S-SOMADS: A NEW SURVEY TO MEASURE STUDENT ATTITUDES TOWARD DATA SCIENCE

April Kerby-Helm[1], Michael A. Posner[2], Alana Unfried[3], Douglas Whitaker[4], Marjorie E. Bond[5], Leyla Batakci[6], and Wendine Bolon[7]
Department of Mathematics and Statistics, [1]Winona State University, Winona, MN, USA
[2]Villanova University, Villanova, PA, USA
[3]California State University, Monterey Bay, Seaside, CA, USA
[4]Mount Saint Vincent University, Halifax, NS, Canada,
[5]The Pennsylvania State University, University Park, PA, USA,
[6]Elizabethtown College, Elizabethtown, PA, USA,
[7]RSM US, Davenport, IA, USA
michael.posner@villanova.edu

Attitudes play an important role in students' academic achievement and retention, yet we lack quality attitude measurement instruments in the new field of data science. This paper explains the process of creating Expectancy Value Theory-based instruments for introductory, college-level data science courses, including construct development, item creation, and refinement involving content experts. The family of instruments consist of surveys measuring student attitudes, instructor attitudes, and instructor and course characteristics. These instruments will enable data science education researchers to evaluate pedagogical innovations, create course assessments, and measure instructional effectiveness relating to student attitudes. We also present plans for pilot data collection and analyses to verify the categorization of items to constructs, as well as ways in which faculty who teach introductory data science courses can be involved.

INTRODUCTION

Data literacy for all citizens is increasingly vital in today's world, and data-savvy professionals are in high demand. The number of undergraduate data science programs has increased 84% over five years, from 37 in 2017 to 68 in 2022 (Swanstrom, n.d.). Introductory data science classes are offered in these institutions, as well as many others, to cater to student requests. The demand is so high that these institutions often struggle to keep up.

How do student attitudes, instructor attitudes, and the context of a data science class relate to student learning? Understandably, there is a lack of research in new fields such as data science. Educators are therefore constrained to assuming the relevance of research from fields closely related to data science, namely computer science and statistics. Across STEM programs, student attitudes are linked with student achievement and future career choices (Schunk, 1991; Simon et al., 2015). In computer science, Gurer et al. (2019) found that attitudes toward programming are associated with achievement in and perceived learning of computer science. Evans (2007) and Budé et al. (2007), found a significant correlation between negative attitudes and poor achievement in undergraduate introductory statistics courses, and the report *Connecting Research to Practice in a Culture of Assessment for Introductory College-level Statistics* (Pearl et al., 2012) identifies understanding attitudes and the outcomes that can be realized when students' attitudes are adequately measured as a research priority. Specific priorities include investigating (a) how attitudes contribute to success in learning; (b) how attitudes contribute to long term engagement; and (c) the important attitude constructs to measure about instructors and how these influence teaching practices and ultimately student outcomes. But do these relationships exist in data science?

Quality research on attitudes requires valid and reliable instruments that are grounded in theory. Computer science education researchers developed surveys of attitudes (Balik et al., 2003; Dorn & Tew, 2015) that have been well validated (Rachmatullah et al., 2020). Hoegh and Moskal (2009) focus on high-level perceptions of the discipline rather than knowledge in computing. In statistics, the Survey of Attitudes Toward Statistics (Schau, 1995) is a commonly used instrument. However, the instrument was not grounded in educational theory and therefore lacks validity evidence, has structural issues, and misses important constructs from Expectancy Value Theory (Whitaker et al., 2019). This critical omission means that it can be very difficult or impossible to connect student responses to either student learning or attitude constructs documented in the educational psychology literature. In data science, Liu

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org

and Huang (2017) and Dichev and Dicheva (2017) each used small, self-created sets of attitudinal questions that were not rigorously validated.

For readers not familiar with instrument design, an *instrument* is a survey, an *item* is an individual question asked on the survey, *negatively-coded items* are items where high numbers indicate a low score on the construct (e.g., "I hate data science" if the construct is liking data science) and are used for survey validity, and *constructs* are groups of similar questions that are grouped together to obtain a valid and consistent (usually mean) score.

The goal for this project is to develop six instruments assessing student and instructor attitudes toward data science and statistics and the related learning environment. Our hope is that the S-SOMADS (Student Survey of Motivational Attitudes toward Data Science) instrument will fill the void of well-designed instruments in data science attitudes that will empower teachers and researchers to understand and improve the teaching and learning of data science. The target audiences for all these instruments are introductory, college-level courses. This paper details the process of creating the S-SOMADS instrument by describing the theoretical model and two workshops of faculty and practitioners that informed our definition of data science, the scope of topics in an introductory data science class, and the item writing process for survey development. Details about the statistics instruments are previously published (Batakci et al., 2018; Whitaker, Unfried, & Batakci, 2018; Whitaker, Unfried, & Bond, 2019).

CONSTRUCTS AND DEFINITIONS

The theoretical framework used to develop the S-SOMADS is Expectancy Value Theory (EVT), a psychological theory of motivation that has been thoroughly researched and widely adopted in many disciplines (Eccles & Wigfield, 2020; Wigfield & Eccles, 2020). In EVT, one's achievement-related choices and performance are affected by Subjective Task Values [STVs] and what one expects to happen (expectancy), which mediate all other psychological components of the model (Wigfield & Eccles, 2020). EVT was chosen as the framework because STVs and expectancy have been shown to predict choices and performance on tasks. EVT has been used extensively in research with undergraduate students (Wigfield & Eccles, 2020) and has been researched in statistics education (Ramirez et al., 2012). Using methods such as confirmatory factor analysis, empirical results of the S-SOMAS, the statistics version of our student survey, further support the decision to use EVT with the analogous student statistics instrument (Unfried et al., 2021). Overall, there is both theoretical and empirical support for using EVT to model students' attitudes toward data science.

The development of S-SOMADS focused on measuring nine EVT constructs: Expectancy, four STVs (Utility Value, Interest/Enjoyment, Attainment Value, and Cost), Academic Self-Concept, Intrinsic Goal Orientation, Beliefs and Stereotypes about Data Science, and Perception of Difficulty. Brief definitions of these constructs, which are based on the EVT literature (Eccles & Wigfield, 2020) are given in Table 1. Although the EVT framework has many constructs to fully account for achievement motivation, the constructs displayed in Table 1 were chosen for inclusion on the S-SOMADS because they are theoretically more salient for predicting choice and performance and for measurement considerations. For example, the cultural milieu in which a student was raised is ultimately related to their data science choices and performance but is mediated through at least three other constructs and would be exceedingly difficult to appropriately measure on a survey with the intended use of the S-SOMADS; therefore, the development team opted not to measure this aspect of EVT.

Many of the EVT framework components are interrelated and distinguishing among them may not be possible with a particular instrument. On the statistics student survey (S-SOMAS), the Intrinsic Goal Orientation and Beliefs and Stereotypes constructs were not empirically distinguishable from the other constructs and were ultimately dropped from the instrument (Unfried et al., 2017).

ITEM DEVELOPMENT

We began by gathering experts in data science to discuss the instrument via two workshops in Summer 2021. These workshops brought together faculty from fields including computer science, statistics, education, and political science who taught introductory data science as well as a data scientist from industry who worked with undergraduate students. We met for a total of four days to talk about what data science is (and is not), identify content in an introductory data science class, and discuss specific items for S-SOMADS. We also categorized the types of students in an introductory data science class to consider different student perspectives, which we termed lenses.

Table 1. Construct definitions for SOMADS

Construct	Brief Definition
Expectancy	How an individual thinks they will perform in the field of data science
Utility Value	How much an individual values data science for serving or achieving their goals
Interest/Enjoyment	An individual's interest in data science, or their enjoyment from it
Attainment Value	How important data science is to an individual's identity
Cost	Factors that deter an individual from learning data science
Academic Self-Concept	Self-perception in academic settings (general and data science-specific)
Intrinsic Goal Orientation	Reasons for learning data science related to learning data science for its own sake or receiving favorable evaluations from others
Perceptions of Difficulty	How difficult an individual perceives data science to be
Beliefs / Stereotypes	An individual's beliefs and stereotypes about the field of data science

One challenge that arose was the lack of a universal definition of data science. Through review of curriculum guidelines from various professional organizations (American Statistical Association, 2014; Association for Computing Machinery Data Science Task Force, 2021, De Veaux et al., 2017; Gould et al., 2018; National Academies of Sciences, Engineering, and Medicine, 2018), we identified 72 competencies for data science, although clearly not all are intended for an introductory class. These competencies evolved into 60 topics that might be taught in an introductory data science course. Workshop participants determined the importance of each topic using these possibilities; required for, prerequisite to, optional for, or excluded from an introductory data science class. To determine the importance of a topic, nine students lenses were adopted by workshop participants: 1) a data science major, 2) a statistics major, 3) a computer science major, 4) a business major, 5) a mathematics major, 6) a student who dreads taking the course but who must take a data science course for their major whether their major is related to data science or not, 7) a student with some other major that has an interest/minor in data science, 8) someone who is mid-career or in a certificate program, or 9) a general data science viewpoint; each participant was assigned two lenses to use. These lenses represent the breadth of student backgrounds and career paths for students who might take an introductory data science course. We partitioned this into a general education course vs. a course designed for majors, ranking and discussing topics for inclusion in our content list for an introductory data science course. Since there is not a common definition of the term, data science, for the student survey, we chose to allow each student to answer questions using their own, personal definition of data science. The instructor and environment surveys will gather information on what material was taught in a class, so we will be able to determine the primary content being delivered.

The workshop participants also helped to write and review the initial item pool. We began with the 38 items on the Student Survey of Motivational Attitudes toward Statistics (S-SOMAS), replacing "statistics" with "data science." Each item was measured on a 7-point Likert-type scale of level of agreement (1=Strongly disagree, 7=Strongly agree), with some of the items written as negatively-worded items. We added items identified by workshop participants (written between the workshops), which was done through two student lenses with which the authors were most familiar. Once assembled, each item was rated as *keep*, *modify*, *drop*, or *move to another construct*. We reviewed the results in group discussions and identified an initial pool of items. For example, "I can use efficient methods and algorithms for processing data" was simplified to "I can write code to accomplish a data science task." Examples of items that were dropped were "Too much noise in the data makes me uncertain about my conclusions," which was more content mastery than attitudes, and "School is hard for me," which was too vague.

This pool of items was reviewed by seven subject matter experts (SMEs) from academia and industry in computer science, statistics, data science, and related fields such as business analytics. These individuals were identified based on authorship records and activity in discipline-based education

research. For each item, SMEs rated its necessity as essential; useful, but not essential; or not necessary, and provided qualitative feedback. Lawshe's (1975) Content Validity Ratio was calculated for each item and used to determine inclusion in the pilot survey. Pilot 1 of S-SOMADS began Spring 2022 and contains 87 items, such as "I enjoy using data to answer questions," "I know how to apply data science techniques," "Programming is easy for me," and "If I learn data science, I will look smart."

FUTURE WORK

The survey will be piloted in Fall 2022. Instructors from statistics, computer science, and other departments who teach introductory data science will be recruited to have their students participate in the pilot data collection. Psychometric analyses including exploratory factor analysis, confirmatory factor analysis, item response theory, and reliability/validity will be performed. These data will be used to determine if the items are aligning in the anticipated structure, and to identify items which are problematic and need further investigation. The Motivational Attitudes in Statistics and Data Science Education Research (MASDER) team will make the necessary changes based upon the pilot data collection to refine the S-SOMADS instrument. After the S-SOMADS is complete, we will begin work on instructor and environmental surveys (I-SOMADS and E-SOMADS, respectively).

Our end goal is to improve data science education. The grant that supports this work also supports the creation of a web portal to facilitate instructors registering themselves so that they and their students will receive the appropriate surveys. This portal will also provide reports to instructors on their students' performance, after stripping off identifying information. These reports will compare students' average scores to the distribution of scores from a nationally representative sample of classes, which the grant also supports obtaining. The website will also allow full access to de-identified data for researchers and provide research guides for how to best use the data from the project. Updates to this work can be found at http://sdsattitudes.com/.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation DUE-2013392. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

ACM Data Science Task Force. (2021). *Computing competencies for undergraduate data science curricula*. Association for Computing Machinery. https://doi.org/10.1145/3453538

American Statistical Association. (2014). *Curriculum guidelines for undergraduate programs in statistical science*. https://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

Balik, S., Nagappan, N., Williams, L., Petlick, J., Miller, C., Ferzli, M., & Wiebe, E. (2003). *Pair programming in introductory programming labs* [Paper presentation]. 2003 American Society for Engineering Education Annual Conference & Exposition. Nashville, TN. https://doi.org/10.18260/1-2--12489

Batakci, L., Bolon, W., & Bond, M. E. (2018). A framework and survey for measuring instructors' motivational attitudes toward statistics. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan*. ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_4J3.pdf

Budé, L., Van de Wiel, M. W. J., Imbos, T., Candel, M. J. J. M., Broers, N. J., & Berger, M. P. F. (2007). Students' achievements in a statistics course in relation to motivational aspects and study behaviour. *Statistics Education Research Journal, 6*(1), 5–21. https://doi.org/10.52041/serj.v6i1.491

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L.Z., Francis, A., Gould, R., Kim, A.Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., … , & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, *4*, 15–30. https://doi.org/10.1146/annurev-statistics-060116-053930

Dichev, C., & Dicheva, D. (2017). Towards data science literacy. *Procedia Computer Science*, *108*, 2151–2160. https://doi.org/10.1016/j.procs.2017.05.240

Dorn, B., & Tew, A. (2015). Empirical validation and application of the computing attitudes survey. *Computer Science Education*, *25*(1), 1–36. https://doi.org/10.1080/08993408.2015.1014142

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61,* Article 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Evans, B. (2007). Student attitudes, conceptions, and achievement in introductory undergraduate college statistics. *The Mathematics Educator*, *17*(2), 24–30.

Gould, R., Peck, R., Hanson, J., Horton, N., Kotz, B., Kubo, K., Malyn-Smith, J., Rudis, M., Thompson, B., Ward, M. D., & Wong, R. (2018). *The two-year college data science summit: A report on NSF DUE-1735199*. American Statistical Association. https://www.amstat.org/docs/default-source/amstat-documents/2018tycds-final-report.pdf

Gurer, M. D., Cetin, I., & Top, E. (2019). Factors affecting students' attitudes toward computer programming. *Informatics in Education*, *18*(2), 281–296. https://doi.org/10.15388/infedu.2019.13

Hoegh, A., & Moskal, B. M. (2009). Examining science and engineering students' attitudes toward computer science. In *Proceedings of the 39th IEEE Annual Frontiers in Education Conference: Imagining and engineering future CSET education* (pp. 1–6). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/FIE.2009.5350836

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Liu, M., & Huang, Y. (2017). The use of data science for education: The case of social-emotional learning. *Smart Learning Environments, 4*, 1. https://doi.org/10.1186/s40561-016-0040-4

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. National Academies Press. https://doi.org/10.17226/25104

Pearl, D. K., Garfield, J. B., delMas, R. C., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). *Connecting research to practice in a culture of assessment for introductory college statistics*. https://www.amstat.org/docs/default-source/amstat-documents/researchreport_dec_2012.pdf

Rachmatullah, A., Akram, B., Boulden, D., Mott, B., Boyer, K., Lester, J., & Wiebe, E. (2020). Development and validation of the middle grades computer science concept inventory (MG-CSCI) assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *16*(5), Article em1841. https://doi.org/10.29333/ejmste/116600

Ramirez, C., Schau, C., & Emmioğlu, E. (2012). The importance of attitudes in statistics education. *Statistics Education Research Journal*, *11*(2), 57–71. https://doi.org/10.52041/serj.v11i2.329

Schau, C., Stevens, J., Dauphinee, T. L., & Del Vecchio, A. (1995). The development and validation of the survey of attitudes toward statistics. *Educational and Psychological Measurement*, *55*(5), 868–875. https://doi.org/10.1177/0013164495055005022

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*(3–4), 207–231. https://doi.org/10.1080/00461520.1991.9653133

Simon, R. A., Aulls, M. W., Dedic, H., Hubbard, K., & Hall, N. C. (2015). Exploring student persistence in STEM programs: A motivational model. *Canadian Journal of Education*, *38*(1), 1–27. https://journals.sfu.ca/cje/index.php/cje-rce/article/view/1729

Swanstrom, R. (n.d.). *College & university degrees.* Retrieved October 12, 2017 and July 20, 2022 from http://datascience.community/colleges.

Unfried, A., Posner, M. A., Bond, M. E., Kerby-Helm, A., Bolon, W., Whitaker, D., & Batakci, L. (2021, August 12). *Why do we need yet ANOTHER instrument measuring student attitudes?* [Conference session]. Statistics, data, and the stories they tell. 2021 Joint Statistical Meetings Virtual Conference. https://douglaswhitaker.ca/talk/why-do-we-need-yet-another-instrument-measuring-student-attitudes/unfried-2021-jsm-slides.pdf

Whitaker, D., Unfried, A., & Batakci, L. (2018). A framework and survey for measuring students' motivational attitudes toward statistics. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan*. ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_C200.pdf

Whitaker, D., Unfried, A., & Bond, M. (2019). Design and validation arguments for the student survey of motivational attitudes towards statistics (S-SOMAS) instrument. In J. D. Bostic, E. E. Krupa, &

J. C. Shih (Eds.), *Assessment in mathematics education contexts: Theoretical frameworks and new directions* (pp. 120–146). Routledge. https://doi.org/10.4324/9780429486159

Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In A. J. Elliot (Ed.), *Advances in motivation science* (Vol. 7, pp.161–198). Academic Press. https://doi.org/10.1016/bs.adms.2019.05.002