-012750053-9/50006-1

Figure Notes

Figure 1. Model A: Unconditional Parallel Process Model. Expectancy and utility value self-reports measured at Time 1, Time 2, and Time 3 are used to calculate growth parameters (i.e., intercepts and slopes) as well as their covariances. Double-headed arrows represent covariances. Single-headed arrows represent path coefficients. Bolded, solid lines represent intercept path specifications, which are each set to 1.00 for expectancy (E1, E2, E3) and utility value (V1, V2, V3). Bolded, dashed lines represent slope path specifications which are set to 0.00, 1.00, and 1.83 for

Figure 2. Relationship between Expectancy and Utility value Slopes. Individual trajectory estimates from Expectancy Slopes (x-axis) and Utility value Slopes (y-axis) were strongly correlated in the unconditional growth model. Although the variance of utility value slopes was non-significant, it significantly co-varied with the expectancy slopes. As a result, our later models included utility value slopes as an estimated parameter.

Figure 3. Trimmed Path Model. The final trimmed model here represents the statistically significant paths that were present in Model B. Straight arrows represent regression paths. Curved , double-headed arrows represent correlations. Utility values presented are unstandardized path coefficients with standard errors displayed in parentheses. Bolded, solid lines represent intercept path specifications, which are each set to 1.00 for expectancy (E1, E2, E3) and utility value (V1, V2, V3). Bolded, dashed lines represent slope path specifications which are set to 0.00, 1.00, and 1.83 for expectancy and utility value times 1, 2, and 3 respectively.

Table 1

Means, Standard Deviations, Reliabilities, and Correlations between Expectancy, Utility Value, Exam Scores, and Interest

	(N = 389)	M	SD	1	2	3	4	5	6	7	8	9	10	11	12
1	Utility value T1	4.97	1.07	.85											
2	Utility value T2	4.83	1.12	.61	.86										
3	Utility value T3	4.76	1.26	.61	.75	.89									
4	Expectancy T1	5.46	0.83	.32	.29	.28	.83								
5	Expectancy T2	4.92	1.16	.21	.42	.40	.56	.90							
6	Expectancy T3	4.68	1.34	.16	.35	.44	.48	.84	.92						
7	Exam 1	44.85	8.09	.00	.23	.18	.27	.62	.67	--					
8	Exam 2	42.22	8.20	.01	.24	.24	.26	.64	.68	.70	--				
9	Exam 3	42.10	8.83	-.02	.17	.21	.23	.55	.68	.76	.68	--			
10	Final Exam	43.41	8.00	.01	.23	.19	.20	.51	.60	.76	.71	.75	--		
11	Interest T1	3.47	1.51	.47	.42	.46	.27	.19	.16	.00	-.01	.04	-.00	.87	
12	Interest T3	2.95	1.71	.41	.49	.56	.22	.43	.48	.30	.30	.33	.27	.68	.91

Note. Descriptives were calculated using Full Information Maximum Likelihood Estimation to account for missing data using Mplus. Self report measures all used a 7 point response scale ranging from 1 to 7 (all response options were used). Cronbach’s alpha is listed on the diagonal for all applicable scales.

Table 2

Expectancy and Utility Value Unconditional Growth Parameter Estimates

				95% Plausible value Range		Correlations			
		Estimate	Variance	Lower Bound	Upper Bound	1	2	3	4
1	Expectancy Intercept	5.44*	0.47*	4.08	6.84				
2	Expectancy Slope	-0.31*	0.27*	-1.47	0.79	0.12			
3	Utility value Intercept	4.97*	0.65*	3.40	6.54	0.52*	-0.05		
4	Utility value Slope	-0.09*	0.08	-0.69	0.49	0.06	0.83*	0.36	

Note. *p < .05. Plausible values calculated using estimate +/- (1.96 · √variance estimate) $\chi 2$ (8) = 24.7, RMSEA = .08, CFI = .98, TLI = .95, SRMR = .07

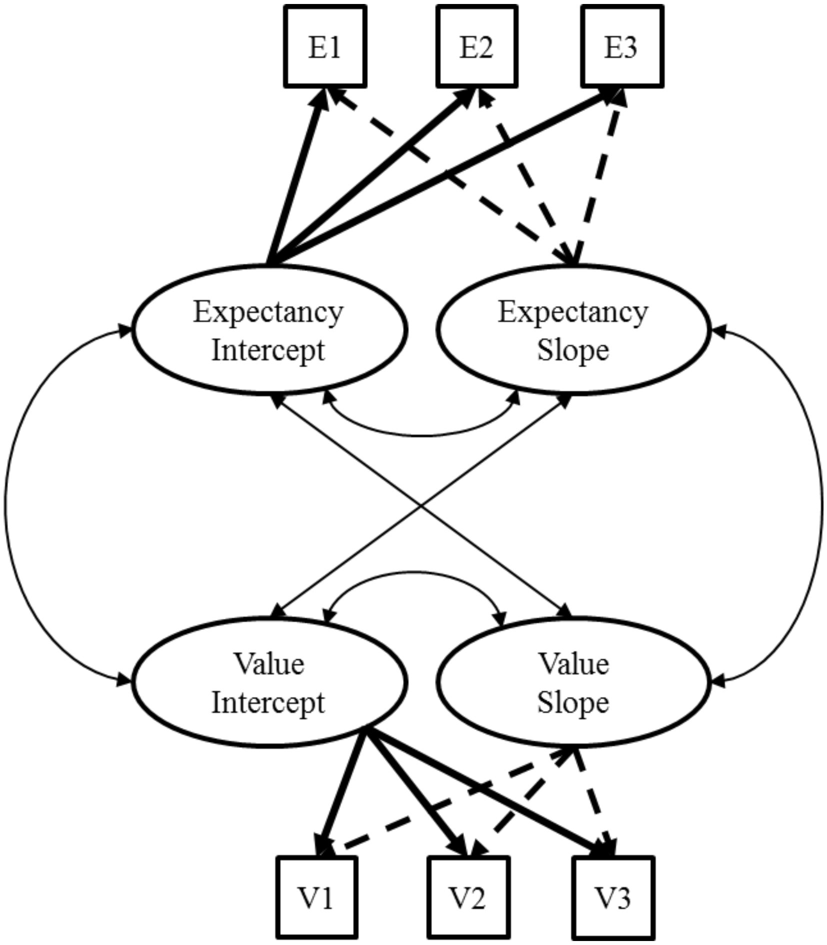
E1
E2
E3
Expectancy Intercept
Expectancy Slope
Value Intercept
Value Slope
V1
V2
V3

Figure 1

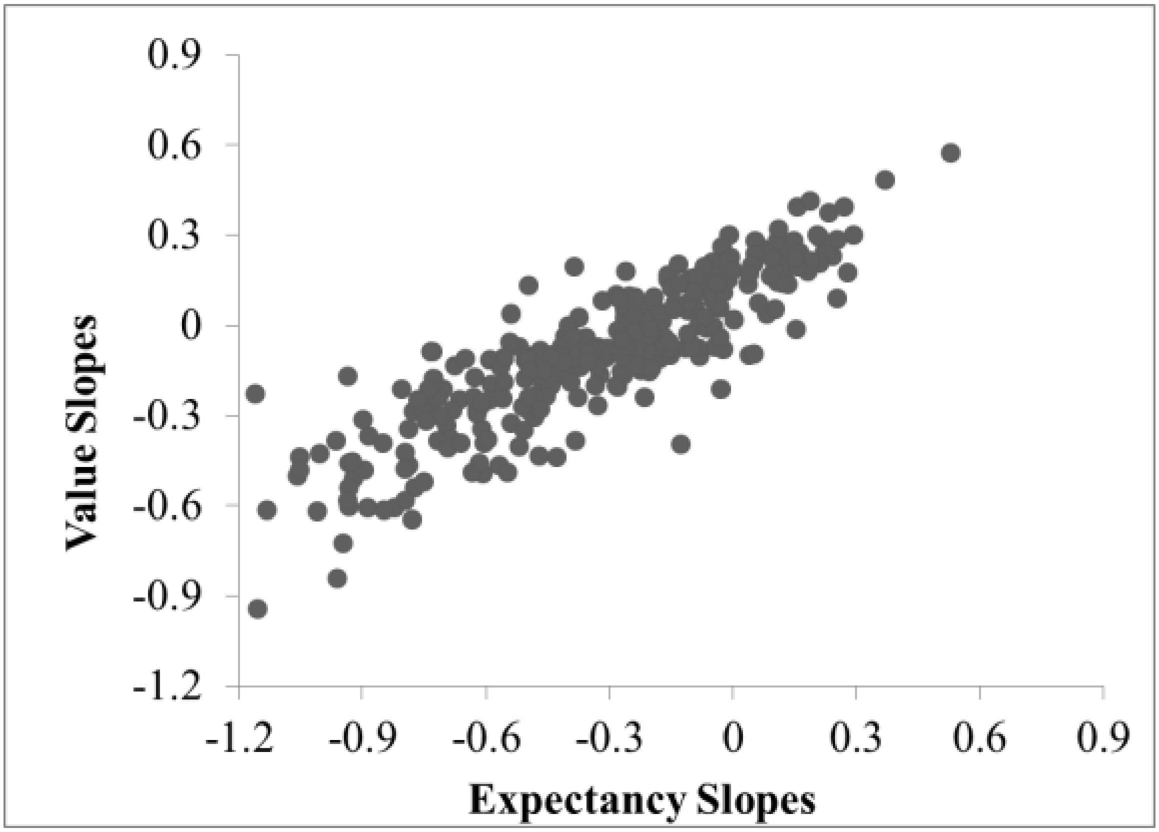
0.9
0.6
0.3
0
-0.3
-0.6
-0.9
-1.2
Value Slopes
-1.2
-0.9
-0.6
-0.3
0
0.3
0.6
0.9
Expectancy Slopes

Figure 2

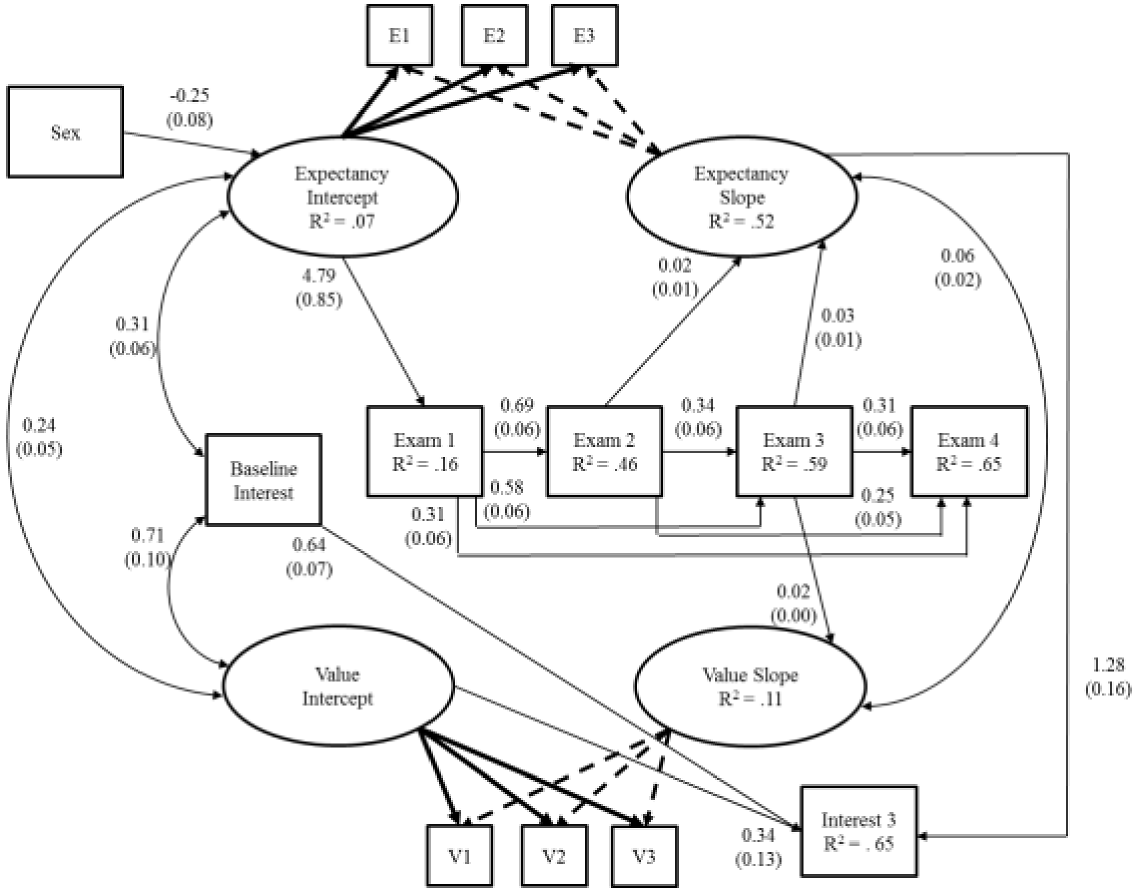
E1
E2
E3
Sex
-0.25
(0.08)
Expectancy
Intercept
$R^2 = .07$
Expectancy
Slope
$R^2 = .52$
4.79
(0.85)
0.02
(0.01)
0.06
(0.02)
0.31
(0.06)
0.03
(0.01)
0.24
(0.05)
0.69
(0.06)
0.34
(0.06)
0.31
(0.06)
Exam 1
$R^2 = .16$
Exam 2
$R^2 = .46$
Exam 3
$R^2 = .59$
Exam 4
$R^2 = .65$
Baseline
Interest
0.58
(0.06)
0.25
(0.05)
0.31
(0.06)
0.71
(0.10)
0.64
(0.07)
0.02
(0.00)
1.28
(0.16)
Value
Intercept
Value Slope
$R^2 = .11$
Interest 3
$R^2 = . 65$
0.34
(0.13)
V1
V2
V3

Figure 3

Paper 2: A Practical Measure of Student Motivation: Establishing Validity Evidence for the Expectancy-Value-Cost Scale in Middle School

Jeff J. Kosovich

University of Virginia

Chris S. Hulleman

University of Virginia

Kenneth E. Barron

James Madison University

Steve Getty

BSCS

Abstract

We present validity evidence for the Expectancy-Value-Cost (EVC) Scale of student motivation. Using a brief, ten-item scale, we measured middle school students' expectancy, value, and cost for their math and science classes in the Fall and Winter of the same academic year. Confirmatory factor analyses supported the three-factor structure of the EVC Scale, as well as measurement invariance across gender, academic domain, and time. Predictions of the EVC Scale's relationship with domain specific future interest and achievement provide convergent and discriminant validity evidence. The practical utility of the survey is highlighted by the short administration time and the alignment between observed and latent means, indicating that practitioners can use raw scores rather than latent values. Finally, we discuss methods of how to use the EVC scale to provide actionable information for educational practitioners, such as identifying which motivation interventions are most needed for students and if those interventions are working.

A Practical Measure of Student Motivation: Establishing Validity Evidence for the Expectancy-Value-Cost Scale in Middle School

Striking a balance between developing high psychometric qualities, on the one hand, and providing actionable information for practitioners, on the other hand, is a conundrum faced by many researchers. For example, scale developers typically recommend measuring constructs with numerous items to maximize scale reliability and content breadth. However, as the number of items being measured increases, the usability of the measure in the field is quickly limited due to such factors as time constraints and respondent fatigue (Yeager, Muhich, Hausman, Morales, & Bryk, in press). A tension arises because focusing too much on technical specifications can result in a meaningful but unusable assessment, whereas focusing too much on practical concerns can result in a usable but potentially meaningless assessment. These tensions can produce confusion among researchers, evaluators, and practitioners about how best to assess important educational processes and outcomes.

Although a number of motivation scales have been developed, existing measures have a variety of limitations for routine, widespread use. First, a proliferation of theoretical constructs can make it difficult for practitioners to know which measures to use (Murphy & Alexander, 2000). Second, motivational measures are often not validated across academic contexts (e.g., math and science), across populations with different characteristics, or across time. Third, the length of previous measures is not always practical for use in classrooms to quickly assess student motivation at a single time point, or to sample/measure motivation repeatedly over a longer period of time. Similarly, the lack of a reliable and easy to use motivation measure renders it difficult for researchers or program evaluators to assess the effectiveness of educational interventions designed to enhance student motivation.

In this paper, we address technical concerns as well as practical applications for a rapid measure of student motivation, the Expectancy-Value-Cost (EVC) Scale. We argue that a scale's practical utility results from balancing technical properties and practical concerns. Using latent variable modeling, we demonstrate that the ten-item EVC Scale can measure three theoretically separate and important motivational constructs. Furthermore, we provide evidence that practitioners can draw similar conclusions about motivational differences without sophisticated statistical modeling and without sacrificing a large amount of class time.

Expectancy-Value Motivation and Assessment

Of the numerous motivation theories and constructs that appear in contemporary educational psychology, expectancy-value models (Eccles et al., 1983) offer a comprehensive framework for understanding student motivation (Brophy, 2010). The model proposes that motivation consists of two key factors that predict important educational outcomes (Eccles, et al., 1983; Feather, 1988): *expectancy* and *value*. Expectancy, which is linked to achievement outcomes (e.g., grades), reflects the extent to which a student thinks he or she can be successful in a task. Value, which is linked to other academic outcomes (e.g., future interests), reflects the extent to which a student thinks a task is worthwhile (Wigfield & Cambria, 2010).

Assessments of expectancy-value motivation have a long history in education research (Wigfield & Cambria, 2010). Eccles and her colleagues proposed that expectancy and value are separate factors that each can be further distinguished into several dimensions (e.g., Eccles & Wigfield, 1995). Specifically, Eccles and colleagues argued for two dimensions of expectancy and four dimensions of value (see Eccles et al., 1983). The dimensions of expectancy included *ability beliefs* (what students think they can do now) and *expectancy beliefs* (what students think they will be able to do in the future). The dimensions of value are distinguished according to what enhances or undermines a student's overall value for the activity. Positive contributors include *intrinsic value* (engaging in an activity because it is inherently enjoyable), *utility value* (engaging in the activity because it helps achieve other short-term or longer term goals), and *attainment value* (engaging in the activity because it affirms an important aspect of a student's identity). In contrast, *cost* reflects negative aspects of engaging in an activity, such as perceptions of the effort and time required to be successful, the loss of engaging in other valued activities, or negative psychological states from struggling or failing at the activity. Based on prior research, expectancy and value (with the exception of cost) are typically positive correlated to each other and educational outcomes such as achievement or student persistence (e.g., Durik, Vida, & Eccles, 2006). Alternatively, cost is negatively related to expectancy, value, and learning outcomes (for a review see Barron & Hulleman, in press).

Research on existing expectancy-value scales indicated several challenges that informed our current work. First, dimensions within each construct often are highly correlated or load onto one factor (Eccles & Wigfield, 1995). As a result, researchers often pool items across dimensions into an overall, combined expectancy or value scale (e.g., Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). Second, combined scales are often pulled from a larger pool of items that can vary across studies and contexts (e.g., items used in middle school may differ from those used in high school or college). Third, while Eccles experimented with cost scales as a dimension of value (see Parsons, 1980), cost scales generally have not been investigated further in subsequent empirical work (for reviews see Barron & Hulleman, in press; Flake, 2012). As a result, less is known about how cost actually functions, alongside expectancy and value, to influence student motivation. A newer conceptualization of cost in expectancy-value models, however, emphasizes an important distinction between value and cost and re-labels the model as expectancy-value-cost to elevate the unique role of cost (Barron & Hulleman, in press). Recent empirical work across diverse samples also suggests that cost separates into its own factor, and is negatively related to expectancy and value (Conley, 2012; Flake et al., 2011; Grays, 2013; Trautwein et al., 2012).

Despite the theoretical suggestion that expectancy, value, and cost represent separate factors of motivation, relatively little research has focused on the psychometric quality of these scales. Most often, a single study conducted by Eccles and Wigfield (1995) is cited as providing support for expectancy-value scales. However, the items in their study did not formally include items to measure cost. In addition, the items were developed and validated to capture multiple dimensions of expectancy and value, which differs from how the items are frequently used in practice when researchers select one item from each value dimension and then combine them into a single value scale. The Eccles and Wigfield study did not provide empirical evidence or tests of a shorter, more practically useful scale that assesses expectancy, value, and cost as three separate, unidimensional scales. In this paper, we present our initial work to develop and validate a three-factor scale of Expectancy-Value-Cost (EVC) motivation to fill this gap.

The proposed EVC Scale builds on a body of developmental work piloting the scale in a variety of classes with different student populations, ranging in age from middle school through early college. Careful attention was paid to the wording of items so the EVC Scale could be used across a variety of age levels without additional editing for age. Similarly, the academic domain measured by the scale can be easily altered by changing the reference (e.g., “math class” to “science class”). The initial version of the EVC Scale included 24 items used in undergraduate general education courses (Flake et al., 2011; Grays, 2013). This larger pool of items was reduced to 12 for use in evaluating an online intervention in high school science (Getty et al., 2013). Subsequently, as part of a National Science Foundation grant, a panel of experts critically evaluated the items, including an analysis of whether or not the items align with the intended theoretical constructs (i.e., face validity). Other formative data contributing to the EVC Scale development was gleaned from qualitative research on cost (Flake, 2012), and by coding and comparing open-ended responses with Likert-scale item responses. Together, these studies form a basis for the current ten-item EVC Scale.

We also see the EVC Scale serving as a *Tier 1 Support scale* as discussed in the School-Wide Positive Behavioral Support approach to prevention (Gresham, 2004). We intentionally draw a parallel between the EVC Scale and Tier I Supports because the latter work provides a useful framework for identifying areas to target for intervention. Similar to Tier I behavioral supports for all students (Stewart, Benner, Martella, & Marchand-Martella, 2007), we offer the EVC Scale as a Tier I support to capture motivation levels for students in a classroom, grade, or school. As such, the EVC Scale can provide a quick “pulse” of what the motivational profile of a student, classroom, or school could be. For example, teachers and administrators could use this scale to determine which motivational issues are of particular concern. In one classroom, expectancy for math may be the issue; in another classroom, value for science may be the issue. With such formative data, targeted interventions could be adopted or developed to address the particular motivational issues that are identified (expectancy, value, and/or cost).

An emerging body of research highlights the potential for targeted psychological interventions to focus on certain elements of student motivation (e.g., Yeager & Walton, 2011). Furthermore, different interventions should be used to impact different motivational concerns. For example, Dweck and colleagues' growth mindset intervention changes student perceptions about whether they can be successful in a class through effort rather than inherent ability (Blackwell et al., 2007; Dweck, 2006; Yeager et al., 2013). As such, the growth mindset intervention could be considered a type of expectancy intervention (Hulleman, Barron, Kosovich, & Lazowski, in press). Similarly, Hulleman and colleagues' relevance intervention increases perceived utility value for learning material (Hulleman & Harackiewicz, 2009). In addition, Cohen and colleagues' self-affirmation intervention decreases stereotype threat (a cost for learning) and improves subsequent performance school (Cohen et al., 2009) and could be considered a cost intervention. Recognizing that different types of interventions are needed for different motivational concerns, we developed the EVC Scale to help practitioners identify beneficial interventions for their students.

A Blended Approach to Measuring Motivation

In order to develop a scale that is both technically sound and practically usable, we pursued a blended approach to measurement to evaluate the EVC Scale. First, we focus on the technical specifications (i.e., validity evidence, Messick, 1995) that address scale structure and item functionality from a psychometric perspective. Second, we examine usability concerns such as administration time and readily useable results from an applied perspective (i.e., social validity; Gresham & Lopez, 1996). Our approach to measurement strengthens the argument that the EVC Scale is both meaningful (technically and theoretically sound) and usable (easy to implement).

We use confirmatory factor analysis (CFA) to test the congruence between the theoretical and observed scale structure. Based on different theoretical possibilities about the relationships between expectancy, value, and cost, four scale structures were tested for the EVC Scale (see Figure 1). In following best practice guidelines for CFA model comparisons (Kline, 2011), the first model tests a one-factor structure in which all of the items represent a single construct. The second model tests an additional two factor structure with an Expectancy factor and a combined Value-Cost factor (Eccles' original conceptualization of value; Eccles et al. 1983). The third model tests a two-factor structure in which expectancy and value form a single factor (Positive Motivation) and a separate Cost factor. The final model tests a three-factor structure with distinct Expectancy, Value, and Cost factors as proposed in the revised Expectancy-Value-Cost framework (Barron & Hulleman, in press).

An important issue when considering the practical utility of a scale is whether or not it can provide accurate information without sophisticated statistical modeling. Unfortunately, there are a number of problems associated with raw item information

(Marsh, Scalas, & Nagengast, 2010), including variation in how students respond to items across different groups (e.g., gender, age), that can influence the factor structure, factor loadings, item intercepts, and item error variances. One solution is to sequentially test increasingly restrictive CFA models that constrain items parameters to be equal (measurement invariance; Vandenberg & Lance, 2000). Such tests of measurement invariance allow us to examine the extent to which variation in scale scores across students is due to real changes in a construct rather than systematic variation across groups or across time. *Configural invariance* tests whether the same factor structure exists across groups or time. *Metric invariance* tests whether factor loadings (of the items on their respective expectancy, value, or cost factor) are equal across groups or time. *Scalar invariance* tests whether item intercepts are equal across groups or time. Finally, *equal error invariance* tests whether the error variances (or item uniqueness) are equal across groups or time. Establishing measurement invariance is integral to a scale's practical utility because it tests for differential item functioning between groups or time and allows total scores to be calculated using raw data. If all four types of invariance are tenable, strong evidence exists that observed scores represent latent scores and differences in those scores are representative of actual construct differences.

The Science, Technology, Engineering, and Math Context

Although we designed the EVC scale to apply across content areas, the current study focuses on student motivation in STEM classes (Science, Technology, Engineering, and Math) in middle school mathematics and science. STEM subjects play a vital and direct role on a student's ability to be successful in school, to graduate from school, and should they wish, and to become a part of the STEM-related workforce. Research in

STEM education highlights declines in student motivation for math and science (Global Science Forum, 2006). The proportion of students pursuing STEM fields in the United States has decreased despite a growing need for individuals with STEM training for continued economic competitiveness and success (National Research Council, 2010). One fruitful approach to understanding, and possibly reversing, this trend is the systematic study of non-cognitive skills and attitudes like student motivation in STEM domains (Easton, 2013). Domain-specific motivation can predict course choice, educational persistence, and achievement (Durik, Vida, & Eccles, 2006; Jacobs et al., 2002). Well-developed and practitioner-friendly scales of motivation are needed toward this end.

Goals of the Current Study

Using a brief, ten-item scale, we measured middle school students' expectancy, value, and cost for math and science in the Fall and again in the Winter of the same academic year. There are four major goals of the current study. First, we examined the dimensionality of the EVC Scale using CFA to determine whether the scale adheres to a three-factor structure or to an alternative one- or two-factor structure. Second, we used measurement invariance models to test if observed scores on the EVC Scale can accurately reflect observed scores across gender, academic domain, and time. Third, we examined convergent and discriminant evidence by testing the extent to which it correlates with standardized test scores, future interest in that academic domain, and EVC Scale scores across domains (i.e., math and science). Finally, we examined average survey completion times to determine if the scale was administered quickly.

Method

Participants

The current study uses data collected from students at a diverse, public middle school in the southeast (59% free/reduced lunch, 45% limited English proficiency, 40% white, 39% Hispanic, 10% African-American). Of the 547 students enrolled in the middle school's sixth, seventh, and eighth grades, approximately 100 had no EVC Scale data due to logistical problems during the first year of scale administration. Additionally, some students were missing data due to absences or data collection issues (e.g., illness, early release days, snow days). Because we were interested in comparing math and science motivation, our sample was further restricted to students who either had complete EVC Scale data in math and science in Fall 2012 or Winter 2012. This resulted in an overall sample of 401 students. Demographically, boys and girls were evenly represented in the final sample (51.4% girls). The racial distribution was 51% White, 31% Hispanic, 9% Black, 4% Asian, and 5% Mixed or Other.

Due to relatively small numbers of students within grades, the sample included sixth (n = 114), seventh (n = 137) and eighth grades (n = 150) students. Prior research suggests that sixth, seventh, and eighth grade students follow similar trajectories of motivation over time (Jacobs et al., 2002). Preliminary CFAs and invariance analyses also suggested some support for invariance across grades (Supplementary Materials), but due to small sample size, we did not have sufficient statistical power for a formal test of invariance across grade levels.

Procedure and Measures

Expectancy-Value-Cost (EVC) Scale. Students completed the brief, ten-item EVC Scale (see Appendix A) through an online survey platform before participating in benchmark tests about that respective domain (math or science). Some students who did not have access to a computer completed the survey using paper and pencil. Response times were collected for all computer administrations. Sample items include: expectancy (three items, e.g., *I know I can learn the material in my math/science class*), value (three items e.g., *I value my math/science class*), and cost (four items e.g., *My math/science classwork requires too much time*). Items used a six point, Likert-type scale ranging from 1 (*Strongly Disagree*) to 6 (*Strongly Agree*).

Outcome measures. Separate three-item scales were used to assess future interest in math (α_{fall} = .81; α_{winter} = .86) and science (α_{fall} = .87; α_{winter} = .88) , which were based on similar measures of future interest (Hulleman, Godes, Hendricks, & Harackiewicz, 2010), and comprised of the items: "*I look forward to learning more about math/science*", "*I want to take more math/science classes in the future*", and "*I want to have a job that involves math/science someday*". Standardized test scores (i.e., the state's No Child Left Behind assessments) from spring of the prior year were obtained from school district data.

Data Analysis Plan

Confirmatory factor analysis. To test different conceptualizations of the EVC Scale, the four competing CFA models presented in Figure 1 were tested using Mplus 7.1. In all models, the items were constrained to load only on their respective factor (no cross loading), and the factors were allowed to correlate. CFAs were conducted separately for Math and Science in both the Fall and Winter Samples. To assess the adequacy of global fit, we examined χ^2 difference tests (Satorra & Bentler, 2010) to compare nested models. We also examined RMSEA, CFI and SRMR values which provide information about model misspecification, improvement over a null model, and an average estimate of correlation residuals, respectively. Recommendations for good fit were ≤ .06 for RMSEA, ≥.95 for CFI, and ≤ .08 for SRMR (for a review, see Hooper, Coughlan & Mullen, 2008). In addition, we further investigated model misspecification by examining model residuals. Following recommendations from Kline (2011), we focused particularly on correlation and standardized residuals with values greater than |.10| and |3.0| respectively. Although these guidelines are provided, it is largely left to researcher's judgment to determine if the local misfit is a concern and which index to use. We treated our measures as continuous indicators (Rhemtulla, Brosseau-Liard, & Savelei, 2012) and used maximum likelihood estimation with robust standard errors (MLR) for all CFA and invariance tests.

Measurement invariance. Three sets of measurement invariance also were conducted using Mplus 7.1. First, gender invariance was tested separately for Math and Science in both the Fall and Winter Samples. Second, cross-domain invariance was conducted to determine if the EVC Scale functions the same in math and science. Finally, longitudinal invariance analyses were conducted separately for math and science. In specifying the longitudinal invariance models, latent standardization (i.e., setting latent means to zero and latent variances to one) was used to set the scale of the latent factors. This method of standardization is particularly useful in longitudinal invariance because the parameter estimates of the latent means that are produced are analogous to Glass's Δ (Little, Slegers, & Card, 2006). The production of latent Glass's Δ also enables us to compare latent and observed effect sizes (i.e., examines whether or not practitioners can use the scale without advanced statistics).

Rather than testing if more restrictive models (i.e., metric and scalar) improve fit, invariance testing is concerned with maintaining fit. Thus, more restricted invariance models are considered to fit if they do not drastically decrease the fit of models as measured by chi-square difference tests ($p > .05$) and changes in CFI ($\leq .01$; Cheung & Rensvold, 2002). Instead of using strict accept/reject decision rules (Iacobucci, 2009; Steiger, 2007), we used assessments of global model fit, local model fit, and change in model fit between competing models to make holistic judgments about the adequacy of tested models.

Convergent and discriminant evidence. One method for establishing convergent and discriminant evidence for validity is assessing whether or not responses to the EVC Scale for science differ from responses for math. To do this, we examine the latent correlations of the EVC Scale in math class and science class. To further assess convergent and discriminant evidence, we also correlate expectancy, value, and cost with another measure of domain-specific motivation (future interest in each respective academic domain) as well as a measure of domain-specific achievement (prior state standardized test scores). These correlations can be used to assess if the EVC Scale relates to future interest and prior achievement in expected directions at expected magnitudes. The correlations with science achievement are based on smaller samples due to missing data and the fact that only 6th grade science achievement scores were available.

Completion time. To assess if the EVC Scale could be administered with relatively little time commitment, we calculated descriptive statistics for student completion time of the scale.

Results

Descriptive Analyses

Individual item descriptive statistics were similar across domains and time points. Expectancy and value items tended to have means of near 5.0 and standard deviations of 1.0, whereas cost items tended to have means near 2.5 with standard deviations of 1.3 (Supplementary Materials). Given past research on student levels of motivation (Jacobs et al., 2002), these high means are not without precedent among middle-school-aged students. The full range of responses was used except for Winter math expectancy which did not use the strongly disagree option. In terms of percentage of people who chose each response category, expectancy and value items tended to have the three disagree options chosen 5% to 15% of the time, whereas cost items had slightly more spread across response options (Supplementary Materials). Finally, when examining item correlations for math and science in Winter and Fall (Supplementary Materials), intra-construct items tended to be more highly correlated with each other than with items from other constructs.

Confirmatory Factor Analyses

Model fit was assessed for all factor structures tested (see Figure 1). The one-factor model (Model 1) and the two factor model of Expectancy vs. Eccles' Value (Model 2) did not display adequate fit. Although the two factor model of Positive Motivation vs. Cost (Model 3) typically displayed adequate fit, the three factor Expectancy-Value-Cost model (Model 4) was championed in all instances because it displayed superior fit (see Table 1). In the Fall, Model 4 displayed good fit in measuring math, χ^2 (32) = 30.46, *p* = .54; RMSEA < .01; CFI > .99; SRMR = .02, and science, χ^2

(32) = 44.87, p = .07; RMSEA =.04; CFI = .99; SRMR = .03, and demonstrated significantly better fit than Model 3 in both math, $\Delta\chi^2$ (2) = 56.96, p < .01) and science, $\Delta\chi^2$ (2) = 53.82, p < .01. Similarly in the Winter, Model 4 displayed good fit in measuring math, χ^2 (32) = 40.58, p = .14; RMSEA = .03 CFI = .99; SRMR = .03, and science motivation, χ^2 (32) = 42.70, p = .10; RMSEA =.04; CFI = .99; SRMR = .03, and demonstrated better fit than Model 3 in both math, $\Delta\chi^2$ (2) = 147.36, p < .01, and science, $\Delta\chi^2$ (2) = 35.65, p < .01. Based on these findings we tested for invariance of the three factor model (Model 4), which represented separate factors of expectancy, value, and cost.

Gender, Cross-Domain, and Longitudinal Invariance

Gender invariance. We concluded that some measurement invariance was present for when comparing gender (See Table 2, Gender). It is important to note that in all cases, the global fit indices for the Scalar models decreased only slightly from the Metric models. For example, when comparing the metric model, χ^2 (71) = 104.53, p = .01; RMSEA = .06; CFI = .97; SRMR = .06, to the scalar model, χ^2 (78) = 115.46, p < .01; RMSEA = .06; CFI = .97; SRMR = .07; and $\Delta\chi^2$(7)= 11.13, p = .13, the fit indices were virtually identical. Despite areas of local misfit, we determined that the EVC Scale effectively displayed scalar invariance for gender. However, for equal error invariance models, the SRMR values and percentages of large correlation residuals increased substantially. Further inspection of correlation residuals revealed that the model was overestimating the correlations between the first and third expectancy items. The model also appeared to be overestimating expectancy and cost item correlations, and underestimating the correlations between some value and cost items, though these values tended to be closer to |.10| and likely less problematic. The boys' and girls' models both displayed similar areas of local misfit, but the values tended to be larger for boys. As a result, we caution comparing mean differences between boys and girls without latent variable modeling.

Academic domain invariance. Academic domain invariance (i.e., math vs. science) was more clearly supported than gender invariance (see Table 2, Invariance Models: Academic Domain). We judged the fit of the equal error invariance model to be adequate in the Fall, χ^2 (169) = 253.01, $p < .01$; RMSEA = .04; CFI = .97; SRMR = .06, and $\Delta\chi^2(10) = 16.29$, $p = .09$; as well as in the Winter, χ^2 (169) = 284.29, $p < .01$; RMSEA = .05; CFI = .95; SRMR = .06, and $\Delta\chi^2$ (10) = 20.90, $p = .02$. Despite a decline in fit from the scalar model to the equal error variance model, global and local fit were still quite good for the latter. In terms of local misfit, the majority of larger residuals indicated over- or under-estimation within domain (though the actual residual values were still quite small), rather than between domains.

In addition to testing measurement invariance, the academic domain invariance models provided information about the domain specificity of motivation constructs in the form of latent variable correlations. The patterns of latent correlations demonstrate that math expectancy, value, and cost are only somewhat related to science expectancy, value, and cost (see Table 2, Academic Domain Invariance by Time). Cross-academic domain, same-factor motivation correlations (e.g. Fall math and science expectancy, $r = .29$) were typically larger than cross-academic domain, different-factor correlations (e.g., Fall math value and Fall science expectancy, $r = .19$). One notable exception was that cross-academic domain cost was moderately correlated in both Fall, $r = .57$, and Winter, $r =$

.57. The larger cross-domain cost correlations suggest that either cost is measured at a more domain-general level, or that cost is a construct that is less dependent on domain. Following these results, we tested for longitudinal invariance from Fall to Winter in science and math.

Longitudinal invariance. The EVC Scale also demonstrated observed longitudinal invariance (See Table 2, Longitudinal Math and Longitudinal Science). Both the math, $\chi^2(169) = 234.03$, $p < .01$; RMSEA = .03; CFI = .98; SRMR = .05; and $\Delta\chi^2(10) = 17.35$, $p = .07$, and science, $\chi^2(169) = 223.69$, $p < .01$; RMSEA = .03; CFI = .98; SRMR = .05; and $\Delta\chi^2(10) = 14.84$, $p = .14$, equal error invariance models displayed excellent global fit (see Supplementary Materials). Local misfit tended to be weak or nonexistent, though there may be some underestimation between some expectancy and cost items as well as between different cost items at different time points. As with the domain invariance models, the χ^2 difference tests suggest that the scalar models fit statistically significantly worse that the metric models. However, both latent invariance and observed invariance models display good fit in their own right.

Mean change over time. The results of the longitudinal invariance analyses allowed us to compare the latent mean differences (Δ_L) to the observed mean differences (Δ_O). Small or negligible discrepancies would allow practitioners to use the observed change as reliable indicators of construct change. The only notable discrepancy was between the latent and observed effect size for science expectancy. Observed and latent change was otherwise comparable. Math expectancy did not change ($\Delta_L = 0.01$, $p = .92$; $\Delta_O = 0.08$, $p = .36$), but science expectancy showed a positive trend ($\Delta_L = 0.08$, $p = .16$; $\Delta_O = 0.20$, $p = .01$). Math value did not change ($\Delta_L = -0.05$, $p = .32$; $\Delta_O = 0.08$, $p = .49$), but science value again showed a small increase ($\Delta_L = 0.14$, $p < .01$; $\Delta_O = 0.24$, $p < .01$). In contrast, there was a decrease in both math cost ($\Delta_L = -\ 0.18$, $p < .01$; $\Delta_O = -0.25$, $p < .01$) and science cost ($\Delta_L = -0.17$, $p < .01$; $\Delta_O = -0.21$, $p = .01$). These findings indicate that latent and observed change is reasonably similar.

Reliability estimates. Reliability estimates were calculated using coefficient ω which is a more accurate index of internal consistency than α (Yang & Green, 2011). Because the error variances were constrained to be equal, reliabilities for each construct were the same in the Fall and Winter. Reliabilities were good for math, $\omega_{\text{expectancy}} = .88$, $\omega_{\text{value}} = .84$, $\omega_{\text{cost}} = .86$, as well as science, $\omega_{\text{expectancy}} = .88$, $\omega_{\text{value}} = .88$, $\omega_{\text{cost}} = .87$. The EVC Scale also displayed moderate to strong test-retest reliability at the latent level (see Table 3, Longitudinal Invariance by Domain), with the longitudinal correlations being slightly higher, for example, when comparing math expectancy, $r_{\text{math expectancy}} = .74$, to science expectancy, $r_{\text{science expectancy}} = .68$.

Convergent and Discriminant Evidence

In an effort to test for convergent and discriminant evidence, we examined the relationships between expectancy, value, and cost, future academic domain interest, and domain-specific standardized achievement scores. Table 3 presents the latent within- and between-domain correlations for expectancy, value, and cost. As expected, correlations among the subscales of the EVC Scale were more strongly related within domain than across domain. For example, math expectancy and math value are moderately correlated $r = .55$, whereas math expectancy and science value are less strongly correlated, $r = .31$. Similarly, math expectancy and science expectancy are also weakly correlated, $r = .29$, indicating evidence for cross-domain discrimination between constructs.

Results indicate that the predicted pattern of correlations held between future interest and math expectancy, value, and cost in the Fall ($r_{i.e}$ = .59, $r_{i.v}$ = .68, $r_{i.c}$= - .36) as well as in the Winter ($r_{i.e}$ = .53, $r_{i.v}$ = .70, $r_{i.c}$= - .44). In science, a similar pattern of correlations emerged between future interest and expectancy, value, and cost with interest in the Fall ($r_{i.e}$ = .61, $r_{i.v}$ = .76, $r_{i.c}$= - .38) as well as in the Winter ($r_{i.e}$ = .70, $r_{i.v}$ = .76, $r_{i.c}$= - .47).

In addition to interest, the EVC subscales were correlated with math and science achievement. Reduced samples of students were used to calculate these correlations because we were unable to obtain achievement scores for all students. For example, the science test was only completed by 6th graders. For math achievement, there were there were expected correlation patterns with expectancy, value, and cost in the Fall ($r_{m.e}$ = .18, $r_{m.v}$ = .14, $r_{m.c}$= - .17; n = 305) and in the Winter ($r_{m.e}$ = .18, $r_{m.v}$ = .15, $r_{m.c}$= - .17; n = 262). There were similar correlation patterns for science achievement with expectancy, value, and cost in the Fall ($r_{s.e}$ = .39, $r_{s.v}$ = .26, $r_{s.c}$= - .29; n = 49) and in the Winter ($r_{s.e}$ = .47, $r_{s.v}$ = .35, $r_{s.c}$= - .41; n = 86).

Scale Completion Time

A desirable property of a practical scale is a short completion time—a property displayed by the EVC Scale. The average completion time in minutes for both math, M = 4.00, SD = 0.97, and science, M = 3.36, SD = 0.89, was less than five minutes. Times also decreased slightly after the first implementation of the scale. It is important to note that as part of the administration of the EVC Scale additional demographic questions were included. As a result, the time estimates are conservative and the EVC items alone would have taken less time. Based on these results, practitioners could gain a relatively large amount of information in little time.

Discussion

Building upon prior work, the EVC Scale offers a rapid, practical means to measure student motivation. The current study provides validity evidence for the Expectancy-Value-Cost (EVC) Scale in math and science classes. First, we examined the structural properties of the EVC Scale. Second, we assessed if the observed values could be used by practitioners without latent variable modeling by testing for measurement invariance. Third, we examined whether or not responses to the EVC Scale correlated to other measures of motivation and achievement in expected ways. Finally, we assessed how efficiently the scale could be administered by examining completion time.

Structural Properties of the EVC Scale

A revised, three-factor EVC framework (Barron & Hulleman, in press) was supported through CFA, suggesting that expectancy, value, and cost are separate factors in both math and science. Furthermore, invariance testing offered preliminary support that the three-factor model was accurately reflected the observed data. Thus, observed score results of the EVC Scale can be used by practitioners without having to rely on modeling data at the latent level with advanced statistical techniques. In addition, the invariance analyses revealed that the EVC Scale can be used to compare student motivation over time in both math and science. Reliability estimates were also favorable at each measurement occasion. As mentioned, the EVC scale is intended to be used as a regularly-administered, primary support measure that can provide practitioners, evaluators, and researchers with a general pulse of students' motivation.

Convergent and Discriminant Evidence

The correlations presented between the EVC Scale in math and science, as well as those between the EVC Scale and domain specific achievement and future interest provide convergent and discriminant evidence for the EVC scale. First, in line with the expectancy-value framework (Wigfield & Cambria, 2010), correlating the EVC Scales for math and science provide evidence that motivation is domain specific. Second, because future interest is promoted by seeing value (Eccles et al., 1983), value should be highly correlated with future interest; in the current data, value and future interest were highly correlated, providing some convergent evidence for the EVC Scale. In contrast, expectancy and cost were more weakly correlated with future interest, providing discriminant evidence. Moreover, correlations between future interest, expectancy, and value also highlight how expectancy and value are unique constructs, despite high expectancy-value correlations. Finally, the expected pattern of relationships between EVC Scale subscales and achievement were present—achievement correlated more strongly with expectancy than value.

Practicality and Usability of the EVC Scale

In terms of practicality, the EVC Scale can be completed quickly with minimal intrusion on class instruction. Most administrations occurred via an online survey platform and required, on average, less than five minutes. This is encouraging because two major barriers to psychological measurement in applied settings are time limitations and delivery constraints (Yeager et al., in press). A rich body of work has investigated the role of non-cognitive skills in learning and achievement, such as motivation, perseverance, academic behaviors, or academic mindsets toward success in school (for reviews see Pintrich, 2003; Snipes, Fancsali, & Stoker, 2012). As such, the development of the EVC Scale provides practitioners, researchers, and program evaluators with a tool for quickly assessing three types of non-cognitive attributes.

Using the EVC Scale could help researchers identify how types of motivation connect to academic performance, or to a student's future interest in those domains. It is also possible that the EVC Scale could be used to track or measure the effectiveness of a motivation intervention. For example, the EVC Scale was used to evaluate the effectiveness of an online intervention on student learning outcomes (Getty et al., 2013; Lazowski, Hulleman, Barron, & Getty, 2012). In this case, the intervention replaced typical science instruction while the EVC Scale made it possible to assess student reactions to online instructional materials. Not only did students react differently to the computerized instruction across classrooms, but motivation during the three week period predicted changes in learning outcomes. Specifically, expectancy positively predicted science content knowledge, value positively predicted future science interest, and cost negatively predicted science content knowledge.

Having information on expectancy, value, or cost problems can help teachers tailor instruction based on knowing classroom-wide or individual motivation deficit. Interventions that increase student expectancy are different than interventions focused on increasing value or decreasing cost. As noted earlier, growth mindset interventions (e.g., Blackwell et al., 2007) help convince students that they can learn and get smarter through effort and engaging in academic challenges. This type of intervention, which promotes expectancy, is very different than an intervention designed to enhance students' perceptions of value for the material. In contrast, a value intervention might ask students to focus on how the material they are studying relates to their life (Hulleman & Harackiewicz, 2009). These interventions improve motivation by targeting different psychological processes which could be identified as expectancy or value deficits. We propose administering the EVC Scale to determine which intervention is needed by identifying which general motivational factors are most at risk for a particular classroom.

Finally, prior research has demonstrated that student motivation declines over time as students progress through grades K-12 (e.g., Jacobs et al., 2002). In particular, students undergoing academic transitions often suffer the largest motivational declines, such as when transitioning to middle school (Eccles & Wigfield, 2000), high school (Casillas et al., 2012), or college (Silva & White, 2013). The EVC Scale could be utilized before, during, and after these key transition points to help identify struggling students in need of motivational remediation. For example, the Carnegie Foundation (Silva & White, 2013) has identified several key indicators of student success in developmental mathematics courses in community colleges by using brief, practical measures of student beliefs and attitudes about learning. These indicators have then been used to develop a set of interventions to boost students' expectancies and reduce perceived costs of learning mathematics. This work has contributed to an increase in the completion rate of developmental math courses from 15% to 50% in just over three years.

Limitations and Future Research

Despite the promising results present in the current study, there are several limitations to this work. First, further testing is needed to assess whether comparisons should be made across gender without latent variable modeling. We plan to conduct qualitative studies to aid in further refinement of the EVC Scale. Such data could provide valuable information about the differences between how boys and girls complete the EVC Scale. However, scalar invariance results were strong, implying that the EVC Scale can be readily used by researchers to investigate gender differences in motivation among middle school students with latent modeling.

Second, although the global fit indices were generally good or acceptable, there was some indication of local model misfit. The presence of such local misfit suggests that, although the EVC Scale works well overall, there are some specific items that could be further revised to improve the scale. Specifically, we found presence of some positive correlation residuals (which can suggest item redundancy), and presence of negative correlation residuals (which can suggest a lack of unidimensionality). However, the relatively inconsistent pattern of residuals may also indicate that the large residuals are simply chance findings. Therefore, a next step is to cross-validate these models on another sample.

Third, small sample sizes for sub-groups of students limited our ability to conduct more fine-grained invariance testing, such as comparing sixth, seventh, and eighth grade. These small sample sizes restrict the generalizability of our overall models because the observation/parameter ratios were large. For example, the gender invariance models compared groups relatively small groups of students in the Fall (n_{boys} = 147, n_{girls} = 164) and Winter (n_{boys} = 125, n_{girls} = 145). We also note that because of the small sample of students for which science achievement and interest data were available, these results need to be replicated in larger samples, and with additional age groups and domains.

Fourth, we did not have access to students’ classroom membership and could not account for the nested structure of the data. Ignoring nesting can result in smaller standard errors and increased Type I error rates as the intraclass correlation (ICC) increases. As such, the effects of nesting on the EVC scale need to be addressed in future studies.

Finally, it will be important to determine how well the EVC Scale functions in other academic domains and samples, as well as to examine changes in student motivation over periods of time longer than were investigated in the current study. All of the frequency distributions for expectancy and cost items were skewed in the current sample, resulting in relatively off-center means, potentially masking change in some participants. Low usage of a response category also could indicate that there may be too many response categories, or that they need to be changed. In prior samples (Getty et al., 2013), the full range of responses was used and the distributions were more normal. Thus, further testing with middle school samples is required to assess the efficacy of different response scales and students' understandings of them. It is possible that the time at which students are measured could also affect responses. For example, testing key transition points, such as at the beginning of middle school or high school, may highlight students who are at particular risk for negative educational outcomes.

Conclusion

In this study, we present initial evidence for a tool that can be used by researchers, practitioners, and program evaluators to get a pulse of three types of domain-specific motivation: expectancy, value, and cost. In addition, we offered a blended approach to measuring motivation that balanced traditional psychometric standards of what makes a good scale along with practical considerations to ensure it could be useable by a wider range of stakeholders, most notably practitioners. We present this tool for researchers, practitioners, and evaluators to determine where motivational interventions could be targeted and potentially to assess the effectiveness of interventions on motivation after they are administered.

References

Barron, K. E., Grays, M. P., Flake, J. K., Hogan, E. A., Lazowski, R. A., Pohto, P. A, … & Hulleman, C. S. (2011, May). *What matters for college students' motivation? Two qualitative studies*. Poster presented at the 4th Annual Meeting of the Society for the Study of Motivation, Washington, DC, May 26.

Barron, K. E., & Hulleman, C. S. (in press). Expectancy-Value-Cost model of motivation. To appear in J. S. Eccles & K. Salmelo-Aro (Eds.), *International Encyclopedia of Social and Behavioral Sciences, 2nd Edition: Motivational Psychology*. Elsvier.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*(1), 246-263. doi: 10.1111/j.1467-8624.2007.00995.x

Brophy, J. E. (2010). *Motivating students to learn*. New York, NY: Routledge. doi: 10.1111/j.1467-8535.2010.01135_2_1.x

Casillas, A., Robbins, S., Allen, J., Kuo, Y. L., Ann Hanson, M., & Schmeiser, C. (2012). Predicting early academic failure in high school from prior academic achievement, psychosocial characteristics, and behavior. *Journal of Educational Psychology, 104(2)*, 407- 420. doi: 10.1037/a0027180

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. doi: 10.1207/s15328007sem0902_5

Cohen, J., McCabe, L., Michelli, N. M., & Pickeral, T. (2009). School climate: Research, policy, practice, and teacher education. *The Teachers College Record*, *111*(1), 180-213.

Conley, A.M. (2012). Patterns of motivational beliefs: Combining achievement goals and expectancy-value perspectives. *Journal of Educational Psychology*, *1*, 32-47. doi: 10.1037/a0026042

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology*, *98*(2), 382-393. doi: 10.1037/0022-0663.98.2.382

Dweck, C. S. (2012). Mindsets and human nature: Promoting change in the Middle East, the schoolyard, the racial divide, and willpower. *American Psychologist, 67(8)*, 614. doi: 10.1037/a0029783

Easton, J.Q. (2013, June). *Using measurement as leverage between developmental research and educational practice*. Center for the Advanced Study of Teaching and Learning, Charlottesville, Virginia.

Eccles, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., Meece, J.L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J.T. Spence (Ed.)*, Achievement and achievement motivation* (pp. 74-146). San Francisco, CA: W. H. Freeman.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin, 21(3),* 215-225. doi: 10.1177/0146167295213003

Eccles, J. S., & Wigfield, A. (2000). Schoolings influences on motivation and achievement. In S. Danzinger and J. Waldfogel (Eds.), *Securing the future: Investing in children from birth to college* (pp. 153-181). New York, NY: Russell Sage Foundation.

Feather, N. T. (1988). Values, valences, and course enrollment: Testing the role of personal values within an expectancy-valence framework. *Journal of Educational Psychology, 80*(3), 381-391. doi: 10.1037/0022-0663.80.3.381

Flake, J. K. (2012). *Measuring cost: The Forgotten component of expectancy value theory*. Unpublished Master's thesis. James Madison University.

Flake, J. K., Barron, K. E., Hulleman, C. S., Lazowski, R. A., Grays, M. P., & Fessler, D. (2011, May). *Evaluating cost: The forgotten component of expectancy-value theory*. Poster presented at the 23rd Annual Convention for the Association for Psychological Sciences, Washington, DC, May 26-29.

Getty, S.R., Hulleman, C. S., Barron, K. E., Stuhlsatz, A. M., & Marks, J.C. (2013). *Factors that Affect Learning in High School Science; Measuring Motivation, Achievement, and Interest in Science*. Paper presented at the meeting of the National Association for Research in Science Teaching, San Juan, Puerto Rico.

Global Science Forum. (2006). *Encouraging Student Interest in Science and Technology Studies*. OECD Publishing.

Grays, M. P. (2013). *Measuring motivation for coursework across the academic career: A longitudinal invariance study*. Unpublished doctoral dissertation. James Madison University: Harrisonburg, VA.

Gresham, F. M. (2004). Current status and future directions of school-based behavioral interventions. *School Psychology Review*, *33*(3), 326-343. http://www.nasponline.org/publications/spr/index-list.aspx

Gresham, F. M., & Lopez, M. F. (1996). Social validation: A unifying concept for school-based consultation research and practice. *School Psychology Quarterly, 11*(3), 204-227. doi: 10.1037/h0088930

Hooper, D., Coughlan, J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6(1),* 53-59. http://www.ejbrm.com/main.html

Hulleman, C. S., Godes, O., Hendricks, B., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102(4)*, 880–895. doi: 10.1037/a0019506

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science, 326*(5958), 1410-1412. doi: 10.1126/science.1177067

Iacobucci, D. (2009). Everything you always wanted to know about SEM (structural equations modeling) but were afraid to ask. *Journal of Consumer Psychology*, *19*(4), 673-680. doi: 10.1016/j.jcps.2009.09.002

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children’s self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, *73*(2), 509-527. doi: 10.1111/1467-8624.00421

Kline, R.B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd Edition). New York: Guilford.

Lazowski, R. A., Hulleman, C. S., Barron, K. E., & Getty, S. (2012, September). *Development of an expectancy-value scale for an online science curriculum*. Paper presented at the Motivation Retreat, University of Tubingen, Germany.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying latent variables in SEM and MACS models. *Structural Equation Modeling*, *13,* 59-72. doi:10.1207/s15328007sem1301_3

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22(2),* 366-381. doi: 10.1037/a0019225

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. doi: 10.1037/0003-066X.50.9.741

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology*, *25*(1), 3-53. doi: 10.1006/ceps.1999.1019

National Research Council [NRC] (2010). *Rising Above the Gathering Storm, Revisited: Rapidly Approaching Category 5.* Washington, DC: The National Academies Press.

Parsons, J. E. (1980) *Self-perceptions, task perceptions, and academic choice: Origins and change*. Unpublished final technical report to the National Institute of

Education, Washington, DC. (ERIC Document Reproduction Service No. ED 186577)

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667-686. doi: 10.1037/0022-0663.95.4.667

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354-373. doi: 10.1037/a0029315

Satorra, A., & Bentler, P.M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75,* 243-248. Doi: 10.1007/s11336-009-9135-y

Silva, E., & White, T. (2013). *Pathways to improvement: Using psychological strategies to help college students master developmental math.* Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Snipes, J., Fancsali, C., & Stoker, G. (2012). *Student academic mindset interventions: A review of the current landscape*. San Francisco: The Stupski Foundation.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*(5), 893-898. doi: 10.1016/j.paid.2006.09.017

Stewart, R. M., Benner, G. J., Martella, R. C., & Marchand-Martella, N. E. (2007). Three-Tier Models of Reading and Behavior A Research Review. *Journal of*

Positive Behavior Interventions, 9(4), 239-253. doi: 10.1177/10983007070090040601

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology*, *104(3)*, 763-777. doi: 10.1037/a0027470

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70. doi: 10.1177/109442810031002

Wigfield, A., & Cambria, J. (2010). Expectancy-value theory: Retrospective and prospective. In Urdan, T.C. & Karabenick, S.A. (Eds.), *The decade ahead: Theoretical perspectives on motivation and achievement, 16,* (pp. 74-146). Emerald Group Publishing Limited.

Yang, Y., & Green, S. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29,* 377–392. doi:10.1177/0734282911406668

Yeager, D., Muhich, J., Hausman, H., Morales, L., & Bryk, A. (in press). The case for practical measurement. *Review of Educational Research.*

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They’re not magic. *Review of Educational Research*, *81*(2), 267-301. doi: 10.3102/0034654311405999

Table 1.
CFA Global Fit and Local Fit Summary Values for Math and Science in Winter and Fall

	Model Tested	χ^2	***df***	**RMSEA**	**CFI**	**SRMR**	**Residuals**		$\Delta\chi^{2\dagger}$
							^{3}SR	^{4}CR	
Math Fall 2012	**Model 1**	296.30*	35	.16	.72	.12			
	^{1}**Model 2**	274.63*	34	.15	.74	.12			
	^{2}**Model 3**	85.10*	34	.07	.95	.04			
	Model 4	*30.46*	*32*	*.00*	*1.00*	*.02*	*0%*	*7%*	*56.96**
Science Fall 2012	**Model 1**	362.65*	35	.17	.66	.12			
	^{1}**Model 2**	314.15*	34	.17	.70	.10			
	^{2}**Model 3**	97.99*	34	.08	.93	.04			
	Model 4	*44.87*	*32*	*.04*	*.99*	*.03*	*4%*	*4%*	*53.82**
Math Winter 2013	**Model 1**	289.98*	35	.16	.68	.10			
	^{1}**Model 2**	398.13*	34	.20	.57	.12			
	^{2}**Model 3**	189.69*	34	.13	.81	.08			
	Model 4	*40.58*	*32*	*.03*	*.99*	*.03*	*7%*	*7%*	*147.36**
Science Winter 2013	**Model 1**	273.02*	35	.16	.72	.11			
	^{1}**Model 2**	187.49*	34	.13	.81	.09			
	^{2}**Model 3**	74.71*	34	.07	.95	.04			
	Model 4	*42.70*	*32*	*.04*	*.99*	*.03*	*9%*	*0%*	*35.65**

Note. Total N = 401. n = 311 for Math and Science in Fall 2012. n = 270 for Math and Science and Winter. * $p < .05$. Italicized lines indicate championed models. [1] This model tests Eccles' original two factor model in which expectancy items form one factor while value and cost items form a second factor. [2] This model tests a conventional wisdom two factor model in which expectancy and value items form one factor while and cost items form a second factor. [3] SR indicates the percent of standardized residuals greater than |3.00|. [4] CR indicates the percent of correlation residuals greater than |.10|. †Difference tests were only conducted to compare models that displayed acceptable fit. Models 1 and 2 displayed inadequate fit in all cases and fit values.

Table 2.
Invariance Modeling Global Fit Indices, Local Fit Summaries, and $\Delta\chi^2$ Change

Invariance Models		Global Fit					Local Fit		Change	
	Model^1	χ^2	*df*	**RMSEA**	**CFI**	**SRMR**	**SR^2**	**CR^3**	$\Delta\chi^2$	*df*
Gender	**C**	94.88*	64	.06	.97	.04				
Math	**M**	104.53*	71	.06	.97	.06			9.62	7
Fall	_**S**_^{4}	_115.46*_	_78_	_.06_	_.97_	_.07_	_2%_	_19%_	_11.13_	_7_
	_**E**_^{4}	_121.68*_	_88_	_.05_	_.97_	_.09_	_0%_	_24%_	_13.71_	_10_
Gender	**C**	81.32	64	.04	.98	.04				
Science	**M**	85.80	71	.04	.99	.05			4.30	7
Fall	_**S**_	_101.28*_	_78_	_.04_	_.98_	_.06_	_6%_	_20%_	_18.81*_	_7_
	E	126.92*	88	.05	.96	.10	4%	31%	39.59*	10
Gender	**C**	89.28*	64	.05	.97	.05				
Math	**M**	91.55	71	.05	.98	.05			1.96	7
Winter	_**S**_	_101.84*_	_78_	_.05_	_.97_	_.06_	_2%_	_17%_	_10.70_	_7_
	E	120.10*	88	.05	.97	.10	8%	20%	30.95*	10
Gender	**C**	57.53	64	.00	1.00	.03				
Science	**M**	65.30	71	.00	1.00	.06			7.78	7
Winter	**S**	69.98	78	.00	1.00	.07	1%	12%	3.88	7
	_**E**_	_81.71_	_88_	_.00_	_1.00_	_.07_	_3%_	_9%_	_24.68*_	_10_
Academic-	**C**	157.41	145	.02	1.00	.03				
Domain	**M**	172.07	155	.02	.99	.05			14.37	10
Fall	**S**	235.40*	159	.04	.97	.05	4%	6%	136.22*	4
	_**E**_	_253.01*_	_169_	_.04_	_.97_	_.06_	_3%_	_6%_	_16.29_	_10_
Academic-	**C**	213.97*	145	.04	.97	.04				
Domain	**M**	218.63*	155	.04	.97	.05			6.54	10
Winter	**S**	256.45*	159	.05	.96	.05	3%	7%	93.72*	4
	_**E**_	_284.29*_	_169_	_.05_	_.95_	_.06_	_3%_	_11%_	_20.90*_	_10_
	C	178.83*	145	.02	.99	.03				
Longitudinal	**M**	191.15*	155	.02	.99	.05			12.32	10
Math	**S**	214.29*	159	.03	.98	.05	4%	6%	49.07*	4
	_**E**_	_234.03*_	_169_	_.03_	_.98_	_.05_	_4%_	_6%_	_17.35_	_10_
	C	195.14*	145	.03	.98	.03				
Longitudinal	**M**	201.07*	155	.03	.98	.05			6.35	10
Science	**S**	206.41*	159	.03	.98	.05	1%	4%	5.52	4
	_**E**_	_223.69*_	_169_	_.03_	_.98_	_.05_	_2%_	_6%_	_14.84_	_10_

Note. *p < .05 ^{1} C = Configural Model, M = Metric Model, S = Scalar Model, E = Equal Error Model. ^{2} "SR" refers to the percentage of standardized residuals greater than |3.0|. ^{3} "CR" refers to the percentage of correlation residuals greater than |.10|. Italicized lines indicate championed models. ^{4}Gender in Fall math shows two plausible models.

Table 3.
Latent Correlation Matrices For Academic Domain Invariance and Longitudinal Invariance

Academic Domain Invariance by Fall (below diagonal) and Winter (above diagonal)[1]

		Math			Science		
		Expectancy	**Value**	**Cost**	**Expectancy**	**Value**	**Cost**
Math	**Expectancy**		.55	-.62	.38	.31	-.29
	Value	.80		-.54	.36	.43	-.27
	Cost	-.45	-.50		-.35	-.36	.57
Science	**Expectancy**	.29	.19	-.28		.86	-.63
	Value	.26	.32	-.31	.82		-.56
	Cost	-.29	-.28	.57	-.56	-.47	

Longitudinal Invariance by Math (below diagonal) and Science (above diagonal)[2]

		Time 1 (Fall)			*Time 2 (Winter)*		
		Expectancy	**Value**	**Cost**	**Expectancy**	**Value**	**Cost**
Time 1	**Expectancy**		.85	-.54	.68†	.64	-.60
(Fall)	**Value**	.80		-.46	.56	.73†	-.51
	Cost	-.51	-.57		-.48	-.45	.62†
Time 2	**Expectancy**	.74†	.49	-.55		.84	-.64
(Winter	**Value**	.59	.75†	-.50	.64		-.58
	Cost	-.50	-.51	.82†	-.60	-.56	

Note. N = 401. [1]Values below the diagonal in the Domain correlations represent the Fall Sample (n = 311) and values above the diagonal represent the Winter Sample (n = 270). [2] Values below the diagonal in the Longitudinal correlations represent Math and values above the diagonal represent Science. † Denotes latent test-retest correlations for the construct.

Figure 1. Competing Factor Structures of the EVC Scale

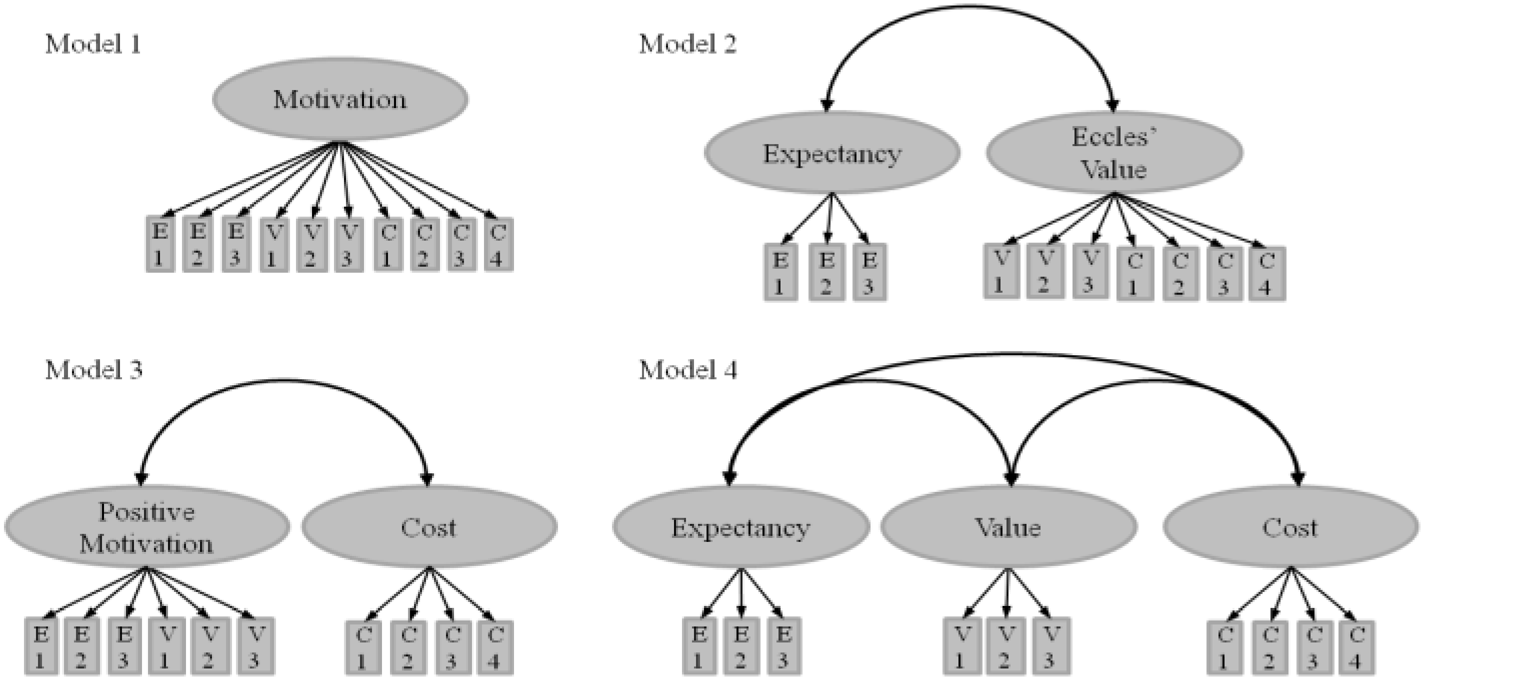

Figure 1. Diagram showing four possible models for the relationships among expectancy, value, and cost. Model 1tests a single-factor model comprising all ten items. Model 2 tests Eccles' original two-factor model in which expectancy items form a single factor while value and cost items form another factor. In Model 3,expectancy and value items form a single factor, and cost items form a separate factor. Model 4 tests a framework in which Expectancy, Value, and Cost are distinct latent factors (i.e., the EVC Scale)

Paper 3: A Pragmatic Measurement Approach: Using Argument-Based Validation in Applied Education Sciences

Jeff J. Kosovich and Chris S. Hulleman

University of Virginia

Jessica K. Flake

York University

Abstract

Measurement researchers have compiled decades of research on ideal validation practices (i.e., ensuring the quality a measure). However, state-of-the-art perspectives on measurement are not necessarily integrated with common practice, particularly in applied research domains. The result is a tension between practical needs and technical standards. We introduce the pragmatic measurement approach. Though pragmatic measurement is largely an adaptation of contemporary thinking on argument-based validation, the notable changes are an explicit focus on the practical, contextual factors that can undermine measurement quality and heavier reliance on theory for predictions. The current study tests a pragmatic measurement approach to validating a short measure (three single-item scales) of students' expectancy, value, and cost in two samples from community college math. We identified four typical uses of such scales in motivation research. To validate these four uses we collected validity evidence that the measures could be used as intended. Overall, it was possible to identify instances where single-item scales were appropriate and instances where they were not. Moreover, reducing the measure from 13 to 3 items only produced a 4% loss of variance explained in students' self-reported interest. Multi-dimensionality appeared to be the main culprit when inconsistent results were present, though it is important to consider how shorter scales may interact with other study characteristics (e.g., sample size). We recommend the pragmatic approach as a validation method for low-stakes contexts when users can compromise between technical standards and practical needs.

Pragmatic Measurement: Using Argument-Based Validity in Applied Education Sciences

Education researchers often require quick and efficient assessments of various student characteristics (e.g., motivation) to use in classroom settings. Unfortunately, guidelines for addressing measurement obstacles, such as scale length, are ambiguous at best and non-existent at worst (Csikszentmihalyi & Larson, 2014; Deno, 1985; Gogol et al., 2014; Stanton, Sinar, Balzer, & Smith, 2002; Yeager, Walton, & Cohen, 2013). Many measures may be questionable because of a general lack of evidence for the quality of the data they produce (Paulhus & Vazire, 2007), and measures with few items have received particular criticism from measurement experts (e.g., Subar et al., 2001; Widaman, Little, Preacher, & Sawalani, 2011). To further strain matters, there is a growing call to action to improve the quality of validity evidence for measurement in theory, research, and practice (Flake, Pek, & Hehman, 2017; Graham, 2015). The result is a tension between technical and practical constraints when conducting measurement in field research (Yeager, Bryk, Muhich, Hausman, & Morales, 2013). To address these tensions, the technical and practical sides of measurement need to be considered.

The current paper is a step towards developing a coherent set of measurement recommendations in educational science through the lens of *pragmatic measurement*. Pragmatic measurement (Kosovich, Hulleman, Barron, & Getty, 2015; Yeager, Bryk, et al., 2013) is an argument-based validation approach that is theoretically-based, minimally intrusive, and maximally informative. Argument-based validation (Cronbach, 1988; Kane, 2013b) requires scale users to make explicit how the scale will be used, and to provide evidence that those uses are supported. Pragmatic measurement is based largely on contemporary measurement philosophy, but emphasizes the consideration of contextual constraints and the need for strong theoretical a-priori predictions. This research is aimed at alleviating the tensions between the technical and practical aspects of measurement in motivation research. The methods adopted represent a proof of concept for the pragmatic measurement approach and a step towards cohesive guidelines for use. We discuss the presence or absence of validity evidence for a measure of motivation using the pragmatic measurement perspective. Importantly, we discuss the need to consider the validity evidence in aggregate, rather than rely on any single analysis.

Measurement Validation: Disconnects and Definitions

Despite the centrality of measurement to education science, measurement experts have discussed the need for a stronger scientific basis in the field of measurement as recently as 10 years ago (Schmeiser & Welch, 2006). Schmeiser and Welch note that test developers long viewed scale development as more of an art than a science. Though the measurement field has progressed over the past two decades in developing its scientific foundations, the advances come at an inopportune moment. Unfortunately, research found a severe shortage of methodologists and methodological training in the measurement field between 1990 and at least 2008 (Aiken et al., 1990; Aiken, West, & Millsap, 2008; Clay, 2005). As a result, it is likely that many measurement advances have not been widely disseminated due to the shortage of experts in the field.

For example, a survey of psychology PhD programs in the U.S. revealed that only 64% of doctoral programs offered any sort of measurement courses, and only 24% offered a full course devoted to the topic, with a total median of 4.5 weeks of measurement training in the typical PhD curricula (Aiken et al., 2008). Also concerning was the fact that less than half of the survey respondents indicated their students could appropriately assess the quality of measurement tools. In their earlier publication, Aiken and colleagues (Aiken et al., 1990) noted that one of the primary sources of statistical and measurement training was from students' mentors—a troubling notion for the dissemination of cutting-edge methodology. Why does the state of measurement training and knowledge matter? The lack of current knowledge about measurement theory and the lack of strong guidelines for validation in applied settings may be a reason that researchers and other users rarely address validity in the literature (e.g., Flake et al., 2017).

A Brief Primer on Validity

According to the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education (NCME), *validity* is the "degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). Validity theory has gone through numerous revisions in its history before arriving at contemporary argument-based approaches (Cronbach, 1988; Kane, 1992, 2013a). Early conceptualizations of validity focused primarily on *criterion evidence* or the degree to which a measure correlated with or predicted some outcome; the criterion approach was later deemed insufficient (Cronbach, Lee & Meehl, 1955). The following era of measurement theory and use focused around specific types of validity that confirmed different aspects of a given test, such as using experts to assess the *content* of the scale, or factor analyses to assess the *structure* of the scale. These eras of measurement theory were developed mostly for the purposes of accountability testing and theory building

(Yeager, Bryk, et al., 2013). During this time, various types of validity were seen as important to specific types of tests and unimportant to others (Messick, 1989). Messick noted several examples in which different validity types were tied to a particular testing aim:

> For example, content validity was deemed appropriate to support claims about an individual's present performance level in a universe of tasks or situations, criterion-related validity for claims about a person's present or future standing on some significant variable different from the test, and construct validity for claims about the extent to which an individual possesses some trait or quality reflected in test performance. (1989, p. 6)

Most recently, measurement experts have argued for a unified conceptualization of validity (Messick, 1989) in which test developers and users compile multiple types of evidence to support the intended uses and interpretations of the tests (Kane, 2013a). While the methods (some of which will be described later) for obtaining validity evidence have remained relatively similar to those described half a century ago, the philosophy behind the validation process as well as the curation of validity evidence has changed drastically. We introduce pragmatic measurement as one potential approach for validation in applied settings. Though it is largely based on contemporary thinking on validation, the pragmatic measurement approach has two notable changes: an explicit focus on the contextual factors that can undermine measurement quality, and a heavier reliance on theory to make a-priori predictions about construct relationships as a substitute for more typical reliability and validity information.

Argument-Based Validity: Use and Interpretation

The pragmatic measurement approach focuses on building evidence-based arguments to support specific uses and interpretations of a scale (Kane, 1992, 2013a, 2013b). In order to develop a validity argument, it's helpful to consider the *overall purpose* that is driving the need to conduct measurement. The overall purpose simply refers to the research design and research questions of a particular data collection endeavor (e.g., to determine if students are proficient in mathematics). Next, measure users need to specify the intended *use* of the measures (Chapelle, Enright, & Jamieson, 2010). In other words, users need to articulate how the measures will support the overall purpose (e.g., the math test will be used to classify students into proficiency groups). However, each intended use makes several *assumptions* about the characteristics of the measure. The assumptions delineate what must be true of the measure in order for sound interpretations of the data (e.g., the measure can accurately classify students into different proficiency groups). To demonstrate that the assumptions are met, users need to collect evidence from various *sources of validity*. Sources of validity are the empirical and theoretical supports present in the development and ongoing utilization of the measures (e.g., experts have reviewed the measure and repeated piloting shows it can discriminate between proficiency levels). In summary, the general research purpose is supported by specific uses of a measure, the uses make several assumptions about the characteristics of a measure, and validity evidence is collected to assess whether or not the assumptions are met.

One example of insufficient validation comes from a meta-analysis by Hulleman and colleagues (Hulleman, Schrager, Bodmann, & Harackiewicz, 2010) in which conflicting results from various studies were attributed to differences in wording of the items and manner in which constructs were operationalized (Marsh, 1994). Arguably, better validity evidence (e.g., more construct specificity) could have prevented confusion in the motivation field. It is worth noting that the pivotal difference in pragmatic measurement and standard measurement approaches is primarily in how validity evidence is collected and curated. In the following section, we review several areas where researchers had specific research goals that required a pragmatic approach to collecting validity evidence.

Collecting Validity Evidence for Pragmatic Measures

To identify validity sources that can be used for pragmatic measurement, we examined three separate programs of research with varying overall purposes. Considering how each group of researchers supported the uses of their particular measures demonstrates how others have approached measurement pragmatically. Table 1 summarizes four major sources of validity evidence. The first, *differences*, refers to a-priori hypotheses between various contrasted groups. Naturally existing groups (e.g., gender) are sometimes characterized by different levels of a construct. Experimental groups may be theorized to show construct differences after the experimental stimulus is introduced. Other constructs change with time in expected ways (i.e., repeated measures).

The second, *relationships,* refers to a-priori hypotheses regarding the correspondence between multiple constructs in different forms. Expected correlations show that interrelationships between constructs are of the expected direction and intensity. Test-retest correlations show that the construct is relatively stable or not across time points. Prediction demonstrates that a scale can meaningfully explain variation in another construct. The third, *proxy scales,* are simple approximations of longer scales or are indirect measures that are thought to be highly correlated with variation in a construct. Finally, *targeted expert knowledge,* helps to guide scale selection and scale reduction, and to identify expected differences and expected relationships. The sections below consider these sources of validity evidence collected in the context of field research programs, including improvement science, experience sampling, and curriculum-based measurement.

Improvement science. Yeager and colleagues (Yeager, Bryk, et al., 2013) provide the most complete description of pragmatic measurement for the purposes of improvement science. *Improvement science* (also known as continuous improvement) refers to a systematic approach to solve a problem in practice through brief iterative cycles of testing and study (Lewis, 2015). Compared to typical programs of applied research, improvement science focuses on the ultimate goal of effecting change in practice and implementation (Bryk, Gomez, Grunow, & LeMahieu, 2015). In pursuing measurement for improvement science, Yeager and colleagues proposed an approach called *practical measurement*. We view the pragmatic measurement approach as a generalization of Yeager and colleagues' practical measurement—their research serves as a foundation for the thinking presented in the current paper.

They identified a number of specific methods for collecting validity evidence to ensure alignment between intended uses and measures (Yeager, Bryk, et al., 2013). First, to ensure the content of their items was theoretically sound, they conducted an extensive review of existing scales for each of their constructs of interest (Table 1: Targeted Expert Knowledge). Existing items were selected or adapted based on theoretical bases and empirical evidence that they effectively represented their respective construct. Their second goal was focused on measuring construct change following a new potential improvement to classes (Table 1: Differences—Experimental). Student mindsets (i.e., attitudes about personal behaviors and beliefs about math) were measured at the beginning of the semester and after a mindset intervention was introduced. They examined mean differences between scales before and after the intervention, which showed that negative mindsets (e.g., anxiety) decreased and positive mindsets (e.g., interest) increased, as would be hypothesized based on theory (e.g., Table 1: Differences – Change Over Time, Experimental). Their third goal focused on predicting student success in math classes. Predictive analyses suggested that their brief scales were able to identify students who were most at-risk for failing to complete the course (Table 1: Relationships—Prediction). Yeager and colleagues work provides a detailed example of the benefits of pragmatic measurement for the purposes of improvement science.

Experience sampling. The *experience sampling methodology* was developed to assess frequencies and patterns of psychological states, daily life experiences, and thinking (Csikszentmihalyi & Larson, 2014). This methodology exemplifies the idea of minimal intrusion because the method was designed to be smoothly integrated into everyday life. Although the repeated process can be taxing on participants, any single assessment is minimized to prevent disruption. Experience sampling collects brief assessments from individuals throughout days, weeks, or months using a beeper or similar signaling technology in combination with a data-recording apparatus (e.g., notebook, cellphone). The method is so efficient at collecting data that the amount of data that could be collected was even considered a disadvantage of the method (Csikszentmihalyi & Schiefele, 1994; Hormuth, 1986) before modern computing technology was available (Dimotakis, Ilies, & Judge, 2013).

Experience sampling methods also use limited-length scales, with one recommendation requiring two minutes or less per measurement occasion (Csikszentmihalyi & Larson, 2014). Because these scales are collected frequently, the researchers also acknowledge that multi-item scales would likely lead to participant disengagement given the frequency of collection. As a result, the methods for assessing validity and reliability are demonstrative of what pragmatic measurement users may utilize. Ultimately, experience sampling researches may eschew reliability coefficients of internal consistency (e.g., alpha) in favor of test-retest reliability (Table 1: Relationships—Test-retest reliability), or the correlation between a scale at two different time points. An additional method for assessing the validity of the items was to examine multiple group comparisons to see if responses differed in expected ways—e.g., participants without schizophrenia were coded as having more ordered thoughts than participants with schizophrenia (Table 1: Differences—Known Groups). Overall, experience sampling methods are unique in the degree to which scale length is an obstacle.

Curriculum-based measurement. Sharing similarities with both improvement science and experience sampling, curriculum-based measurement (Deno, 1985; Fuchs & Fuchs, 2004; Shinn, 2013) represents another approach to measurement that prioritizes situational and contextual factors. Whereas the experience sampling method is generally a tool for researchers, curriculum-based measurement was developed to aid practitioners and administrators. These models of measurement in practice assess student performance in the context of their experienced curriculum.

Curriculum-based measurement users demonstrated several validation methods for pragmatic measurement. Using targeted expert knowledge, they were able to identify relationships and differences that (if found) could support the validity of their measures (Table 1: Differences, Relationships, Targeted Expert Knowledge). For example, they examined expected gaps between existing groups such as comparing reading fluency among standard and special education students (Deno, Mirkin, & Chiang, 1982). They also used expert knowledge to identify potential proxy scales (Table 1: Proxy Scales) in place of more complicated assessments. Contextually, the researchers were interested in measuring students' reading comprehension (Deno et al., 1982). However, the traditional scales were far too cumbersome to use for intensive longitudinal collection (i.e., many times through the semester). As a result, they tested other indicators that would correlate strongly with more traditional scales, finally settling on the number of mistakes while reading out loud. The combination of these methods provided a strong basis of validity evidence for measurement.

Summary and Importance of Theory

A common theme among the discussed research domains is the problem of scale length. Though we acknowledge that other validation approaches are likely necessary for observation, behavioral, or qualitative methods, we focus our discussion of validation methods to those most relevant to short self-report scales[5]. The general wisdom is to avoid short scales unless no other options are available (Widaman et al., 2011). One caveat to consider is that theoretical and accountability measurement research has

[5] As similar analyses are used for many types of measurement endeavors, these sources of validity evidence are potential options for other measurement choices beyond short measures and self-report.

unevenly spawned research on the merits of longer scales. In each of the programs of research discussed above, the researchers' validity evidence selection was contingent on the overall purpose of their research. The pragmatic measurement approach is a step toward developing explicit guidelines for short scales that is more useful than the recommendation to not use them unless necessary.

The Role of Theory in Pragmatic Measurement

Of critical note in the case of pragmatic measurement is that theory and prior research on a particular domain can help provide predictions about how specific constructs should function when measured. All three of the research domains previously discussed (i.e., improvement science, experience sampling, and curriculum-based measurement) leveraged theory and expert knowledge (researcher or practitioner) to fill in gaps created by obstacles in the field. Theory is often cited as critical for test development in typical validation practice (AERA et al., 2014), but Kane's (Kane, 1992, 2013a) argument-based approach was developed in part to aid those who did not have strong theoretical backing for their measures. Because conducting measurement in the field often requires users to eschew standard measurement practices, theory can help to fill the validity gaps. For example, typical measure development recommendations suggest several iterations of expert feedback and pilot testing before final items are selected (Schmeiser & Welch, 2006). However, measure users often lack the time or resources to conduct long-term content validation studies. Strong theory or prior research on how the construct of interest relates to other variables, and how it typically differs among groups, can help to make predictions before the measure is deployed (Yeager, Bryk, et al., 2013). To this point we have discussed practical measurement in the abstract, but the specific approaches that users adopt will depend on the overall purposes of their own research—the validation is always focused on a specific construct. In the current research, adopt the pragmatic approach validate a measure of motivation.

Setting the Stage: Achievement Motivation in Education

The expectancy-value-cost framework (Barron & Hulleman, 2015) is a version of more general models of expectancy-value frameworks (Atkinson, 1964; Vroom, 1964) adapted for the education context (Eccles et al., 1983). According to the expectancy-value-cost framework, students' achievement and choices are most proximally predicted in part by three key factors: expectancies for success, subjective task values, and perceptions of cost. *Expectancies for success* refer to individuals' beliefs in their ability to succeed in a situation or task. *Subjective task values* refer to the importance, usefulness, or enjoyment an individual associates with a situation or task. Finally, *perceptions of cost* refer to the perceived psychological, temporal, or effort-based obstacles that prevent an individual from succeeding. Expectancies, values, and costs are determined by a multitude of factors from the general cultural milieu, to parental attitudes, to past achievement experiences (Eccles et al., 1983). The framework is helpful for organizing motivational theories and constructs in more practice-based contexts (Hulleman, Barron, Kosovich, & Lazowski, 2016; Murphy & Alexander, 2000; Pintrich, 2003).

Primarily adapted by developmental researchers, much of the early expectancy, value, and cost research[6] focused on change in student motivation during foundational

[6] We note that for the purposes of this paper, we refer to the body of expectancy-value research based off of the framework adapted by Eccles and colleagues and later refined by Barron and Hulleman. There are

academic years. As the framework became more prominent in the field, it was applied to older students up to and beyond college years. Because of its roots in developmental and educational research, expectancy, value, and cost studies are typically characterized by two important features—they are often longitudinal in design, and they are almost always conducted in naturalistic settings (i.e., they are not laboratory-based). The resulting scales developed for studying the expectancy-value-cost framework serve as examples of theory-based measures that were borne of pragmatic considerations.

Expectancy, value, and cost studies span a range of achievement domains and utilize numerous psychological constructs. For example in their seminal measurement paper, Wigfield and colleagues (Wigfield, Harold, & Blumenfeld, 1993) measured student perceptions of ability (3 items), expectancy (2 item), difficulty (1 item), usefulness (1 item), importance (1 item), and liking (2 items) in math, reading, and sports for a total of 10 items per domain. The resulting set of scales was 30 items that exclusively focused on expectancy, value, and cost constructs. As a part of a larger longitudinal study, the data were eventually aggregated to examine the growth trajectories of student motivation from first grade through the end of high school (Archambault, Eccles, & Vida, 2010; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). Over the past three decades, these scales directly or indirectly contributed to an influential body of educational research. However, during this time, expectancy, value, and cost researchers adopted practices that were not necessarily in line with measurement experts' recommendations.

numerous expectancy-value models including those described by Lewin (Lewin, Dembo, Festinger, & Sears, 1944), Atkinson (Atkinson, 1964), and Vroom (Vroom, 1964).

The result of the departure from recommended measurement practice was a need for more validity evidence. Specifically, two aspects of the expectancy, value, and cost scales required attention. First, reliability (i.e., the reproducibility of responses) of brief scales will likely be worse than a longer scale, or uncalculatable for single-item scales (Crocker & Algina, 1986). Second, it is difficult to provide strong validity evidence that the construct breadth is adequately addressed with only a few items per scale (Widaman et al., 2011). The expectancy, value, and cost researchers (Eccles et al., 1983) partially addressed these validity concerns through their theoretical work. Consider that their 30+ items comprised 18 different scales (six motivational constructs in three different domains) resulting in a number of conceptually distinct but related scales. For example, the subjective task-value construct can actually be divided into three (Barron & Hulleman, 2015; Eccles et al., 1983): attainment value—the inherent importance of a task to an individual; intrinsic value—the enjoyment an individual experiences from the task; utility value—the perceived usefulness of the task to some future goal. These specific sub-constructs that described value were correlated highly enough that they could provide stable reliability estimates, yet were also able to be used as unique scales (Simpkins, Davis-Kean, & Eccles, 2006; Wigfield et al., 1997; Wigfield, Eccles, Mac Iver, Reuman, & Midgley, 1991; Wigfield et al., 1993). The benefit of this approach was that the researchers had theoretically defined value in such a way that the unique conceptual aspects of value were themselves smaller constructs. From a technical perspective, this meant that using several items to measure the sub-constructs was unnecessary because they were unidimensional (Wigfield & Eccles, 2002) or because the items captured a sufficient amount of variance in the overarching construct (Gaspard et al., 2015).

Whether the researchers were aware of these properties during measure development is unclear, as there are few published validity studies on expectancy, value, and cost measures (Eccles & Wigfield, 1995; Flake, Barron, Hulleman, McCoach, & Welsh, 2015; Gaspard et al., 2015; Kosovich et al., 2015; Perez, Cromley, & Kaplan, 2014). For the early research questions in the field, the measures appeared to work effectively. It is critical as the field moves forward and branches into new lines of inquiry that good validation practices are used. The current paper is both a test of the pragmatic measurement approach, and is also an attempt to examine the validity evidence for a measure of expectancy, value, and cost.

Current Research

The pragmatic measurement approach is aimed at the spectrum of education scientists and practitioners who wish to assess students. The purpose of the current study was to test the practical measurement perspective as a validation method for single-item measures. We did so by examining the evidence supporting four typical uses of expectancy, value, and cost scales: measuring those particular constructs in classrooms, examining group differences, monitoring change over time, and monitoring intervention effects (Table 2, "Proposed Uses"). Each proposed use can be considered a research question in the current study. To continue our validity argument, we then identified several assumptions that underlie each specific use along with sources of validity that would help to test those assumptions (Table 2, Column 2 and 3).

We derive evidence that these various assumptions are met by examining different sources of validity (Table 1; Table 2, "Validity Sources"). The first two sources of validity are methodological in nature: *Targeted Expert Knowledge,* and *Proxy Scales.*

Recall that targeted expert knowledge refers to the use of expert knowledge for making predictions about scale functioning (e.g., expected correlations). Proxy scales refer to the use of shortened or alternative measures that capture much of the same variance as a standard measure. These sources are discussed in the methods section. The other two sources, *relationships* and *differences*, are analytic in nature and support specific assumptions. Recall that relationships refer to correlations and predictions made using the measures under study. Differences refer to a measures' sensitivity to differences between existing groups (e.g., gender), experimental groups (e.g., treatment vs. control), or measurement occasions (e.g., beginning and end of semester). These sources are discussed in the results section. It is worth noting that evidence supporting the earlier uses also functions as support for later uses. For example, it is important to demonstrate reliability to show that data generated are consistent, but reliability is also important because it affects whether or not more advanced analyses (e.g., monitoring growth) are accurate. In the current study, we compile validity evidence for the use of single-item expectancy, value, and cost measures; in doing so, we test the feasibility of the pragmatic measurement approach.

Method

Participants

The current study included two samples for the purposes of cross-validation and replication. Both samples were drawn from a large southeastern two-year college in the Spring and Fall of 2016. Both samples were a part of a full-semester intervention study and all included students who provided consent for their data to be used in research. Table 3 summarizes the sample characteristics. Sample 1 (n = 740) was relatively diverse in terms of ethnicity, 35% Hispanic/Latino, 35% White, 20% Black, 9% other (Asian, Hawaiian/Pacific Islander, Alaskan, Indian, Mutiracial), 2% unknown or missing. The majority of students were female (58%) and were not receiving Federal Pell grants (54%). Sample 1's median age was 20 (IQR = 19-22), with a minimum age of 18, and a maximum of 59. Sample 2 (n = 1855) differed slightly in terms of demographic characteristics with 39% Hispanic/Latino, 31% White, 23% Black, 7% other (Asian, Hawaiian/Pacific Islander, Alaskan, Indian, Mutiracial), 1% unknown or missing. The majority of students were female (61%) and were not receiving Federal Pell grants (81%). Sample 2's median age was 19 (IQR = 18-21), with a minimum age of 18, and a maximum of 80.

Nested data. Students in Sample 1 were nested within 34 classrooms, which were nested within 20 instructors. Students in Sample 2 were nested within 74 classrooms, which were nested within 30 instructors. Intra-class correlations calculated for student interest at four time points and for pass rates yielded a range of values from .07 to .11 in both samples (see Table 4). Because this variance was interchangeable between the class and instructor levels (i.e., we could not partition the unique variance of both levels), we elected to include classrooms as the nesting indicator. Given the range of ICCs all analyses used adjusted standard errors to correct for dependencies in the data with the TYPE=COMPLEX function in Mplus.

Measures

Motivation. We collected a 13-item measure of student motivation using items adapted or developed from several sources (Eccles et al., 1983; Kosovich et al., 2015). The scale included four expectancy items (e.g., E1, "How confident are you that you can learn the material in this class?"), four value items (e.g., V1, "How important is this class to you?"), and five cost items (e.g., C1, "How often does this class require too much time?"). A full list of expectancy, value, and cost items are included in Table 5. Truncated versions of each scale were used as immediate baseline measures for the intervention time points (described below). These scales are typically averaged to create composite scores for analyses. All three scales used a 5-point response scale ranging from 1 (Not at all) to 5 (Extremely).

Manipulation check. Two additional items were included to measure relevance (e.g., "How relevant is the course material to your future career plans?"). This scale is typically averaged to create composite scores for analyses. This scale used a 5-point response scale ranging from 1 (Not at all) to 5 (Extremely).

Characteristics. Additional data were also collected to support the primary analyses. Students' race, sex, and age were initially collected from the administrative data at the college. Missing values were filled-in where possibly by student reports. Other administrative data included their college GPA, Pell Grant status (1 = Pell, 0 = No Pell), and the number of credits they had earned at the college. Other student-reported data collected included students' self-reported high school GPA, parents' education, and grade in their most recent prior math course. Finally, their experimental assignment was also included (Experimental Group = 1, Control Group = 0).

Dependent variables. Interest and Pass status were collected as outcome or dependent variable measures. We used a 3-item scale (Hulleman, Godes, Hendricks, & Harackiewicz, 2010; Hulleman, Kosovich, Barron, & Daniel, 2016) to measure students' continuing interest (e.g., "How interested are you in learning about careers involving math?”). All scales used a 5-point response scale ranging from 1 (Not at all) to 5 (Extremely). Students final designation for the class was also collected (Pass = 1; Fail, Withdrew, and Incomplete = 0). Because we were interested in students’ passing, Pass calculated as individuals who passed (1) compared to individuals who failed, withdrew, or received incompletes (0).

Procedure

Both samples followed an identical procedure. Data were collected from students four times throughout the semester: Week 1, Week 3, Week 5, and Weeks 14-15. The first and last collection points were longer questionnaires (approximately 40 questions) that included the measures described above. The middle time points (Week 3 and 5) had three phases: (1) a 10-item pre-intervention questionnaire, (2) an experimental activity, and (3) a 5-item post-intervention questionnaire. The experimental activity had two conditions: the value intervention (experimental) and the control activity (control). Students in the *value intervention* read a series of brief prompts and about math’s usefulness to everyday life, future careers, and hobbies and interests. They were then asked to reflect on the quotes and write a brief essay (three to four sentences) for each prompt for a total of three brief essays during the first intervention (Week 3), and a more general essay during the second intervention (Week 5). Students in the *control activity* were asked to summarize the material they were learning in their class during the first activity and asked to describe their study strategies during the second activity. Students received course credit for participation for all of the data collection points. The interventions were required as coursework, though the students were still able to provide consent for the use of their data in research at no penalty.

Missing data. A substantial amount of missing data was present in these samples, ranging from 20% to 75% missingness, which could prove problematic for estimation procedures. In the case of high rates of missingness, it is important to consider if the missingness is random or systematic. In our case, the missingness was systematic and thus violated assumptions necessaries for standard maximum likelihood estimation. However, bias due to non-random missingness can be attenuated by including extra variables in the model that are correlated with the variables under study. We note that higher rates of missingness were correlated with lower pass rates, gender, and ethnicity. The solution in these cases is to include auxiliary variables, which provide additional information for the maximum likelihood estimation. Also important, most students (approximately 95%) provided responses to at least one wave of questionnaire data, and the questionnaire items were highly correlated over time. Thus, although there was missingness at any given time, there was a substantial amount of auxiliary information available to improve model estimation.

Targeted Expert Knowledge and Proxy Scales

Sources of validity are not necessarily explicit analyses; they are also predictions and choices made regarding the measures being examined. For example expert knowledge and proxy scales both function within the research design rather than as analytic techniques. Specifically, we designed the studies and selected specific analyses that would allow us to use expert knowledge to make predictions about the scale’s performance (e.g., expectancy and value will be positively and moderately correlated). Thus, we view departures from expert knowledge as potential evidence that the validity may be compromised. Similarly, expert knowledge was used to determine which single-item proxy scales would be used in the study to represent the motivation constructs. Single-item proxy scales represent the central focus of the current study as we assess the quality of single-item measures.

Single-item selection. Two sets of items were generated for analysis beyond the composite scores. For randomly-selected scales, items within each construct were randomly ranked using a random number generator and the highest ranked of each was selected to represent the scale (Expectancy = E3, Value = V3, Cost = V3; see Table 4). Expert-selected items followed a specific rating procedure. Two experts independently ranked the items on two or three dimensions: (1) how well the item represented the construct, (2) how well the item would predict interest, and (3) how sensitive the item would be to the intervention. The third dimension was only used for value items. Ratings were then summed across raters and dimensions. The highest ranked item from each scale was used to represent the overall construct (Expectancy = E2, Value = V4, Cost = V4; see Table 4) in all analyses where available (see the next section for exceptions). Table 4 indicates which items were selected as single-items as well as those that formed the full composites.

Random vs. alternate items. Random items were selected as a strong test of expert versus arbitrary item selection in the reliability and prediction analyses. However, the longitudinal analyses and experimental analyses required a different set of comparison items. This is because two items from each scale were previously selected to be used as pre-intervention measures during the intervention time points. In these cases, the expert items were present, but the randomly selected items were not. Thus, for the longitudinal and experimental analyses, we compare the expert scale to the *alternate* scale (Expectancy = E1, Value = V1, Cost = C5; see Table 4) rather than the random scale. These alternate scales represent a weaker test of arbitrary selection, but still compare the expert-selected item to another single-item measure. The expert-alternative item contrast is likely to be less stark than the expert-random item contrasts because the alternative items were chosen previously as more optimal. Table 4 indicates which items were designated as alternate items.

A note on composite scores. The composite scales were created by summing the items within each construct (i.e., expectancy, value, and cost). However, Time 2 and Time 3 used truncated scales (two items per scale rather than four) to minimize the potential for frustrating participants who were completing the intervention. As a result, the composite scale for those two times differs from the first and last times. Supplemental analyses suggest that reducing scales from the first and last time points to match does not drastically impact the models, thus we retain the full composite where possible for increased precision. During Time 2 and Time 3, the composite was created from the expert item (E2, V4, C4, respectively) and the alternate item (E1, V1, C4 respectively) of each construct.

Hypotheses and Disconfirming Evidence

As discussed, one integral part of pragmatic measurement is using expert knowledge and predictions to develop a-priori hypotheses about what should and should not be observed. This section generally and briefly describes what patterns were expected and unexpected from the results for each use.

Use 1 hypotheses. Of central concern to these analyses was the replicability of results by the expert scale from the composite scale. We assumed the composite scale to represent our best estimate of each construct, thus departures from the composite to the single-item scales are particularly egregious disconfirmations of validity evidence. We generally expected the expert scale to yield similar conclusions to the composite. In contrast, it was unclear if the random scale would match the other two.

We expected correlations between expectancy and value to be moderate (e.g., .30 to .50) and positive. In contrast, we expected cost to be negatively correlated with expectancy and value. Disconfirming results would have been positive relationships with cost, or a negative relationship between expectancy and value. We also expected the test-retest correlations to be moderate (.40 or higher) depending on the length of measurement intervals; more distant time points should be more weakly correlated. Based on prior work, we expected value to generally show higher test-retest correlations than expectancy. It was not clear how cost might compare. Disconfirming results would show especially low correlations between time points, particularly for those more distant from one another. Finally, in terms of prediction, we expected value to be most predictive of interest, followed by expectancy and cost. In contrast, expectancy was most likely to be predictive of passing, whereas value is often unrelated. As with the prior analyses, there is little guidance on how correlated cost would be with outcome measures.

Use 2 hypotheses. Prior research suggested that women are likely to underestimate their ability even when they perform as well as or better than their male peers. Thus, in terms of differences, we expected women to report lower expectancy while also displaying higher pass rates than men. It was possible that there would be no difference in expectancy, though such findings would be unexpected. Disconfirming evidence of validity would be that women show higher expectancies. For value and cost, there are no established and consistent patterns of gender differences, though if differences did exist, we expected women to show lower motivation (i.e., less value, more cost) than men.

Use 3 hypotheses. It is well established that expectancy and value decline over time. Thus, we predict that the growth models will show a similar pattern in the current study as well. Although there are no explicit studies of cost change over time, we hypothesize an increase in cost given the constructs' negative relationship with the other two constructs. Disconfirming evidence would include rapidly increasing expectancy or value. Research on trajectory variance is mixed with at least one study showing significant variance (Jacobs et al., 2002) and another showing non-significant variance for value (Kosovich, Flake, & Hulleman, 2017). Thus we expect significant variation in expectancy change, whereas value change may be more stable. There are no predictions regarding cost. There is no obvious disconfirming evidence for trajectory variances and covariances because so little research exists.

Use 4 hypotheses. One of the major concerns about the experimental analyses is that the measure's validity cannot be assessed if the intervention does not work. Thus, we include two additional measures (a manipulation check and an outcome measure) to assess the intervention's effects. Disconfirming evidence would be a substantial departure in conclusions when using the single-item measures compared to the composite measure. Similar disconfirming evidence would be if the measures departed drastically from the manipulation check and outcome measure. Because of the timing of each measure, it is important to consider the results of the Use 4 analyses as a whole.

Results

This section is organized by the intended uses and underlying assumptions as outlined in Table 2. For each use, we first examine the validity evidence for the expert scale in the absence of information about the composite scale or random scale if possible. Doing so allows us to make validity arguments that rely on the analyses and expert knowledge rather than on comparison to a longer scale that may not exist in other circumstances. After presenting the results for the expert scales, we supplement our validity argument by comparing the findings with the composite and the random scale. This process helps to identify areas where these analyses may be insufficient to judge validity using only single-item measures.

Use 1: Measuring Important Motivation Phenomena in Classrooms

The first several assumptions tested for Use 1 include demonstrating reliability (Table 2: Use 1, Assumption 4), demonstrating that measures are correlated as expected (Table 2: Use 1, Assumption 5), and demonstrating that measures correlate with outcomes as expected (Table 2: Use 1, Assumption 6).

Reliability and intercorrelations. Table 6 and Table 7 contain descriptive statistics and correlations for the three versions of the motivation scales. Inter-construct correlations between single-item scales varied in magnitude but tended to preserve the direction of the relationship. Standard deviations tended to get larger for the single-item measures. The expert expectancy scale was correlated with the value composite, Sample 1: r = .40, p < .05, Sample 2, r = .46, p < .05, with the cost composite, Sample 1: r = -.26, p < .05, Sample 2, r = - .25, p < .05, and with interest, Sample 1: r = .43, p < .05, Sample 2, r = .46, p < .05. The expert value scale was correlated with the expectancy composite, Sample 1: r = 50, p < .05, Sample 2, r = .42, p < .05, was weakly or uncorrelated with the cost composite, Sample 1: *r* = -.06, p > -.05, Sample 2, *r* = -.10, p < .05 , and was correlated with interest, Sample 1: *r* = .64, p < .05, Sample 2, *r* = .64, p < .05. The expert cost scale was correlated with the expectancy composite, Sample 1: *r* = -.11, p < .05, Sample 2, *r* = .16, p < .05, was uncorrelated with the value composite, Sample 1: *r* = .01, p < .05, Sample 2, *r* = -.03, p < .05, and was uncorrelated with interest, Sample 1: *r* = -.01, p > .05, Sample 2, *r* = -.01, p > .05. Because the correlations are in expected directions and mostly of expected magnitudes, these correlations can be used as reliability evidence.

Generally, correlations yielded by expert-selected items were quite similar to the values yielded by composite scores. Differences between the three alternative scale compositions would not lead to substantively different conclusions with one exception. The cost composite was more strongly related to expectancy in Sample 1, *r* = .28, p < .05, and Sample 2, *r* = .29, p < .05 than the expert-selected cost item (C4), *r* = -0.11, *p* > .05, *r* = -0.16, *p* > .05, respectively, or the randomly selected cost item (C3), *r* = -.02, *p* > .05, *r* = -.02, *p* > .05, respectively. Overall, it appeared that correlations with single-item cost measures tended to depart most from the composite correlations; the departures were larger for the random item than the expert item.

Finally we examined single-item correlations with their respective composite measures as a secondary check of reliability. As would be expected, the correlation between the single-items and composites were relatively high, suggesting that the reduced measures had at least some reliability. Differences were relatively small, but correlations between expert-selected items and other measures were almost always better approximations of the composite correlations than the random-selected items.

Test-retest reliability. Test-retest reliability was calculated in two ways. The first was between adjacent time points, representing the shortest time intervals between measures. The second was from the Time 1 survey to the Time 4 survey, representing the largest time interval between measures. We calculated these two versions because adjacent test-retest reliability could not be calculated for the random scale (as noted before, the random items were not collected during Time 2 and Time 3). The two methods allow a deeper exploration of reliability between expert and composite scale, and shallower exploration of all three scales. Table 8 contains both versions of test-retest as well as means of the adjacent correlations by construct (e.g., average correlation for expectancy). Generally, expectancies had the lowest test-retest reliability followed by cost; Value demonstrated the highest mean test-retest correlations. Interestingly, expert scale reliabilities were rarely below the composite reliabilities and were often much higher. For example, the largest decrease from composite to expert scale test-retest was between Time 2 and Time 3 where the composite was $r = .51$ and the expert was $r = .46$. For value and cost, the expert scales almost always demonstrated higher reliability than the composite scales.

The pre-post reliability showed a similar pattern; the expert scale outperformed the composite and random scales in every case (in several cases only by a negligible amount). Interestingly, the random item also tended to show higher test-retest reliability than the composite measure. As with the adjacent reliability, the pre-post reliability was highest for expert scales of value and cost.

Prediction analyses. The final assumption tested for Use 1 was demonstrating that a reduced number of items can represent the underlying construct (Use 1, Assumption 7). The previous analyses show some evidence that the scales retain their construct representation even when shortened to a single-item. This set of analyses provides a quantifiable cost-benefit analysis for scale reduction with the cost measured as loss of total variance explained compared with the benefit of number of items removed from the scale.

Table 9 contains the set of unstandardized coefficients, R-squared values, and standard error estimates. One path model was computed for each scale version (composite, expert, and random). Each model included three predictors (expectancy, value, and cost) and three outcomes (Time 1 Interest, Time 4 Interest, and Pass Rate). Several important factors characterize the results predicting interest. First, more variance was explained in Sample 2 interest ($\overline{R^2} = .38$) than Sample 1 interest($\overline{R^2} = .30$). This is likely due to the fact that the larger sample size increases statistical power (affecting standard errors), and the fact that more respondents increased the reliability of the measures (affecting point estimates). Second, the amount of variance explained in interest decreased from Time 1 to Time 2 (see Figure 1). It is typical for measures to become less correlated as the duration of time between them increases. Third, in all cases, the composite scale yielded the highest variance explained, followed by the expert scale, followed by the random scale (see Figures 1). The average reduction in percent of variance explained was 4% (range 1% to 7%) from composite to expert, 7% (range 2% to 10%) from expert to random, and 11% (range 7% to 15%) from composite to random. In other words, reducing the full 13-item scale to three items (this 10 item decrease in scale length amounts to a 77% reduction in the proportion of the full scale) corresponded with a decrease in total variance explained from .39 to .35 (this .04 decrease in variance explained amounts to a 10% reduction in the proportion of total variance explained)[7]. Thus, the benefit of reducing the scale was far larger than the cost paid in variance explained.

The results were generally similar in terms of variance explained in pass rates, but on a much smaller scale. First, more variance was explained in Sample 2 ($\overline{R^2} = .026$) than Sample 1($\overline{R^2} = .017$). The variance explained did not consistently decrease from composite to expert to random, but all differences were less than 0.5%. In fact, the primary differences in pass rate R^2 were between samples in terms of statistical significance; the r-squared value suggested statistical significance for all three versions of the scale for Sample 2 (the expert scale p-value was .052), but none were significant in Sample 1. In other words, the reduction in scale length had minimal impact on standard errors, whereas the difference in sample size changed the conclusions of the statistical tests. In terms of construct path coefficients, composite versus single-item measures did not appear to substantively change the results either. Constructs remained statistically significant or non-significant regardless of which measure was used.

Use 2: Examine Student Group Differences

Use 2 was supported by two major assumptions: that at least some groups should show mean-level differences of motivation, and that a reduced number of items would remain sensitive to these differences. This use builds off Use 1 because it also requires the scales to be reliable and to measure the theorized underlying construct. Table 5 contains means and standard deviations for composite, expert, and random scales, as well

[7] On average, the composite explained 39% of the variance in interest. Thus, a 4% decrease from composite to expert scales represents a reduction in the proportion of total variance explained of 10% (39/4 = 9.75).

as the inter-construct correlations for the expert and random scales. Differences between the single-item and composite scale means ranged from -0.17 to 0.25 points different.

Gender differences. In addition to examining descriptive statistics and correlations, primary analysis for this group concerned known-group differences for the expectancy sub-scale. Specifically we assessed gender differences. Sample 1 showed that women had lower expectancies on the composite scale (d = .33, p < .05), the expert scale (d = .37, p < .05), and the random scale (d = .24, p < .05). Similarly, Sample 2 showed that women had lower expectancies on the composite scale (d = .33, p < .05), the expert scale (d = .36, p < .05), and the random scale (d = .17, p < .05).

We also include gender difference analyses for value and cost for the sake of completeness. Sample 1 showed that women and men were similar on the cost composite scale (d = .00, p > .05), the expert scale (d = .03, p > .05), and the random scale (d = .02, p > .05). Similar results manifested on the value composite scale (d = .01, p > .05), the expert scale (d = .00, p > .05), and the random scale (d = .01, p > .05). Sample 2 showed that women and men were similar on the cost composite scale (d = .02, p > .05), the expert scale (d = .02, p > .05), and the random scale (d = .02, p > .05). Similar results manifested on the value composite scale (d = .00, p > .05) and the random scale (d = .01, p > .05), though the expert scale actually showed women with lower value than men (d = .13, p < .05).

Use 3: Monitor Student Motivation Change

Use 3 builds off of both of the prior uses and their underlying assumptions which tested whether the scales under study represent the underlying constructs as expected and whether they can be shortened or not. Particularly pertinent to Use 3 are the test-retest reliabilities which showed varying degrees of stability for motivation change over time. The group difference analyses also show that at least the expectancy subscale is sensitive to differences. Use 3 has three assumptions: that student motivation changes during the semester, that the change can be captured by the scales, and that reduced scales can show similar results.

Modeling and model specification[8]. Latent growth curve modeling functions by estimating a latent, random-coefficients regression equation to describe construct change in a sample. The model estimates an intercept, a slope, as well as variance estimates for each parameter. The variance indicates whether or not individuals are uniform in their latent trajectories, or if they vary. Data with more than three time points can also produce quadratic terms which represent change in the slope in each time interval (acceleration or deceleration). When growth parameters have variance, they can be correlated with other growth parameters or variables. An extension of the latent growth-curve model is the dual-process model, which allows a user to estimate growth parameters for multiple constructs simultaneously. These dual process models further allow for covariance between growth parameters in different constructs (e.g., is change in Construct A related to change in Construct B). In the current study, we estimated simultaneous growth models that included expectancy, value, and cost (we also included experimental condition as a covariate based on the results of the experimental analyses).

[8] Although the individual items are ordinal and not continuous scales, they were treated as continuous. Some research suggests that with 5 or more response categories, ordinal data can be treated as continuous without substantial problems. Part of the reason for this decision was for model comparability. Composite scores included too many values to be treated as ordinal indicators, whereas single-item measures could be treated as such. Additional analyses suggested that although accounting for the ordinal nature of the indicators resulted in near perfect fit for several of the models, there was no substantive difference between treating items as categorical and continuous for the purposes of model selection.

Because of the sheer volume of models it would require to arrive at the most appropriate model for each scale version (composite, expert, alternate) independently, we instead used the composite scale in Sample 1 to determine the most appropriate specifications for each individual construct before computing the full three-construct model. We then re-calculated each model using the expert and alternate versions of the scale and compared fit. The original plan was to estimate all of the models in Sample 1, and then to cross-validate those models in Sample 2. Because the full three-construct growth model was ultimately unable to converge in Sample 1, we instead cross-validated the individual construct models and estimated the three-construct model in Sample 2 only. The individual construct models generally replicated in Sample 2 except for slope variances; the expectancy and value slopes did not significantly vary in Sample 1, but did in Sample 2. As with previous analyses, this is likely a matter of statistical power—inspecting raw individual means suggested variability among students. Based on our initial model building phase, we arrived at a linear growth model for expectancy (intercept and slope), and a non-linear growth model for value and cost (intercept, slope, and quadratic). Although including the quadratic terms led to improved model fit, it was not clear that the parameters were necessary; the correlations between slope and quadratic terms were near 1.0 for value and cost. Because these data were collected from an intervention study, we included a dummy code for experimental condition as a covariate for all items.

Model fit. To determine whether or not models adequately represented the data, we examined several fit indices including χ^2, RMSEA, CFI, TLI, SRMR, AIC, BIC, and ABIC (see Table 10 note for recommended fit values). Generally speaking, models using composite, expert, and alternate scales did not differ in whether or a given model would be considered acceptable in absolute terms (i.e., "is this model an accurate representation of the data). Interestingly, it was not even necessarily clear that the composite-scale model always fit better than the single-item-scale models. For example, the composite value-only model in Sample 2 demonstrated lower fit, χ^2 (2) = 38.75, *p* < .05; RMSEA = .11; CFI = .98*;* TLI = 1.07; SRMR = .05, than the expert value-only model, χ^2 (2) = 2.55, *p* > .05; RMSEA = .01; CFI = 1.00*;* TLI = 1.00; SRMR = .03. However, these five fit indices are not necessarily meant for cross-model comparison and thus are not as informative as indices developed explicitly for that purpose (AIC, BIC, ABIC). In virtually every case, the composite scale model showed better AIC, BIC, and ABIC values (i.e., lower values) than the single-item scale models. This suggests that while the global fit indices for making independent model decisions support the use of any of the three scales, the composite versions generally fit better in comparison to the single-item versions.

Parameter comparison. Model fit inspection suggested that single-item measures could be used in place of composite measures without changing model-selection. However, such conclusions do not necessarily mean that the results of the models were identical. We next examined the substantive conclusions that would be drawn based on each version of the scales in the full three-construct growth model in Sample 2. Table 11 contains the intercepts, slopes, quadratic terms, and associated variances estimated in the model. Of the 16 growth parameters estimated in each model (we consider covariances later), there were five instances in which the expert and alternate models did not match the composite (i.e., all parameters were statistically significant except in the following five cases). First, the expert-scale model suggested that the value slope did not have statistically significant variance. Second, the expert-scale model suggested that the quadratic term for value was not statistically significant. Third, the alternate-scale model suggested that the expectancy slope was not statistically significant. Fourth, the alternate-scale model suggested the cost slope was not statistically significant. Fifth, the alternate-scale model suggested the cost quadratic term was not statistically significant. It is important to consider that in each instance where the slope or quadratic terms were deemed non-significant, the associated variance was considered statistically significant, which would suggest retaining an estimate of the slope in the model. The only case in which a single-item measure would lead to selection of a different model was the non-significant value slope variance for the expert measure.

Table 12 includes the correlations between growth parameters in each model. Whereas the growth parameters were relatively stable across scale versions, the covariances between the growth parameters yielded some stark contrasts. For example, the correlation between expectancy intercept (EI) and value intercept (VI) was .51 using the composite scales and .54 using the expert scales, but it was .28 using the alternate scales. Among the expectancy and value covariances, the composite and expert scales performed similarly. Among the covariances that included cost, there was a greater departure from the composite scale. For example, the correlation between expectancy intercepts and cost slopes was -.24 and statistically significant when the composite scale was used, but was near zero and non-significant for both single-item measures.

Use 4: Monitoring Experimental Processes

Use 4 builds off of the other three uses and their underlying assumptions that support the use of the scales to measure motivation. The current study addresses two assumptions for monitoring experimental processes: that during the course of an intervention, relevant scales should demonstrate differences between experimental groups (Table 2, "Use 4, Assumption 2), and that items most-aligned with the intervention should yield the largest differences (Table 2, "Use 4, Assumption 3). The final set of analyses estimated the effects of an experimental manipulation on different formulations of the value scale. Prior to considering these analyses, it is important to point out that experimental manipulations may not be present for a number of reasons beyond poor measurement quality. One possibility is that the intervention does not work at all. Another possibility is that in a particular occasion, the intervention was not implemented well enough to produce effects. Thus, in order to determine whether or not a scale is sensitive to experimental manipulations, it is necessary to include other indicators that signal that the intervention did or did not produce the intended effects. To address this concern, we first examined three measures to assess whether or not the intervention produced effects. The first two measures were manipulation checks of the degree to which students found the material relevant to their lives. These measures were created specifically to assess effects of the value intervention. We conducted the manipulation checks immediately after the intervention during Time 2 and during Time 3 (one to seven days after Time 2). We then examined the effects of the intervention on pass rates, which were the intended outcome of interest for the intervention.

Model specification. Path models were calculated for each version of the value scale (composite, expert, alternate), for each sample (Sample 1 and Sample 2), for each dependent variable (manipulation check, pass rate, value scales). All twelve models were identical in that they contained four predictors: baseline composite measures of expectancy, value, and cost, and a dummy code for experimental condition (0 = control condition, 1 = value condition). All models included one dependent variable, either one version of the value scale, the manipulation check, or pass rates.

Manipulation check and outcomes. Overall, the intervention did not appear to yield the expected impacts (see Table 13). In Sample 1, there was a positive but non-significant effect (b = .159) of the intervention on the manipulation check measure immediately following the intervention. The manipulation became slightly negative (b = -.035) during the next time point. In Sample 2, there was a positive and significant effect (b = .246) of the intervention on the manipulation check measure immediately following the intervention. The manipulation became smaller and non-significant (b = .113) during the next time point. Finally, Sample 1 (b = .106) and Sample 2 (b = .065) showed positive but non-significant effects of the intervention on pass, suggesting no overall effect. Given these results, it is difficult to determine the results of the value measure. Thus, we could only assess whether intervention effects were in the same direction and of relatively similar magnitude.

Value. The post-intervention value measures were collected during the time point after the intervention was delivered (Time 3). For Sample 1, results differed from those found with the relevance measure at Time 3; all three versions of the value scale suggested a negative effect in Sample 1. Unlike the relevance measure, the composite and expert value scales suggested the negative effect was statistically significant or nearly so. For Sample 2, results were also similar to those found with the relevance measure at

Time 3; all three versions of the value scale suggested a non-significant near-zero effect. Overall, it appears that the intervention did not work. However, the general direction and pattern of results was consistent across the manipulation check, the value measures, and the pass-status indicator.

Discussion

The foundation for conducting pragmatic measurement is providing validity evidence for the specific interpretation and uses of the scale. In the current study we identified four common uses of expectancy, value, and cost scales. We then identified several assumptions underlying each of the uses, which we aligned with various sources of validity evidence. We used theoretical knowledge to determine whether or not we could draw useful conclusions using only single-item scales. We were also able to check the accuracy of our conclusions by comparing conclusions from single-item scales to conclusions from composite scales. In the following sections, we briefly discuss the presence or absence of validity evidence for each of the proposed uses. We also note that no single analysis is sufficient evidence for validation. Instead, we advocate for considering the results in aggregate.

Use 1: Measuring Important Motivation Phenomenon

The results tended to support the use of the expert version of the expectancy, value, and cost scale for measuring student motivation. Correlations between constructs were generally of the expected direction and magnitude and most departures from those predictions were minor. There was one major exception to these supportive findings: the correlation between value and cost was near zero. Prior research (Eccles et al., 1983; Gaspard et al., 2015; Perez et al., 2014) and theory (Eccles et al., 1983) suggests that value and cost should be negatively related. On its own, this finding is potentially concerning because it would suggest that the single-item scale is not reproducing prior findings. However, the differences between single-item and composite scales suggest that this anomaly is not a product of scale length reduction. Item-total correlations suggested that the cost scale had somewhat lower reliability than the other two scales, but not to an alarming degree. Furthermore, the test-retest reliabilities for all three constructs implied similar patterns as previous longitudinal studies (Kosovich et al., 2017; Perez et al., 2014). Two more likely hypotheses are that the low correlation represents variation due to the inclusion of a different sample, or that the cost scale is more generally invalid. Given that this is the first time this cost scale has been used to assess community college students, sample differences (from prior studies) seem to be the most likely explanation. In other words, the relationships among constructs may be different for students with these characteristics than those in previous studies—the result being different relationships among the constructs under study. The additional analyses provided additional validity evidence in favor of scale use.

The prediction analyses were consistent with the reliability check and provided further evidence that the expert scales could act as proxies for the full scale. As expected, there was a decrease from the composite to the expert scale in terms of variance explained in outcomes. However, the removal of 10 scale items only resulted in a decrease in variance-explained of 4%. In these cases it is up to the researchers to determine what an acceptable benefit-cost ratio would be. We judged the expert scale as showing acceptable losses. The longer-term outcomes (end of semester continuing interest and pass status) were actually more promising than the concurrent measures because all three versions of the scales were identical. Interestingly, it appeared that reducing sample size (by 60%) or reducing scale length (by 77%) resulted in similar variance reduction. Thus, deciding on the appropriate balance between psychometric quality and pragmatism, it is important to consider the interaction between scale length, if outcomes are short-term or long-term, sample size, and scale composition (e.g., which item/s are selected).

Use 2: Examining Student Group Differences

Whereas there is established evidence and theory behind gender differences in expectancy (Wang & Degol, 2013; Wang, Eccles, & Kenny, 2013), value differences are inconsistent across studies and cost simply does not have much research to draw from. Thus, the only validity tests of expected-group-differences pertain specifically to expectancy. In short, there was evidence for the use of the expectancy scale at least for detecting gender differences. As is the case with prior research, the current study showed that women reported lower expectancy than men despite performing better than men. The analyses related to experimental differences provided a similar test of group differences, but was inconclusive because of a lack of intervention effects.

Use 3: Monitoring Motivation over Time

The third use showed a complex pattern of results that ultimately supported the use of the expert scale with some important caveats. The expert model partially replicated the results of prior research on short-term expectancy and value trajectories (Kosovich et al., 2017), showing that both constructs started out higher and declined over the course of the semester. Interestingly, the expert scale indicated no variance in value slopes, which precluded an estimate of the covariance between expectancy and value slopes. Although Kosovich and colleagues (2017) found a strong correlation between the two trajectories, their models also showed that value slope variance was not statistically significant. In contrast, the model using the alternate scale suggested that there was variance in the value slope. The composite scale model showed similar results to the alternate scale. Although this may seem troubling, it is likely another case of multi-dimensionality. Specifically, Kosovich and colleagues measured one dimension of value which corresponds to the expert-selected item. The other additional departure was the presence of a statistically significant quadratic term indicating a curved trajectory. The prior research could not have estimated a quadratic term because they only used three time points of data (quadratic requires at least four). However, the quadratic term may be an artifact of the data collection design because the time interval between Time 3 and Time 4 was several times longer than any other time interval.

For expectancy and value growth parameters, the expert item was able to sufficiently replicate the composite results in terms of direction, magnitude, and statistical significance. In contrast, covariances with cost's growth parameters differed between the three scale versions hinting at weaker validity. Again, we suggest that multi-dimensionality may be one of the most important factors to consider for developing single-item measures. Differences in sub-constructs can lead to different conclusions, depending on which item is used. Importantly, these differences are not signaled by model fit indices which looked generally good for all three versions of the scales.

Use 4: Monitoring Intervention Processes

Despite the inconclusive results of the experimental portion of the current research, the measures appeared to be sensing real effects. These can only be seen if the nuance of the study design is understood. Ignoring the pragmatic measures being tested in the study, the manipulation checks showed a decline in value in both samples. In one sample, the manipulation check of value was slightly positive but no different than zero two weeks after the first intervention. In the other sample, the manipulation check of value was slightly negative, but no different from zero. The pragmatic measures of value matched the general trend in both samples but showed a steeper decline in both cases. The pragmatic measure was collected between the two manipulation checks, and an additional exposure to the intervention occurred immediately before the second manipulation check. Because of the timing of these measures, it is arguable that the sharp decline indicated by the pragmatic measure was accurate. The important consideration is that the value scale led to a similar conclusion as the manipulation check in each sample. However, further work is necessary to determine if the measures really are sensitive to experimental manipulation.

The major caveat to testing Use 4 is the study's statistical power. The level of randomization occurred at the classroom level, meaning that the likelihood of finding experimental differences (if they existed) was restricted. As with some of the previous limitations, power concerns are not strictly limited to pragmatic measurement. However, the current study demonstrated repeatedly that sample size seemed to have a similar impact on results as scale reduction. Thus, it is important to keep many of the standard concerns about research conduct in mind, because they may become particularly relevant in conjunction with pragmatic approaches.

Limitations

There are a number of important limitations to the current research that both restrict which conclusions can be drawn and point to necessary work moving forward.

The first limitation deals with the motivational measures specifically. Both the value and cost scales (particularly the latter) seemed to perform differently than expected. It is entirely possible that value and cost are qualitatively different among the students in this sample and relative to prior research contexts. There are a number of reasons why the findings may have been more discrepant from prior research than expected. The sheer age range (17-80) indicates an wide breadth of life experience. The racial and ethnic diversity further complicate the matter as different cultural backgrounds may favor different goals and behaviors, or may interpret the constructs in entirely different ways. As mentioned earlier in the paper, more fine-grained distinctions between facets of the constructs under study need to drive measurement and the current value and cost measures are not developed for that purpose. The result may be that student heterogeneity (i.e., variability in characteristics such as age, gender, culture) yielded different interpretations of the scale items. In other words, the sample includes a different population, or draws from several different populations, than prior research. For example, does *important* have the same connotation to teenagers and middle-aged students? Does life experience increase individual capacities for balancing various priorities and reduce cost for some? Do different ethnic backgrounds lead to different opinions about whether or not a class is useful? Additional qualitative research could help to illuminate different interpretations of these different scales. More in-depth quantitative analyses (e.g., measurement invariance) could unravel the possibility of sample heterogeneity. This also leads to a more general limitation regarding measure type.

The second limitation is related to the near-exclusive use of self-report data in the current research. Although the patterns of observed correlations conform to theory, most of the results are ultimately correlations among self-report. There are many options that can be used to improve validity evidence of self-report measures by incorporating other modes of measurement (e.g., observational or behavioral). For example, students who are more motivated to succeed in a class are likely to participate more or and have fewer absences. Including measures of class participation, homework submission rates, attendance, or future course selection would be useful metrics for determining the true utility of these pragmatic motivation measures—particularly if there are differential results from different versions of the measure. Observational measures focused on student engagement may also provide useful information in showing that variation in student behavior is visible by others. Such correlations are particularly important validity evidence for practitioners. If self-report and observational measures do not converge to some degree, it may difficult for practitioners to see results. In the current research, the only measure used in this manner was students' likelihood of passing. The lack of corroboration between different types of measures may also lead to concerns about bias.

The third limitation deals with pragmatic measurement's inability to identify item bias. It is well documented that individual items may show differential functioning for one group over another. It is also an inherent part of measurement theory that participant responses (observed scores) do not perfectly correspond with individuals true levels of a construct (true scores). Although there are methods for identifying bias and estimating individuals' true scores, such methods are not available for single-item measures and may not be ideal for pragmatic measurement more generally. Until we identify alternative methods for identifying bias, utilizing more advanced methods during scale development when possible will bolster the measure's validity evidence. Related to bias, an insufficient number of items precludes many other advanced techniques used to study measure qualities.

The fourth limitation concerns measurement error. Despite the promise of scale reduction offered by the current and past research, it is not possible to partition measurement error of single-item measures. All of the analyses conducted in the current study used statistical models that assumed no measurement error in the items. This is a problematic practice because observed scores are assumed to be a combination of the underlying construct and various sources of error. Statistical models such as multiple regressions assume that items are free of error. Thus, using raw items likely violates such assumptions. We make no claims that pragmatic measurement can solve this problem, though we do argue that high quality items that have been vetted by the validation process may be able to function adequately. This is perhaps the most difficult aspect of measurement for the pragmatic approach to overcome. Although much of this paper has focused on single-item measures, pragmatic measures can include multiple items—thus this concern may be able to be addressed by using a few items rather than one item. Ultimately, it is up to measure users to estimate how much an impact error may have on conclusions, and to decide the appropriate compromise between concerns about error and other pragmatic concerns.

The final limitation of these findings is that it appears that advanced statistical models begin to suffer from compromises made to conduct practical measurement. Although all three versions of the scales yielded similar conclusions about how the constructs change over time, the relationships among the growth parameters were variable. In particular, it was unclear if versions of the cost scale could be meaningfully integrated into a simultaneous growth model because of substantial departures of the single-item measures from the composite. The results provided additional evidence that multi-dimensionality is perhaps one of the most important considerations when adapting a pragmatic measure, even for constructs like value where the dimensions are normally highly correlated. The results of this study point to the need for strong theoretical backing and a research base from which to draw empirical predictions, even down to the narrowest characteristics of a particular construct.

General Implications

Perhaps the most important consideration is whether or not the current findings are applicable to other samples, constructs, and modes of measurement. As with much scholarly work, the answer depends on many factors. First and foremost, the findings of this study in no way validate other motivation measures in different samples or domains. Second, current research does not unequivocally support the use of single item measures for all constructs (self-reported or otherwise). Single item measures unaccompanied by validity evidence should be particularly suspect for many of the reasons discussed throughout the current manuscript (e.g., likely lower reliability and construct breadth). That being said, there are important lessons that can be gleaned from the current research that generalize beyond the current sample, the current measures, and the current domain. Based on the work and limitations discussed above, we summarize the implications of this work with four major points, discussed in greater detail below. We also discuss potential future directions of work on pragmatic measurement.

- Pragmatic measurement can provide useful validity information.
- The key to pragmatic measurement is a compromise between technical and practical requirements.
- Different phases of research offer different validation opportunities.
- Pragmatic measurement can be complementary to other measurement approaches.

Pragmatic Measurement Can Provide Useful Validity Information

Anecdotally, there are many situations for which pragmatic measures are well suited. A researcher working in the field may need a measure and only have a week to find one, an administrator or instructor may need a measure but lack the knowledge or resources to conduct validation, or a researcher may need to collect supplemental data with limited item space. Many measurement opportunities arise suddenly and simply cannot wait for the lengthy and slow process of developing a measure using standard methods. This is not an excuse for lowing measurement standards; rather it is a fact of conducting research in natural settings. Pragmatic measurement as we describe it in this study is a tool for low-stakes contexts. We developed the pragmatic measurement perspective out of recognition that better validity evidence is needed across the social sciences. It is in these cases where pragmatic measurement can provide an evidence-based approach to conduct measurement in the face of contextual constraints.

It is important to remember that there is a long history of measurement practices in educational research. Prior to modern computing technology and modern statistics, we were still able to conduct validation. Although the philosophy of measurement has evolved over the years, the methods are still applicable. In pushing pragmatic measurement forward, it may prove useful to investigate older approaches to validation that may still be useful (Rudner, Getson, & Knight, 1980).

The Key to Pragmatic Measurement is Compromise

Pragmatic measurement requires compromise between the technical standards of measurement and practical constraints of research. Depending on the overall purpose of ones' research, different balances may be acceptable. Perhaps the most important considerations are the consequences of making an incorrect conclusion using the scale. Data for use in high stakes decisions-making processes, such as determining teacher pay or student advancement, require extremely precise measures (Duckworth & Yeager, 2015). Generally speaking, it is probably best to follow the most technically rigorous and costly validations methods possible for mission-critical measures.

When consequences are less extreme, there is more room for comprome between technical and practical concerns. For example, based on the current research, practitioners would be justified in using the motivation scales for low-stakes purposes without much reservation. The major concerns with the measures pertain to much more complex research questions than most practitioners would need. Even though the cost scale was potentially problematic, it generally functioned as would be expected. Researchers who might use pragmatic measures as covariates would have a substantial amount of evidence suggesting that the measures could be used.

Another way to frame this consideration is to ask whether a scale's values are inherently meaningful or arbitrary? For example, a one million dollar increase in pay has a real, tangible meaning. A half-point increase in student's value is meaningless without providing relevant context (e.g., comparison to another group, time point, or person).

Pragmatic measurement as we currently conceptualize it is simply not equipped to address inherently meaningful response values without latent variable modeling. The values generated by statistical models typically vary depending on the specific items or composites chosen. Because of this, pragmatic measurement is helpful in drawing general conclusions about the direction or general strength of a relationship or difference. We advocate for the use of pragmatic measurement for informal assessments (e.g., “Are students in my class relatively high or low in expectancy?”), or as supplementary or exploratory measures (e.g., “I want to control for expectancy in my statistical model.”). Whether it concerns a short measure or a long measure, a measure of motivation or of some other psychological construct, there is simply not enough evidence at any level that self-reports measures can be used for high stakes purposes (Duckworth & Yeager, 2015), nor that such measures could be used to label an individual. The overall purpose for which measures are being used may dictate how best to validate those measures. Ultimately, it is up to the measure users to determine the best balance of breadth vs. depth, and of precision vs. adequacy. Moving forward, it will be essential to the development of pragmatic measurement to test the method with other constructs and other modes of measurement to determine when compromise is possible and not.

To maintain the technical quality of a measure when reducing items, users need to relinquish other benefits—chief among them, construct breadth. This highlights the need for specificity. With a limited number of available items, measure users need to carefully consider which aspects of a construct should be most related to the outcomes of interest. In the current study, different items purported to measure the same motivational construct showed different relationship patterns with other constructs (i.e., led to different conclusions). This specificity is one of the key compromises in developing pragmatic measures. Using the value items from the current research as an example, the item capturing to *utility* (i.e., the perception that a task is useful to the future) was more strongly correlated with expectancy than the item capturing *general value*. Given that utility has been cited as a unique aspect of value that is more directly correlated with achievement (Wigfield & Cambria, 2010), the differential relationships make sense theoretically. Such differential relationships should also be expected outside of motivation research, too. For example, the Classroom Assessment Support System measure (CLASS; Hamre, Hatfield, Pianta, & Jamil, 2014) was developed to measure teacher-student interactions in the classroom and comprises three factors (Emotional Support, Classroom Organization, and Instructional Support). Each of the factors is made up of several dimensions which are in-turn made up of several behavioral indicators. Researchers interested in a pragmatic version of the CLASS to test the relationship between emotional support and students perceptions that their instructor respects them may need to look at one of the dimensions (e.g., regard for student perspectives) or even one of its behavioral indicators. Measures like the CLASS include behavioral indicators that vary drastically and may not be as correlated with each other as the items within expectancy-value-cost constructs. This fact means that combining many items from multiple dimensions into a single measure would be more likely to attenuate relationships between particular aspects of the constructs and other outcomes. The tension here is shortening a measure narrows its scope of usability; an item or indicator that is predictive of outcomes may not be most representative of the present level of the overarching construct (i.e., only one aspect of the larger construct relates to a particular outcome).

Moving forward, it will be important to determine if limiting measure lengths impede the larger program of research.

Different Phases of Research Offer Different Validation Opportunities

Validation is an ongoing process that occurs in every stage of research from research design through secondary data analyses. In recognition of this point, it is worth considering where measure users may have opportunities to bolster their validity evidence. For this purpose, we suggest two unique phases of validation—the study development phase and the study conduct phase. The development phase encompasses initial decisions about the study and possibly pilot work. In this phase, some of the more sophisticated statistical techniques may be viable before the full-scale study is launched and the final pragmatic measures are needed. In the development phase, it is likely that measure users will have at least a few items for each construct—even if they do not have the expansive item pools that some measurement experts suggest. The study conduct phase encompasses the actual study time after measures have already been selected and the research design is in place. In the conduct phase, users will need to make do with the existing measures which may have already been deployed. The current research falls into the second phase of validation.

The first phase is where additional validation opportunities may present themselves. For example, the recommended approach to assessing group differences requires users to establish *measurement invariance* (Cheung & Rensvold, 2002; Wu, Li, & Zumbo, 2007). Measurement invariance is a practice of latent variable modeling that tests for bias in scale items. Researchers sequentially test whether different characteristics of items (e.g., the strength with which the item represents the construct) are equal across groups. Doing so allows researchers to ensure that differences are due to actual variations in the construct rather than unrelated factors. Testing such concerns with single-item measures is not possible once measures have been selected, but may be possible during the study development phase.

Even with relatively short scales, users may have several items per construct during the development phase of a project. Pools of at least three items offer the opportunity to conduct many of the psychometric analyses (e.g., invariance testing) that cannot be done effectively with one- or two-item scales. Thus, even when users have relatively urgent timelines for launching a study or an assessment, it is worth the effort to test some of the items to assess their performance. If it is not possible to get a sufficiently large sample to conduct statistical analyses, users may also consider collecting qualitative data from students.

Pragmatic Measurement Can Be Complementary

As we have previously suggested (Kosovich et al., 2015), measurement approaches are not mutually exclusive and it may be beneficial to conduct various validation inquiries. Many measures are often collected to supplement primary research questions, but the number of supplemental measures is often confined by space restrictions. When being included for covariates or as exploratory measures, pragmatic measures may open new horizons by making it possible to include more factors than usual. For example, it might be possible to create a practical measure of education covariates that captures a large amount of variance in typical educational outcomes. Pragmatic and more in-depth measurement approaches can and should be used together to ensure validity for all of the measures in a study. A benefit of pragmatic measurement is that users can distribute resources needed for validity to spend more time on the primary measures without abandoning validity evidence for the secondary measures. In far too many cases it is nearly impossible to find any validity evidence at all, let alone a compilation of validity evidence. In further developing the pragmatic measurement perspective, we hope to increase both the general frequency of validation practice, as well as the reporting of validity evidence.

Context Drives a Need for Pragmatic Measures

One criticism that may have occurred to readers throughout this paper was whether or not the reduction from four items per construct to one item per construct is really necessary. The answer, based on our particular measurement obstacles is unequivocally, “Yes.” Student behavior monitored during the course of these studies suggests sensitivity to research activities and practices. Whether through discussions with faculty and administrators, or through students’ qualitative data, or non-response patterns, evidence suggests that the need for minimal intrusion is particularly salient.

For example, the research design was altered mid-semester during pilot testing because of the volume of student complaints regarding survey frequency. These reports were largely anecdotal and were not collected in a systematic way. However, they were corroborated a) by a large number of non-responses to student essays, b) by a number of student essays that ignored the writing prompt to lecture the researchers about wasting students’ time with repetitive questioning, and c) by a consistent and rapid decline in response rates after several weeks into the semester. This constellation of feedback suggests that pragmatic measures are not only important, but is indispensable in some cases.

Although severe sensitivity to scale length may seem unique to the current research and the population in the current study, it highlights perhaps the most important consideration for pragmatic measurement. Users of all backgrounds (e.g., researchers, practitioners, administrators, community organizers) need to consider the context under which they intend to conduct measurement. Many situations will share similar characteristics, and also present unique challenges. Whether adopting the pragmatic measurement perspective or more commonly used measurement perspective, it is critical to consider what setting characteristics could undermine the quality of the data collected.

Conclusion

At the end of the day, we endorse the pragmatic measurement approach as a method for validating and improving measurement practices. The decision to use this pragmatic measurement approach over lengthier and more technically demanding approaches is not a simple decision. When the overall purpose of a measure is to provide high-stakes feedback about an individual or a policy, measure users need to budget significant resources to the validation process. Pragmatic measurement is not a seamless replacement for typical best practice recommendations. Instead, when setting constraints have a strong potential to undermine data collection efforts, it is crucial to weigh the costs and benefits of different measurement approaches. Just as taking a person's pulse may not be able to diagnose a rare heart disease, a pragmatic measure may not detail why a student is unmotivated. However, just as someone's pulse can signal a doctor that more tests may be needed, a practical measure of motivation can signal where researchers and practitioners may want to ask more questions.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *The American Psychologist*, *63*(1), 32–50. http://doi.org/10.1037/0003-066X.63.1.32

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., … Sherman, S. J. (1990). Graduate training in statistics, methodology, and measurement in psychology: A survey of PhD programs in North America. *American Psychologist*, *45*(6), 721–734. http://doi.org/10.1037//0003-066X.45.6.721

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology*, *102*(4), 804–816. http://doi.org/10.1037/a0021075

Atkinson, J. W. (1964). *An introduction to motivation*. Princeton, NJ: D. Van Nostrand Company, Inc.

Barron, K. E., & Hulleman, C. S. (2015). Expectancy-Value-Cost Model of Motivation. In *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 8, pp. 503–509). Elsevier. http://doi.org/10.1016/B978-0-08-097086-8.26099-6

Bryk, A. S., Gomez, L., Grunow, A., & LeMahieu, P. (2015). *Learning to improve: How America's schools can get better at getting better*. Cambridge, MA: Harvard Education Publishing.

Chapelle, C. a., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13. http://doi.org/10.1111/j.1745-3992.2009.00165.x

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. http://doi.org/10.1207/S15328007SEM0902_5

Clay, R. (2005, September). Too few in quantitative psychology. *Monitor on Psychology*, 26.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* ERIC.

Cronbach, Lee, J., & Meehl, P. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*, 281–302.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 2–17). Hillsdale, NJ: Lawrence Erlbaum.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology* (pp. 35–54). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-017-9088-8_3

Csikszentmihalyi, M., & Schiefele, U. (1994). Interest and the Quality of Experience. *European Journal of Psychology of Education*, *10*(3), 251–270.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–232.

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*(1), 36–45.

Dimotakis, N., Ilies, R., & Judge, T. A. (2013). Experience sampling methodology. In J. M. Cortina & R. S. Landis (Eds.), *Modern Research Methods for the Study of Behavior in Organizations* (pp. 319–349). Routledge. http://doi.org/10.4324/9780203585146

Duckworth, A. L., & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, *44*(4), 237–251. http://doi.org/10.3102/0013189X15584327

Eccles, Jacquelynne S., Wigfield, A. (1995). In the Mind of the Actor: The Structure of Adolescents' Achievement Task Values and Expectancy-Related Beliefs. *Personality & Social Psychology Bulletin*, *21*(3), 215–225.

Eccles, J. S., Adler, T., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation* (pp. 75–146). San Francisco, CA: W. H. Freeman.

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology*, *41*, 232–244. http://doi.org/10.1016/j.cedpsych.2015.03.002

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations.

Fuchs, L. S., & Fuchs, D. (2004). Determining Adequate Yearly Progress from Kindergarten Through Grade 6 with Curriculum-Based Measurement. *Assessment for Effective Intervention*, *29*(4), 25–37.

http://doi.org/10.1177/073724770402900405

Gaspard, H., Dicke, A., Flunger, B., Schreier, B., Häfner, I., Trautwein, U., & Nagengast, B. (2015). More value through greater differentiation: Gender differences in value beliefs about math. *Journal of Educational Psychology*, *107*(3), 663–677. http://doi.org/10.1037/edu0000003

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., … Preckel, F. (2014). "My Questionnaire is Too Long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, *39*(3), 188–205. http://doi.org/10.1016/j.cedpsych.2014.04.002

Graham, S. (2015). Inaugural editorial for the Journal of Educational Psychology. *Journal of Educational Psychology*, *107*(1), 1–2. http://doi.org/10.1037/edu0000007

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for General and Domain-Specific Elements of Teacher–Child Interactions: Associations With Preschool Children's Development. *Child Development*, *85*(3), 1257–1274. http://doi.org/10.1111/cdev.12257

Hormuth, S. E. (1986). The sampling of experiences in situ. *Journal of Personality*, *54*(1), 262–293. http://doi.org/10.1111/j.1467-6494.1986.tb00395.x

Hulleman, C. S., Barron, K. E., Kosovich, J. J., & Lazowski, R. A. (2016). Expectancy-value models of achievement motivation in education. In A. A. Lipnevich, F. Preckel, & R. D. Robers (Eds.), *Psychosocial skills and school systems in the Twenty-First century: Theory, research, and applications*. (pp. 241–278). Springer. http://doi.org/10.1007/978-3-319-28606-8

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing

Interest and Performance with a Utility Value Intervention. *Journal of Educational Psychology*.

Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2016). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*. http://doi.org/10.1037/edu0000146

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, *136*(3), 422–49. http://doi.org/10.1037/a0018947

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: gender and domain differences across grades one through twelve. *Child Development*, *73*(2), 509–527. http://doi.org/10.1111/1467-8624.00421

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2013a). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. http://doi.org/10.1111/jedm.12000

Kane, M. T. (2013b). Validation as a Pragmatic, Scientific Activity. *Journal of Educational Measurement*, *50*(1), 115–122. http://doi.org/10.1111/jedm.12007

Kosovich, J. J., Flake, J. K., & Hulleman, C. S. (2017). Short-term Motivation Trajectories: A Parallel Process Model of Expectancy-Value. *Contemporary Educational Psychology*. http://doi.org/10.1016/j.cedpsych.2017.01.004

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2015). A Practical Measure of Student Motivation: Establishing Validity Evidence for the Expectancy-Value-Cost Scale in Middle School. *The Journal of Early Adolescence*, *35*(5–6), 790–816. http://doi.org/10.1177/0272431614556890

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of aspiration. In J. M. Hunt (Ed.), *Personality and the behavior disorders* (pp. 333–378). New York: Ronal Press. http://doi.org/10.1037/10319-006

Lewis, C. (2015). What Is Improvement Science? Do We Need It in Education? *Educational Researcher*, *44*(1), 54–61. http://doi.org/10.3102/0013189X15570388

Marsh, H. W. (1994). Sport Motivation Orientations: Beware of Jingle-Jangle Fallacies. *Journal of Sport and Exercise Psychology*, *16*(4), 365–380. http://doi.org/10.1123/jsep.16.4.365

Messick, S. (ETS). (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, *18*(2), 5–11.

Murphy, P. K., & Alexander, P. A. (2000). A Motivated Exploration of Motivation Terminology. *Contemporary Educational Psychology*, *25*(1), 3–53. http://doi.org/10.1006/ceps.1999.1019

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York: Guilford Press.

Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology*, *106*(1), 315–329. http://doi.org/10.1037/a0034027

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, *95*(4), 667–686. http://doi.org/10.1037/0022-0663.95.4.667

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased Item Detection Techniques. *Journal of Educational Statistics*, *5*(3), 213–233. http://doi.org/10.2307/1164965

Schmeiser, C. B., & Welch, C. J. (2006). Test Development. In *Educational Measurment* (pp. 307–353).

Shinn, M. R. (2013). Curriculum-Based Measurement. In B. J. Irby, G. Brown, R. Lara-Aledo, & S. Jackson (Eds.), *Handbook of Educational Theories* (pp. 783–791). Charlotte, NC: Information Age Publishing, Inc.

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, *42*(1), 70–83. http://doi.org/10.1037/0012-1649.42.1.70

Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, *55*, 167–194. http://doi.org/10.1111/j.1744-6570.2002.tb00108.x

Subar, A. F., Ziegler, R. G., Thompson, F. E., Johnson, C. C., Weissfeld, J. L., Reding, D., … Hayes, R. B. (2001). Is shorter always better? Relative importance of questionnaire length and cognitive ease on response rates and data quality for two dietary questionnaires. *American Journal of Epidemiology*, *153*(4), 404–409. http://doi.org/10.1093/aje/153.4.404

Vroom, V. H. (1964). *Work and motivation. Classic readings in organizational behavior*. New York: John Wiley & Sons, Inc.

Wang, M.T., & Degol, J. (2013). Motivational pathways to STEM career choices: Using expectancy-value perspective to understand individual and gender differences in STEM fields. *Developmental Review*, *33*(4), 304–340.

Wang, M.-T., Eccles, J. S., & Kenny, S. (2013). Not lack of ability but more choice individual and gender differences in choice of careers in science, technology, engineering, and mathematics. *Psychological Science*, *24*(5), 770–775.

Widaman, K. F., Little, T. D., Preacher, K. J., & Sawalani, G. M. (2011). On creating and using short forms of scales in secondary research. In *Secondary data analysis: An introduction for psychologists.* (pp. 39–61). http://doi.org/10.1037/12350-003

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1–35. http://doi.org/10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In *Development of Achievement Motivation* (pp. 91–120). http://doi.org/10.1016/B978-012750053-9/50006-1

Wigfield, A., Eccles, J. S., Mac Iver, D., Reuman, D. a., & Midgley, C. (1991). Transitions during early adolescence: Changes in children's domain-specific self-perceptions and general self-esteem across the transition to junior high school. *Developmental Psychology*, *27*(4), 552–565. http://doi.org/10.1037/0012-1649.27.4.552

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. a., Freedman-Doan, C., & Blumenfeld, P. C. (1997). Change in children's competence beliefs and subjective task values across the elementary school years: A 3-year study. *Journal of*

Educational Psychology, *89*(3), 451–469. http://doi.org/10.1037/0022-0663.89.3.451

Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and Gender Differences in Children's Self- and Task Perceptions during Elementary School. *Child Development*, *64*(3), 830–847.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1–26.

Yeager, D. S., Bryk, A., Muhich, J., Hausman, H., & Morales, L. (2013). Practical measurement. *Palo Alto, CA: Carnegie Foundation for the Advancement of Teaching*, *78712*.

Yeager, D. S., Walton, G., & Cohen, G. L. (2013). Addressing Achievement Gaps with Psychological Interventions. *Phi Delta Kappan*, *94*(5), 62–65. http://doi.org/10.1177/003172171309400514

Table 1

Analytic Sources of Validity Evidence Utilized in Pragmatic Measurement

Sources		Description
Differences	**A-priori hypotheses regarding contrasts group or time-point contrasts.**	
	Known Groups	Naturally existing groups (e.g., gender, grade levels) are characterized by known differences. Demonstrating these known differences with new scales suggest that the scales are sensitive to existing characteristics.
	Experimental	Artificially created groups or situations are sometimes produced for comparisons. This could be comparisons between treatment and control groups or pre-post comparisons. Observing differences following a change, especially one using random assignment, provide active evidence that manipulating a construct can be detected by the scale.
	Change over Time	Some phenomena change as time progresses. Demonstrating scale sensitivity to change bolsters the likelihood that it is useable.
Relationships	**An alternative view of differences, relationships allow the comparison of multiple constructs.**	
	Expected Correlations	Construct interrelationships can signal that the constructs are being assessed. When the correlations are of the expected magnitudes, they can also function as alternative reliability evidence.
	Test-Retest	The relationship between a construct and itself at another point in time can indicate consistency. This may be able to function as an alternative to internal consistence when change in the construct is not expected between measurement intervals.
	Prediction	The scale can meaningfully explain variation in another construct.
Proxy Scales	**Proxy scales may be simple approximations of more complex ones that can serve as alternative scales of a phenomenon of interest.**	
Targeted Expert Knowledge	**Expert knowledge can help to identify expected differences and expected relationships. Furthermore, it should serve to guide scale selection, scale altering, and scale reduction.**	

Table 2

Typical uses for expectancy-value-cost users, associated assumptions, validity sources, and which studies address these uses/assumptions

Proposed Uses	Assumptions	Validity Sources
1. Measure important motivation phenomena in classrooms	1. Instructors have a sense of important motivational experiences in their classrooms	*Prior Research*
	2. Theory-based scales will measure some of the motivation constructs identified by students	*Prior Research*
	3. Theory-based scales will measure some of the motivation constructs identified by faculty	*Prior Research*
	4. Ratings demonstrate reliability	Differences, Relationships, Targeted Exp Knowledge
	5. If measured effectively, motivational constructs should relate to each other in theoretically appropriate ways	Relationships, Targeted Expert Knowled
	6. If measured effectively, motivational constructs should relate to outcomes in theoretically appropriate ways	Relationships, Targeted Expert Knowled
	7. A reduced number of theoretically chosen items can represent enough of the underlying construct to preserve the relationships observed	Proxy Scales, Relationships, Differences
2. Examine student group differences	1. Groups of students should display motivational differences at the mean level	*Prior Research*, Differences
	2. A reduced number of theoretically chosen items preserve the direction and relative intensity of group differences	Proxy Scales, Differences
3. Monitor student motivation change	1. Student motivation changes during the semester	*Prior Research*
	2. Student motivational change can be captured by the expectancy-value-cost scale	Differences, Targeted Expert Knowledge
	3. A reduced set of items can lead to similar conclusions as the longer scale.	Proxy Scales, Differences
4. Monitor intervention processes	1. Motivation-based interventions should operate by facilitating specific types of motivation	*Prior Research*
	2. During the course of a motivation intervention, relevant scales should demonstrate differences between treatment and control groups	Differences, Targeted Expert Knowledge
	3. Items most aligned with the intervention should yield the largest differences	Proxy Scales, Differences

Table 3
Demographics by Sample

		Sample 1	Sample 2
N		740	1877
Age (Median)		20	19
Gender (%)		58	61
Race (%)			
	Black	20	23
	Hispanic	35	39
	Other	9	7
	White	35	31
	Missing	2	1
Parent Education (%)*			
	Did Not Finish High School	7	9
	High School, No College	24	26
	Some College	20	20
	AA or AS	15	15
	BA or BS	22	20
	MA, MS, or MBA	10	7
	Doctorate: Lawyer, Doctor, or PhD	2	2
	Missing (%)	27	22
Pell (%)*		54	81
High School GPA		3.14	3.11
	Missing (%)	15	13
Last Math Class Grade		2.5	2.8
	Missing (%)	28	25
Pass (%)		77	72

Note. Variables marked with a * display percentages of reported data (i.e., missing responses are not included in the percentage breakdown.

Table 4

Intraclass Correlations for Expectancy, Value, Cost, Interest, and Pass Rate

		Expectancy	Value	Cost	Interest	Pass
Time 1	Sample 1	-0.01	0.04	0.02	0.08	
	Sample 2	0.04	0.10	0.00	0.10	
Time 2	Sample 1	0.04	0.10	0.00	0.07	
	Sample 2	0.03	0.07	0.03	0.09	
Time 3	Sample 1	0.01	0.06	0.04	0.04	
	Sample 2	0.03	0.07	0.02	0.09	
Time 4	Sample 1	0.08	0.06	0.03	0.01	0.11
	Sample 2	0.05	0.06	0.06	0.10	0.07
Sample Means	Sample 1	0.03	0.06	0.02	0.05	0.11
	Sample 2	0.04	0.08	0.03	0.10	0.07
Study Mean		0.04	0.07	0.03	0.07	0.09

Note. All values calculated using only available data for each time point

Table 5

Items with Expert, Random, and Alternate Selection

		Collected During				**Used as Single-item**		
Expectancy		T1	T2	T3	T4	Expert	Random	Alternate
E1	How confident are you that you can learn the material in this class?	x	x	x	x			x
E2	How confident are you that you can be successful in this class?	x	x	x	x	x		
E3	How well do you expect to do in this class?	x			x		x	
E4	How confident are you that you can understand the material in this class?	x			x			
Value								
V1	How important is this class to you?	x	x	x	x			x
V2	How useful will this class be to your career?	x			x			
V3	How valuable is this class to you?	x			x		x	
V4	How useful is this class to you?	x	x	x	x	x		
Cost								
C1	How often does this class require too much time?	x			x			
C2	How often do you feel that you just don't have time to put into this class because of other things that you do?	x			x			
C3	How often are you limited in the amount of effort that you can put into this class?	x			x		x	
C4	How often do you feel that you have to sacrifice too much in order to do well in this class?	x	x	x	x	x		
C5	How stressed out are you by your math class?	x	x	x	x			x

Note. "Expert" indicates item most highly ranked by experts across representativeness, prediction of interest, and (among value items) sensitivity to value intervention. "Random" indicates item randomly selected among each scale. "Alternate" indicates the alternative item used as a comparison to expert items when random items were not available for analyses (i.e., interim data collection points). Composites were calculated using all available items at each time point, the composites calculated at Time 1 and Time 4 consisted of more items than the composites calculated at Time 2 and Time 3.

Table 6

Expert and Random Single-Item Means, Standard Deviations, and Correlations by Sample

		Expert		Random				
Sample 1		**M**	**SD**	**M**	**SD**	**1.**	**2.**	**3.**
	1. Expectancy	3.91	0.97	3.97	0.84		.25	.03
	2. Value	3.28	1.15	3.69	1.09	.50		.02
	3. Cost	2.44	1.20	2.38	1.13	-.14	.04	
Sample 2		**M**	**SD**	**M**	**SD**	**1.**	**2.**	**3.**
	1. Expectancy	3.87	0.96	3.925	0.85		.28	-.04
	2. Value	3.42	1.1	3.721	1.05	.39		.01
	3. Cost	2.29	1.19	2.391	1.12	-.16	-.07	

Note. $N_{Sample1}$ = 737. $N_{Sample2}$ = 1855. All descriptives estimated using full information maximum likelihood to account for missing data. Correlations above the diagonal represent correlations between Randomly selected items. Correlations below the diagonal represent correlations between Expert selected items.

Table 7

Reliability Evidence for Composite Scales, Expert- and Random-Selected Items Including Interest and Pass-Status

		Expectancy		Value		Cost		Interest		Pass	
		Sample 1[A]	Sample 2[B]	Sample 1[A]	Sample 2[B]	Sample 1[A]	Sample 2[B]	Sample 1[A]	Sample 2[B]	Sample 1[A]	Sa
Expectancy	Composites			0.42	0.41	-0.28	-0.29	0.45	0.48	.12	.
	Expert Selection	**0.90**	**0.90**	0.40	0.36	-0.26	-0.25	0.43	0.43	.13	.
	Random Selection	**0.82**	**0.85**	0.38	0.36	-0.16	-0.21	0.42	0.42	.13	.
Value	Composites	0.42	0.41			-0.04	-0.07	0.64	0.69	.08	.
	Expert Selection	0.50	0.42	**0.86**	**0.88**	-0.05	-0.10	0.64	0.64	.10	-
	Random Selection	0.29	0.36	**0.82**	**0.85**	-0.08	-0.06	0.55	0.55	.08	-
Cost	Composites	-0.28	-0.29	-0.04	-0.07			0.03	-0.09	.01	.
	Expert Selection	-0.11	-0.16	0.01	-0.03	**0.77**	**0.80**	-0.01	-0.01	.05	.
	Random Selection	-0.02	-0.09	0.06	0.02	**0.68**	**0.67**	0.04	0.04	.01	.

Note. [A] Correlations from Sample 1 greater than .12 are statistically significant (*p* < .05). [B] Correlations from Sample 2 greater than .06 are statistical significant (*p* < .05). All correlations estimated using Full Information Maximum Likelihood Estimation with auxiliary variables. Bolded values indic correlations between single-item measures and composites.

Table 8

Test-Retest Reliabilities for Adjacent and Pre-Post Time Points

			Sample 1 (n = 740)			Sample 2 (n = 1855)		
			Composite	Expert	Random*	Composite	Expert	Random*
Expectancy	Adjacent	T1-T2	0.51	0.56		0.54	0.55	
		T2-T3	0.51	0.46		0.66	0.65	
		T3-T4	0.48	0.50		0.58	0.58	
		Mean	**0.50**	**0.50**		**0.59**	**0.59**	
	Pre-Post	T1-T4	0.43	0.45	0.35	0.45	0.48	0.43
Value	Adjacent	T1-T2	0.59	0.81		0.55	0.71	
		T2-T3	0.59	0.76		0.71	0.86	
		T3-T4	0.57	0.66		0.72	0.83	
		Mean	**0.58**	**0.74**		**0.66**	**0.80**	
	Pre-Post	T1-T4	0.35	0.54	0.47	0.61	0.74	0.58
Cost	Adjacent	T1-T2	0.49	0.79		0.49	0.68	
		T2-T3	0.67	0.74		0.77	0.84	
		T3-T4	0.41	0.51		0.49	0.66	
		Mean	**0.52**	**0.68**		**0.58**	**0.73**	
	Pre-Post	T1-T4	0.48	0.54	0.50	0.41	0.67	0.49

Note. *Random items were not collected during the Time 2 and Time 3.

Table 9

R^2 by Values Scale for Time 1 Interest, Time 4 Interest, and Pass Rate

		Sample 1		Sample 2	
		R^2	SE	R^2	SE
Time 1 Interest	Composite	0.449*	0.047	0.519*	0.022
	Expert	0.400*	0.046	0.448*	0.022
	Random	0.310*	0.047	0.370*	0.023
Time 4 Interest	Composite	0.247*	0.054	0.343*	0.032
	Expert	0.242*	0.055	0.296*	0.031
	Random	0.147*	0.044	0.276*	0.031
Pass	Composite	0.016	0.022	0.029*	0.015
	Expert	0.020	0.021	0.024†	0.012
	Random	0.016	0.016	0.026*	0.013

Note. *p < .05, † p < .10. For each sample, all outcomes were estimated in a single model for each scale and were predicted by (Composite, Expert, or Random) from Time 1.

Table 10

Summary Fit Statistics from Growth Curve Model Building, Cross Validation, and Comparison across Scale Construction

	Description	Scale	χ^2	df	RMSEA	CFI	TLI	SRMR	AIC	BIC	ABIC
Sample 1	E only I,S	Composite	3.63	5	0.00	1.00	1.01	0.02	3721.55	3779.71	3738.43
	E only I,S (Fixed Slope)	**Composite**	**7.47**	**7**	**0.01**	**1.00**	**1.00**	**0.06**	**3721.95**	**3771.16**	**3736.24**
	V only I,S,Q	Composite	9.62*	2	0.08	0.99	0.93	0.04	3660.36	3731.95	3681.15
Model Building Phase	**V only ISQ (Fixed Slope)**	**Composite**	**14.55***	**5**	**0.05**	**0.98**	**0.96**	**0.03**	**3656.62**	**3714.78**	**3673.50**
	C only ISQ	**Composite**	**5.26**	**2**	**0.05**	**0.99**	**0.95**	**0.02**	**3913.63**	**3985.21**	**3934.41**
	EV (Free Slopes)	Composite	69.34*	17	0.07	0.95	0.90	0.04	7217.18	7373.76	7262.64
	EV (V Slope Fixed)	Composite	86.32*	22	0.07	0.94	0.90	0.04	7225.23	7359.45	7264.20
	EV (EV Slopes Fixed)	Composite	138.35*	26	0.08	0.89	0.85	0.05	7274.31	7390.63	7308.08
	EVC (EV Slopes Fixed)							*NO CONVERGENCE*			
Sample 2	E only IS (Slope Fixed)	Composite	61.76*	7	0.07	0.96	0.94	0.05	10510.69	10570.38	10535.44
	E only IS (Slope Freed)	**Composite**	**52.56***	**5**	**0.08**	**0.96**	**0.93**	**0.05**	**10508.72**	**10579.27**	**10537.97**
	V only I,S,Q Slope Fixed	Composite	77.41*	5	0.09	0.96	0.92	0.05	10783.65	10854.19	10812.89
Cross-Validation Phase	**V only I,S,Q (Slope Freed)**	**Composite**	**38.75***	**2**	**0.11**	**0.98**	**1.07**	**0.05**	**10731.15**	**10817.98**	**10767.15**
	C only I,S,Q	**Composite**	**1.89**	**2**	**0.00**	**1.00**	**1.00**	**0.05**	**10951.41**	**11038.23**	**10987.40**
	EV only (free slopes)	Composite	187.71*	17	0.08	0.99	0.95	0.03	20837.95	21027.88	20916.69
	EVC (Free Slopes)	**Composite**	**339.07***	**38**	**0.07**	**0.95**	**0.89**	**0.03**	**31349.18**	**31696.48**	**31493.16**
Sample 1	E Expert	Expert	9.97	7	0.03	0.96	0.94	0.05	4055.35	4104.56	4069.63
	E Alternate	Alternate	10.72	7	0.03	0.96	0.94	0.05	4306.01	4355.23	4320.30
	V Expert	Expert	10.03*	5	0.04	0.96	0.94	0.05	4425.25	4483.41	4442.14
Comparison Phase	V Alternate	Alternate	8.00*	5	0.03	0.96	0.94	0.05	4154.58	4212.74	4171.46
	C Expert	Expert	4.24*	2	0.04	0.99	0.96	0.05	4678.69	4750.27	4699.47
	C Alternate	Alternate	6.20*	2	0.06	0.96	0.94	0.05	4690.53	4762.11	4711.32
Sample 2	E Expert	Expert	30.52*	7	0.05	0.98	0.97	0.05	11657.04	11716.73	11681.79
	E Alternate	Alternate	27.03*	7	0.04	0.98	0.97	0.05	12262.44	12322.13	12287.18

Comparison Phase	V Expert	Expert	2.55	2	0.01	1.00	1.00	0.03	12710.38	12797.20	12746.37
	V Alternate	Alternate	12.30*	2	0.06	0.99	0.95	0.03	12267.80	12354.63	12303.80
	C Expert	Expert	0.36	2	0.00	1.00	1.01	0.05	13535.46	13622.28	13571.45
	C Alternate	Alternate	6.32*	2	0.04	1.00	0.99	0.05	13249.73	13336.55	13285.72
	EVC	Expert	269.74*	38	0.06	0.95	0.89	0.03	37527.63	37874.93	37671.61
	EVC	Alternate	270.10*	38	0.06	0.95	0.90	0.03	36999.02	37346.32	37143.00

Note. E = Expectancy, V = Value, C = Cost. Composite = Composite Scale, Expert = Expert-Selected Scale Alternate = Alternative Scale Item. I = Intercept estimated, V = Slope estimated, Q = quadratic growth estimated. Bolded models chosen for cross-validation and short-scale comparisons. * p < .05. Guidelines vary for fit indices, but following is a list of commonly-used benchmarks: $\chi2$, non-significant; RMSEA < .06; CFI > .95; TLI > .95; SRMR < .08. AIC, BIC, and ABIC have no benchmarks but are used for model comparison with lower numbers being desirable. AIC, BIC, and ABIC should only be considered when comparing models using the same variables.

Table 11

Sample 2 Growth Parameter Estimated Means and Standard Deviations

	Composite			Expert			Alternate		
	Unstandardized	SD	Standardized	Unstandardized	SD	Standardized	Unstandardized	SD	Standardized
EI	3.779*	0.750*	5.041	3.812*	0.774*	4.924	3.695*	0.792*	4.665
ES	-0.036*	0.105*	-0.344	-0.048*	0.130*	-0.365	-0.010	0.126*	-0.08
VI	3.891*	0.691*	5.630	3.540*	0.825*	4.289	4.237*	0.719*	5.892
VS	-0.175*	0.249*	-0.701	-0.082*	0.268	-0.307	-0.264*	0.352*	-0.751
VQ	0.015*	0.032*	0.432	0.004	0.045*	0.115	0.026*	0.045*	0.563
CI	2.477*	0.820*	3.022	2.313*	1.176*	1.968	2.658*	1.207*	2.202
CS	0.095*	0.5668*	0.168	0.181*	0.731*	0.247	-0.012	0.647*	-0.019
CQ	-0.010*	0.071*	-0.139	-0.021*	0.089*	-0.234	0.004	0.077*	0.056

Note. * *p* < .05. E = Expectancy, V = Value, C = Cost, I = Intercept, S = Slope, Q = Quadratic. Unstandardized Growth Parameters were compiled from the full simultaneous growth model using each set of scales. The composite scales at Time 2 and Time 3 were shortened versions of the full scale. Unstandardized parameters should be used to compare results across scale versions (i.e., composite, expert, alternate). Standardized parameters should be used to compare values within individual scale versions (e.g., expert ES is larger than expert VS).

Table 12
Correlations Between Growth Parameters from Simultaneous Growth Model

Composite	EI	ES	VI	VS	VQ	CI	CS
Expectancy Intercept (EI)							
Expectancy Slope (ES)	-0.11*						
Value Intercept (VI)	0.51*	-0.08					
Value Slope (VS)	0.09	0.11	0.37				
Value Quadratic (VQ)	-0.06	0.07	-0.30	-0.93*			
Cost Intercept (CI)	-0.34*	-0.02	-0.09*	0.05	-0.07		
Cost Slope (CS)	-0.24*	-0.16*	0.04	-0.05	0.02	-0.29*	
Cost Quadratic (CQ)	0.27*	0.10*	-0.03	0.04	0.00	0.22*	-0.98*
Expert							
Expectancy Intercept (EI)							
Expectancy Slope (ES)	-0.19*						
Value Intercept (VI)	0.54*	-0.07					
Value Slope (VS)	-0.03	0.06	0.16				
Value Quadratic (VQ)	0.04	0.08	-0.12	-0.92*			
Cost Intercept (CI)	-0.19*	-0.02	-0.07*	0.05	-0.06		
Cost Slope (CS)	-0.04	-0.06	0.09*	-0.13	0.10	-0.58*	
Cost Quadratic (CQ)	0.05	0.01	-0.10*	0.11	-0.09	0.52*	-0.98*
Alternate							
Expectancy Intercept (EI)							
Expectancy Slope (ES)	-0.17*						
Value Intercept (VI)	0.28*	-0.04					
Value Slope (VS)	0.22*	-0.01	0.09				
Value Quadratic (VQ)	-0.20*	0.08	-0.10	-0.94*			
Cost Intercept (CI)	-0.54*	0.02	-0.04	-0.02	0.00		
Cost Slope (CS)	-0.01	-0.13*	0.10*	0.00	-0.01	-0.50*	
Cost Quadratic (CQ)	0.04	0.05*	-0.09*	0.01	0.00	0.42*	-0.97*

Note.* p < .05

Table 13

Multiple Regression Results Showing Intervention Sensitivity in Composite, Expert, and Alternative Items.

			Intervention									
	Scale	Sample	Effect	χ^2	*df*	RMSEA	CFI	TLI	SRMR	AIC	BIC	ABIC
Value	Composite	1	-0.214*	6.33	3	0.04	0.99	0.96	0.03	22390.32	23436.03	22715.23
	Composite	2	0.004	4.29	3	0.02	1.00	1.00	0.02	61027.73	62282.05	61560.87
	Expert	1	-0.171†	4.60	3	0.03	0.99	0.97	0.03	22654.95	23700.66	22979.85
	Expert	2	0.023	4.28	3	0.02	1.00	0.99	0.02	61454.97	62709.29	61988.12
	Alternate	1	-0.114	3.26	3	0.01	1.00	1.00	0.02	22624.28	23669.99	22949.19
	Alternate	2	-0.018	4.29	3	0.02	1.00	0.99	0.02	61620.21	62874.53	62153.35

	Time		Intervention									
	Point	Sample	Effect	χ^2	*df*	RMSEA	CFI	TLI	SRMR	AIC	BIC	ABIC
Manipulation Check	Time 2	1	0.159	4.53	3	0.03	0.99	0.98	0.03	22529.80	23575.51	22854.70
	Time 2	2	0.246*	5.04	3	0.02	1.00	0.99	0.02	61421.70	62676.02	61954.85
	Time 3	1	-0.035	3.08	3	0.01	1.00	1.00	0.02	22502.16	23547.87	22827.06
	Time 3	2	0.113	4.40	3	0.02	1.00	0.99	0.02	61237.02	62491.34	61770.16
Pass	Post-Semester	1	0.106							3073.89	3138.38	3093.93
		2	0.065							11381.87	11459.23	11414.75

*Note.** *p* < .05. † *p* < .10 "Intervention effect" is the unstandardized regression coefficient from a dummy indicator for the treatment condition (0 = control, 1 = treatment). "Value" indicates the *value* item a measured a one to seven days after the intervention; this item was ranked as most sensitive to intervention effects by the expert rater.). "Manipulation Check" indicates the 2-item measure immediately used after the intervention to test for intervention effects. Fit indices were not available for the analyses using Pass as a categorical outcome.

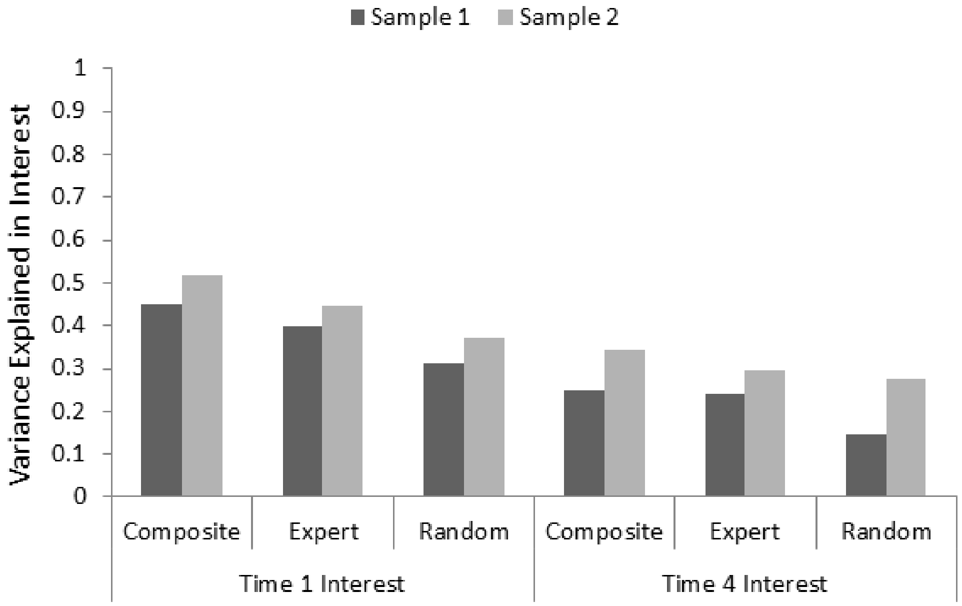

Figure 1. R-Square Values for Time 1 and Time 4 Interest by Scale and Sample

Appendix A: Paper 1—Supplemental Materials

Literature Review and Search Parameters

In an effort to compile empirical knowledge from the literature on expectancy and value change, we conducted an in-depth literature review using focused search parameters. We believe our search parameters to be extensive and to provide some insight into the general state of the literature. Articles were only included in our review if they included at least two time points of data within a single year and at least one of the primary constructs (expectancy or value). The two-time-point restriction was chosen because that is the minimum number of times necessary to assess basic change (i.e., a difference score), though we had hoped to find articles with three or more time points which could be used to assess complex change (i.e., growth modeling). The within-a-year restriction was chosen because classes are typically one year long in elementary through high school and a semester long in late high school and college. Thus, we wanted to assess motivation change within a single learning environment (i.e., beginning of semester and end of semester) or at least within close temporal proximity (i.e., beginning of semester, end of semester, post-semester follow-up). Because we were focused on *natural change* in the constructs, we excluded pre-post measures from interventions (though data collected from control groups was included). We elaborate on the search terms used below.

Search Methodology and Findings

Primary search. Using Google Scholar and PsycNET we included the following search terms “Expectancy-Value,” “College,” “Education,” and “Longitudinal.” The initial search results produced 5,270 articles, though the relevance of the articles rapidly deteriorated by the 100th result (page 10). A modified search included "Expectancy" and "Value" separately, increasing the number of search results to 80,300. Article relevance declined similarly. Both searches produced fewer than 10 articles that met our inclusion criteria, the majority of which were found early in the search results. Because we were simply searching to get a sense of the literature rather than a comprehensive compilation, we ended the search after several pages produced no new studies.

Expanded search. In an effort to bolster our pool of studies, we decided to broaden our search parameters to include additional related constructs (Pintrich, 2003) that overlap with expectancy such as "competence beliefs," "perceived competence," and "self-efficacy," as well as constructs that overlap with value such as "intrinsic motivation," "interest," and "relevance." We also expanded the search to pre-college educational settings ranging from first through twelfth grade. Despite using various combinations of these search terms, the final pool of studies was still only 18.

Final study pool. The list of studies found is included in Table 1 of the main manuscript—it also includes a small list of studies that did not meet the search criteria but provide information about long-term motivation change. Of the studies produced by our searches, 13 included two time points, three included three time points, two included four time points, and one included many time points. We note that one of the two time point studies did not include information from the control group (i.e., natural change) and the many time point study aggregated student motivation measures taken over many days into single composites—thus not reporting any short-term change values in the manuscript. Many of the two time point studies were experimental or quasi-experimental in nature—in those cases we only examined the control group data when available. One of these studies failed to include descriptive statistics for their control group.

Conclusion. Our search was not meant to be comprehensive, nor was it meant to rival a meta-analysis. However, we believe that our search does illuminate the relative lack of studies that are focused on short-term expectancy and value change. We do not make any claims that such studies do not exist—rather we suggest the studies that do exist are not necessarily easy to find. Based on our search, it appears that researchers are most interested in short-term change in the context of pre-post measures of intervention. That is not to say that there is no interest in short-term change, only that the focus has remained relatively narrow.

Expanded Results

Because of the complexity of untrimmed Model B, visual representation of the full model was not plausible. Table S4 contains the full specification for Model B. Each row represents a predictor variable and each column represents a predicted variable. Thus, to determine if one variable predicted another (e.g., if Sex predicted Exam 1), start with the variable Sex in the first row of the table and then examine the 5th column (Exam1); the presence of an arrow (→) in this cell indicates that sex was included as a predictor of Exam 1. To locate the path coefficient for sex predicting exam 1, begin with the column labeled Sex and continue to the row labeled Exam 1; therefore, the effect of sex on Exam 1 is .82. Similarly, some concurrent variables were merely correlated with each other and marked as *C* rather than as a path. For example, the effect of INT1 on VI was a correlation (the second row of the fourth column) of magnitude .68 (the fourth row of the second column. Any variable that was predicted by other variables also had an associated R^2 value (bolded values on the table's diagonal). For example, EI was predicted by Sex and Int1, which explained a total of 6% of the variance ($R^2 = .06$). Below the table, we also include the MPlus model specification for readers familiar with the software.

Table S4
Conditional Growth Model: Model B Path Specifications, Path Coefficients, and Standard Errors

	SEX	INT1	EI	VI	EXAM1	EXAM2	EXAM3	ES	VS	INT3	EX
SEX		→	→	→	→	→	→	→	→	→	
INT1	0.21 (0.23)	**.01**	C	C	→	→	→	→	→	→	
EI	-0.22* (0.04)	0.35* (0.07)	**.06**	C	→	→	→			→	
VI	0.12 (0.09)	0.68* (0.09)	0.24* (0.05)	**.01**	→	→	→			→	
EXAM1	0.82 (0.79)	-0.58 (0.44)	6.65* (1.35)	-1.23 (0.91)	**.21**	→	→	→	→	→	
EXAM2	-0.05 (0.66)	-0.27 (0.38)	2.31† (1.30)	-0.39 (0.80)	0.62* (0.07)	**.48**	→	→	→	→	
EXAM3	0.30 (0.55)	0.53† (0.28)	0.16 (1.21)	-0.94 (0.59)	0.58* (0.06)	0.34* (0.06)	**.60**	→	→	→	
ES	0.03 (0.05)	-0.02 (0.02)			0.01 (0.01)	0.02* (0.01)	0.02* (0.01)	**.52**	C	→	
VS	-0.03 (0.05)	0.01 (0.02)			-0.00 (0.01)	0.01† (0.01)	0.01* (0.01)	0.06* (0.02)	**.18**	→	
INT3	0.12 (0.17)	0.67* (0.08)	-0.22 (0.27)	0.35* (0.17)	0.01 (0.02)	0.01 (0.02)	0.00 (0.02)	1.03* (0.48)	0.30 (0.76)	**.65**	
EXAM4	0.17 (0.57)	-0.12 (0.27)	-1.05 (1.03)	0.55 (0.69)	0.32* (0.07)	0.25* (0.07)	0.30* (0.07)	1.22 (2.15)	0.98 (3.47)		**.**

Note. χ^2 (27) = 64.42, *p* < .01, RMSEA = .07, CFI = .98, TLI = .94, SRMR = .05. Cells above the diagonal indicate whether a (→) or a correlation (C) was estimated for the path model. Blank cells indicate that a path was not estimated. Values on the diagonal (bolded) represent R-squared estimates. Cells below the diagonal represent the unstandardized path coefficient or covariance between the two variables; standard errors are included in parentheses.

MPLUS Syntax

title: Dual process model SEM, conditional growth model

data: file is 202_S07_Complete_small_2_mplus.csv;

variable: names are snum sex Exam1 Exam2 Exam3 Exam4

CumFin XC_new Tot_acad Tot_wXC w1v w2v

```
 w3v w1e w2e w3e w3cint w1cint;

                        usevariables are  w1v w2v w3v w1e w2e w3e
 Exam1 Exam2 Exam3 Exam4 sex w1cint w3cint;
                        missing are all (-99);

Analysis:
bootstrap = 1000;

model:   ei es | w1e@0 w2e@1 w3e@2;
 w3e@0;
 !fixing the residual variance to zero;
 vi vs | w1v@0 w2v@1 w3v@2;
                        !time codes correspond to number of weeks;
 !vs@0;

 !fixing variance of value slope to zero;
 vi with ei;
 vi with vs@0;
 es with ei@0;
 es with vs;
 ei with w1cint;
 vi with w1cint;
 exam4 with w3cint @0;
 exam4 with vs@0;
 w3cint with vs@0;

w3cint on es vi w1cint;
exam4 on exam3 exam2 exam1 ;
exam3 on exam2 exam1;
exam2 on exam1;
exam1 on ei;

es on exam2 exam3;
vs on exam3;

ei on sex;


output: tech4 stdyx mod cinterval(bcbootstrap)
```

Appendix B: Paper 2—The Expectancy-Value-Cost scale

1	2	3	4	5	6
Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree

E1 I know I can learn the material in my [math or science] class.

E2 I believe that I can be successful in my [math or science] class.

E3 I am confident that I can understand the material in my [math or science] class.

V1 I think my [math or science] class is important.

V2 I value my [math or science] class.

V3 I think my [math or science] class is useful.

C1 My [math or science] classwork requires too much time.

C2 Because of other things that I do, I don't have time to put into my [math or science] class.

C3 I'm unable to put in the time needed to do well in my [math or science] class.

C4 I have to give up too much to do well in my [math or science] class.

Appendix C: Paper 2—Supplemental Materials

Table 1: Item Descriptives for Math and Science in Fall and Winter Samples

Table 2: Item Response Category Frequency Distributions by Construct, Academic Domain, and Time

Table 3: Longitudinal Item Correlations for Math and Science Motivation

Table 4: Item Parameters for Math and Science Longitudinal Measurement Invariance

Table 5. Convergent and Discriminant Correlations with Achievement and Future Interest

Figure 1: Observed Longitudinal Measurement Invariance Model.

Table 1.
Item Descriptives For Math and Science in Fall and Winter Samples

	Math					Science				
Fall (n = 311)	n	M	SD	Min	Max	N	M	SD	Min	Max
E1	375	4.96	0.99	1	6	328	5.01	0.91	1	6
E2	375	5.12	1.01	1	6	328	5.08	1.00	1	6
E3	374	4.75	1.13	1	6	328	4.84	1.03	1	6
V1	375	5.12	0.99	1	6	328	4.84	1.01	1	6
V2	375	4.62	1.19	1	6	328	4.61	1.08	1	6
V3	375	4.96	1.10	1	6	328	4.70	1.22	1	6
C1	375	3.36	1.40	1	6	328	2.59	1.28	1	6
C2	375	2.78	1.44	1	6	328	2.49	1.33	1	6
C3	374	2.54	1.37	1	6	328	2.50	1.36	1	6
C4	374	2.61	1.42	1	6	328	2.33	1.23	1	6
	Math					Science				
Winter (n = 270)	n	M	SD	Min	Max	N	M	SD	Min	Max
E1	338	4.99	0.89	2	6	317	5.09	0.95	1	6
E2	345	5.03	0.99	1	6	317	5.14	0.97	1	6
E3	345	4.87	1.07	1	6	317	4.97	1.09	1	6
V1	345	5.10	1.00	1	6	317	5.03	1.06	1	6
V2	342	4.61	1.09	1	6	317	4.79	1.15	1	6
V3	346	4.93	1.08	1	6	317	4.84	1.24	1	6
C1	342	2.93	1.41	1	6	317	2.33	1.22	1	6
C2	345	2.56	1.33	1	6	317	2.28	1.28	1	6
C3	345	2.48	1.34	1	6	317	2.25	1.26	1	6
C4	344	2.40	1.29	1	6	317	2.18	1.17	1	6

Note. N = 401. Students were eligible to be in the Fall Sample if they had data from Math and Science in the Fall. Students were eligible to be in the Winter Sample if they had data from Math and Science in the Winter. In the full sample, students could have two, three, or four instances of data. Only 180 students had full data, thus the number of students with data for any given item may fluctuate.

Table 2.
Item Response Category Frequency Distributions (%) by Construct, Academic Domain, and Time

		Expectancy				Value				Cost			
		Math		Science		Math		Science		Math		Science	
		FA[1]	WI[2]	FA[1]	WI[2]	FA[1]	WI[2]	FA[1]	WI[2]	FA[1]	WI[2]	FA[1]	WI[2]
Item 1	*												
Response Options	**1**	2	0	1	1	2	1	1	2	9	14	21	27
	2	2	3	1	1	2	1	3	2	21	32	34	38
	3	4	3	5	3	2	3	4	4	25	22	21	16
	4	14	14	17	13	11	11	21	14	23	18	15	11
	5	50	52	45	44	45	44	45	41	15	8	6	6
	6	29	28	32	37	39	38	26	37	8	7	2	1
Item 2													
Response Options	**1**	1	1	1	1	3	2	1	2	17	22	27	32
	2	2	2	1	2	5	3	3	3	39	37	33	36
	3	5	2	5	2	7	8	8	7	14	19	16	15
	4	9	14	12	14	22	23	30	21	14	12	16	9
	5	43	47	42	41	42	46	37	38	11	8	5	6
	6	40	33	38	41	22	18	21	30	5	3	3	2
Item 3													
Response Options	**1**	2	2	2	1	2	1	3	2	23	23	24	32
	2	3	1	1	3	2	3	5	5	39	41	39	36
	3	6	7	7	6	4	5	6	5	13	14	15	15
	4	20	14	20	14	15	13	20	16	13	11	11	10
	5	43	47	44	40	43	46	39	35	9	8	8	5
	6	26	28	27	36	34	32	28	36	3	3	3	2
Item 4													
Response Options	**1**									23	23	27	32
	2									36	46	40	39
	3									15	13	13	16
	4									15	8	13	8
	5									5	7	4	4
	6									6	3	2	2

Note. * Values in this column represent the values of the item response scale from 1 (Strongly Disagree) to 6 (Strongly Agree). [1] FA = Fall. [2] WI = Winter. The values presented in the table are in percentages to account for different numbers of responders.

Table 3.
Longitudinal Item Correlations for Math and Science Motivation

Time		**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**
		E1	**E2**	**E3**	**V1**	**V2**	**V3**	**C1**	**C2**	**C3**	**C4**
1	**E1**		.69	.67	.57	.63	.58	-.35	-.40	-.41	-.40
1	**E2**	.73		.65	.57	.61	.58	-.31	-.37	-.33	-.34
1	**E3**	.69	.70		.57	.60	.56	-.27	-.38	-.33	-.42
1	**V1**	.60	.52	.48		.73	.76	-.29	-.38	-.26	-.31
1	**V2**	.60	.58	.54	.67		.71	-.28	-.40	-.33	-.33
1	**V3**	.48	.51	.49	.60	.56		-.25	-.37	-.28	-.30
1	**C1**	-.28	-.25	-.29	-.35	-.35	-.29		.59	.48	.55
1	**C2**	-.33	-.34	-.32	-.34	-.38	-.33	.58		.66	.63
1	**C3**	-.35	-.39	-.36	-.31	-.34	-.30	.46	.62		.58
1	**C4**	-.28	-.36	-.33	-.38	-.39	-.34	.57	.60	.59	
2	**E1**	.55	.54	.58	.26	.33	.37	-.35	-.37	-.44	-.37
2	**E2**	.41	.47	.44	.23	.29	.29	-.31	-.35	-.35	-.31
2	**E3**	.51	.52	.57	.25	.35	.35	-.36	-.34	-.38	-.34
2	**V1**	.36	.42	.34	.47	.53	.45	-.34	-.33	-.30	-.38
2	**V2**	.43	.44	.43	.47	.61	.46	-.33	-.31	-.32	-.34
2	**V3**	.31	.36	.33	.38	.47	.50	-.34	-.29	-.30	-.31
2	**C1**	-.34	-.34	-.35	-.22	-.34	-.27	.54	.55	.46	.53
2	**C2**	-.32	-.39	-.36	-.32	-.42	-.30	.43	.58	.50	.54
2	**C3**	-.30	-.38	-.36	-.30	-.37	-.31	.42	.53	.56	.54
2	**C4**	-.31	-.33	-.34	-.31	-.39	-.28	.43	.49	.50	.57

Time		**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
		E1	**E2**	**E3**	**V1**	**V2**	**V3**	**C1**	**C2**	**C3**	**C4**
1	**E1**	.48	.51	.49	.36	.42	.48	-.44	-.40	-.42	-.42
1	**E2**	.44	.53	.47	.40	.44	.47	-.36	-.36	-.34	-.35
1	**E3**	.49	.52	.50	.49	.47	.45	-.44	-.40	-.45	-.44
1	**V1**	.38	.39	.40	.56	.54	.55	-.32	-.32	-.35	-.31
1	**V2**	.42	.43	.47	.53	.62	.58	-.39	-.38	-.34	-.37
1	**V3**	.39	.39	.42	.55	.53	.67	-.32	-.36	-.32	-.35
1	**C1**	-.30	-.32	-.31	-.24	-.27	-.28	.44	.36	.38	.33
1	**C2**	-.36	-.30	-.39	-.35	-.36	-.35	.51	.52	.49	.42
1	**C3**	-.33	-.29	-.39	-.27	-.26	-.27	.45	.44	.38	.42
1	**C4**	-.39	-.34	-.43	-.28	-.28	-.24	.46	.42	.47	.55
2	**E1**		.77	.77	.71	.65	.61	-.47	-.39	-.45	-.57
2	**E2**	.64		.78	.59	.58	.63	-.44	-.36	-.45	-.52
2	**E3**	.72	.71		.61	.60	.66	-.48	-.43	-.46	-.52
2	**V1**	.40	.40	.40		.77	.73	-.41	-.40	-.38	-.49
2	**V2**	.45	.43	.47	.70		.70	-.36	-.36	-.37	-.43
2	**V3**	.37	.42	.45	.69	.62		-.39	-.37	-.35	-.42
2	**C1**	-.44	-.46	-.44	-.40	-.37	-.36		.67	.64	.64
2	**C2**	-.40	-.32	-.43	-.42	-.37	-.32	.62		.71	.70
2	**C3**	-.40	-.42	-.48	-.34	-.40	-.32	.58	.70		.68
2	**C4**	-.35	-.31	-.42	-.37	-.33	-.35	.55	.68	.65	

Note. Values above the diagonals represent science correlations, values below the diagonals represent math correlations. Because all CFA and Invariance Modeling was performed using full information maximum likelihood, correlations displayed were calculating using pairwise deletion.

Table 4.
Item Parameters for Math and Science Longitudinal Measurement Invariance

		Unstandardized			Standardized		
	Item	Intercept (τ)	Factor Loading (λ)	Error Variance (ε)	Factor Loading Time 1 (λ)	Factor Loading Time 2 (λ)	Auto-correlations
Math Expectancy (ω = .88)	e1	4.96	0.82	0.26	.85	.83	.12
	e2	5.06	0.85	0.32	.83	.81	.00
	e3	4.79	0.96	0.36	.85	.83	.11
Math Value (ω = .84)	v1	5.11	0.83	0.31	.83	.83	.02
	v2	4.62	0.95	0.41	.83	.83	.27
	v3	4.95	0.82	0.54	.75	.74	.25
Math Cost (ω = .86)	c1	3.26	1.02	0.99	.72	.71	.23
	c2	2.80	1.16	0.62	.83	.82	.12
	c3	2.63	1.06	0.74	.78	.77	.13
	c4	2.63	1.08	0.71	.79	.78	.15
Science Expectancy (ω = .88)	e1	5.04	0.77	0.21	.86	.88	.05
	e2	5.09	0.79	0.29	.83	.85	.19
	e3	4.88	0.85	0.32	.83	.85	-.06
Science Value (ω = .88)	v1	4.90	0.88	0.25	.87	.88	.11
	v2	4.66	0.92	0.34	.84	.86	.27
	v3	4.73	1.00	0.45	.83	.84	.50
Science Cost (ω = .87)	c1	2.53	0.93	0.72	.74	.73	.13
	c2	2.47	1.11	0.50	.84	.84	.17
	c3	2.45	1.05	0.64	.80	.79	-.21
	c4	2.32	0.97	0.51	.80	.80	.37

Note. The ω coefficient indicates the subscale’s reliability. Unstandardized factor loadings (pattern coefficients), intercepts, and error variances were constrained to be equal across time, thus only a single value is reported for each. Due to different variances, standardized factor loadings differ and are therefore reported for each time separately.

Table 5.
Convergent and Discriminant Correlations with Achievement and Future Interest

		Math					
			Fall			Winter	
		E	V	C	E	V	C
Math Interest	*r*	.59*	.68*	-.36*	.53*	.70*	-.44*
	n	311	311	311	270	270	270
Math Achievement	*r*	.18*	.14*	-.17*	.18*	.15*	-.17*
	n	305	305	305	262	262	262
		Science					
		E	V	C	E	V	C
Science Interest	*r*	.61*	.76*	-.38*	.70*	.76*	-.47*
	n	311	311	311	270	270	270
Science Achievement	*r*	.39*	.26*	-.29*	.47*	.35*	-.41*
	n	49	49	49	86	86	86

Note. Because of achievement tests are not administered every year, correlations are calculated using pairwise deletion and calculated for students who have appropriate scores available. The science achievement correlations are drastically reduced because the test is only taken by 6th graders. *$p < .05$.

Figure 1. Observed Longitudinal Measurement Invariance Model.

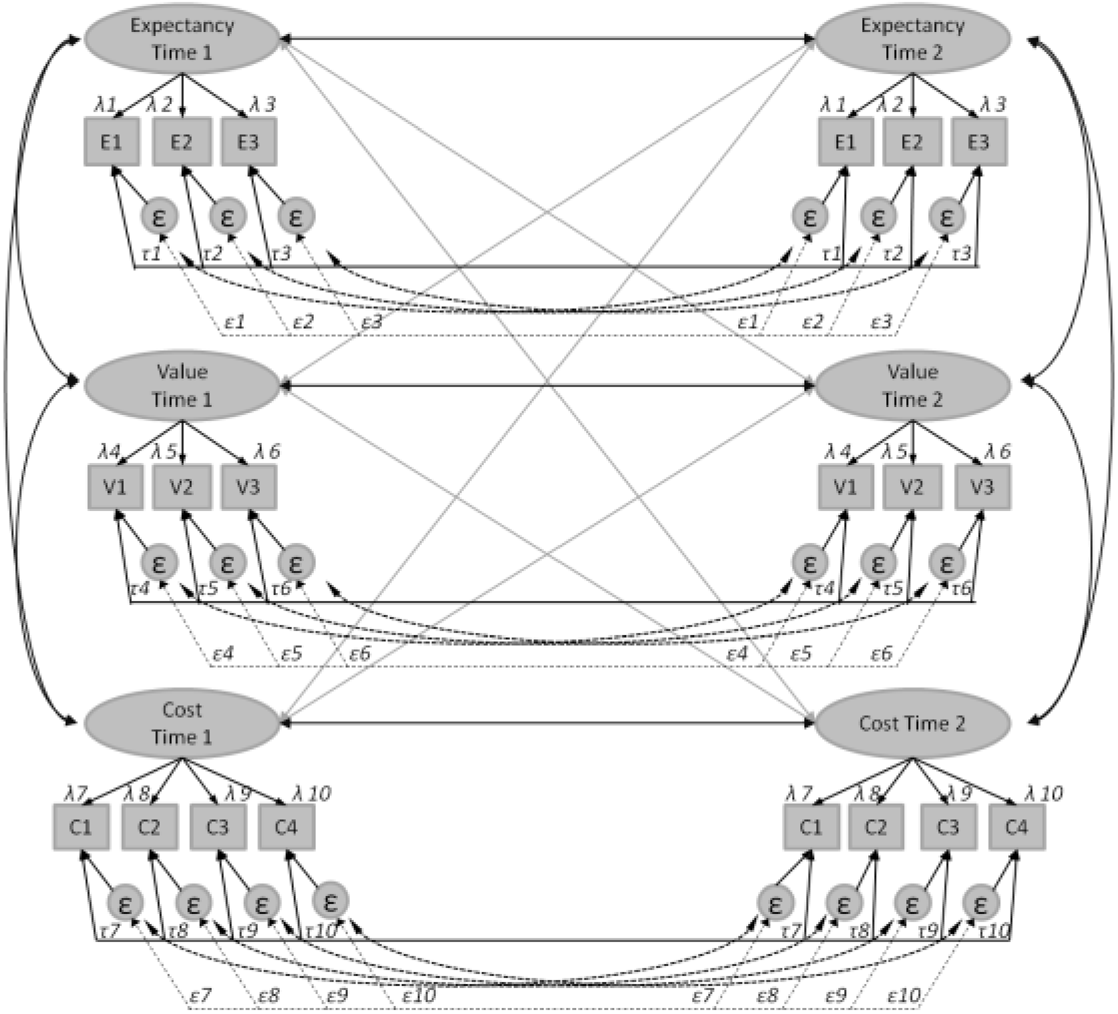

Figure 1. Dark, curved two headed arrows represent latent intercorrelations at the same time point. Dark, straight two-headed arrows represent stability coefficients (test-retest reliability). Light, straight two-headed arrows represent intercorrelations across time. τ scripts represent item intercepts which have been constrained to be equal across time. λ scripts represent factor pattern coefficients which have been constrained to be equal across time. ε scripts represent error variances which have been constrained to be equal across time. Curved, double-headed dashed arrows represent autocorrelations, which account for item dependency on different measurement occasions.