International Society of the Learning Sciences

Who Are the Data Scientists in Education? An Investigation of the Identities and Work of Individuals in Diverse Roles

Joshua M. Rosenberg, University of Tennessee, Knoxville, jmrosenberg@utk.edu
Evelina Galas, University of Tennessee, Knoxville, cxx158@vols.utk.edu
K. Bret Staudt Willet, Michigan State University, staudtwi@msu.edu

Abstract: To understand how the new discipline of data science is impacting education and the learning sciences, we surveyed 68 individuals who identified as or worked in roles related to educational data science. Findings from an analysis of the surveys and follow-up interviews with key participants show that educational data scientists have more confidence with the mathematical and substantive aspects of their work (relative to programming) and that beliefs about what educational data science is has a bearing upon who identifies with this new domain.

Introduction

Across education, several new courses, degree programs, and jobs have been created that share a focus on applying novel research methods to new data sources. Such emergent data sources include students' interactions with digital technologies and teachers' participation in online professional learning communities. Although these courses and areas of work have different names (e.g., learning analytics, educational data mining, and educational data science) that reflect different assumptions about education, they also share a number of common features and can be broadly understood as applications of data science methods in education (Wise, 2020), or *data science in education* (Estrellado et al., 2020).

The role of data science in education has been established through foundational (Lee & Wilkerson, 2018) and forward-looking scholarship (Wise, 2020). In addition, efforts have been made to integrate disparate areas of research. However, some core questions pertaining to this new area of scholarship and work remain. Who identifies as an educational data scientist? Are educational data scientists similar to or different from learning analytics and educational data mining researchers? What do individuals across professional roles (e.g., researchers and analysts in educational organizations or school districts) in education who identify as educational data scientists do?

In this early-stage research, we developed and collected responses to a survey of educational data scientists. We then followed-up our survey with interviews of select respondents in order to learn about the identities and work of those individuals from diverse roles who identify as educational data scientists.

Method

We developed a short survey to understand (a) the backgrounds of educational data scientists, (b) their initial ideas regarding their identification with an educational data scientist role, and (c) the nature of their own work. To administer the survey, we shared the questionnaire directly with interested colleagues and asked influential individuals to reshare in their organizations and with their networks. This resulted in a total of 68 responses from individuals in a variety of professional roles.

In addition to questions on individuals' backgrounds, the survey also included questions on (a) whether individuals saw their work as relating to educational data science, (b) their confidence in three aspects of educational data science that have been used to define the domain (Rosenberg et al., 2020)—mathematics and statistics, computer science and programming, and knowledge about teaching, learning, and educational systems—and (c) what software they used in different contexts (e.g., work, in a degree program). The interview protocol we used with three individuals (with the plan to interview more) included 15 questions organized around (a) identification with educational data science, (b) relevant prior experiences, and (c) present work.

As presented in Table 1, we received responses from 21 faculty or post-secondary instructors, six graduate students, 14 data scientists or analysts, 11 K-12 teachers, and seven individuals from other roles, including non-University instructors of data science courses at the post-secondary level. After being presented with a definition of educational data science as an activity or work involving mathematics and statistics, computer science and programming, and knowledge of teaching, learning, and educational systems, most (82%) respondents reported doing educational data science (or related work); 10% of individuals were unsure.

Table 1: Participants' Professional Roles, Self-Reported Genders, and Identification as Under-Represented

Professional Roles	*n*	Does educational data science	Self-reported gender: Female	Self-reported gender: Male	Identification as under-represented
Data Scientist	18	17	5	9	2 (1)
Faculty	21	16	8	13	2 (2)
Graduate Student	6	5	3	3	1 (2)
Teacher	11	7	7	4	6
Other	7	2	5	2	1

Note. Parentheses following values in the identification as under-represented column denote individuals who preferred not to respond. Four individuals did not report a professional role and gender.

Findings

Individuals in our sample reported (using a 1-5 scale, indicating very little to very great confidence) more confidence with teaching, learning, and educational systems ($M = 4.23$, $SD = 0.85$) and mathematics and statistics ($M = 4.22$, $SD = 0.76$) than computer science and programming ($M = 3.55$, $SD = 1.01$). The most widely used software programs were Microsoft Excel, R, and Google Sheets, all three of which were used frequently in individuals' current work; these were followed by Python, SQL, and SPSS. R was used more than any other software in degree programs; Python was second most-frequently-used in this context.

Preliminary analysis of the three interviews revealed that whether individuals identified as educational data scientists—and, if they identified as educational data scientists, with what parts of the domain they identified—distinguished individuals more than the nature of their work or area of study. For example, one individual noted that they thought of educational data science as being focused on "big data" and machine learning. Although this interviewee often used computer programming in their work, they did not consider themselves to be programmers; if educational data science were defined more broadly, they might have been more inclined to identify with it. Another tentative finding pertained to the challenges of analyzing educational data due to its complexity and the individuals at multiple levels (i.e., teachers and students) involved with its production.

Discussion

Our aim with this study was to understand who identifies as an educational data scientist and what people from education-related roles who identify as educational data science do. Our results showed that even among individuals who reported doing educational data science, they were less confident with aspects related to computer science and programming. In addition, we found that a variety of data science tools were used. While R was widely used by individuals in degree programs, other common statistical software (see McNamara, 2018) were also used. Interviews suggested that identifying as an educational data scientist is distinct from doing educational data science (or being in a data science job or degree program), which highlights how both one's identification with a domain as well as recognition by others contribute to individuals identifying as educational data scientists.

References

Estrellado, R. A., Freer, E. A., Mostipak, J., Rosenberg, J. M., & Velásquez, I. C. (2020). *Data science in education using R*. London, England: Routledge.

Lee, V. R., & Wilkerson, M. (2018). *Data use by middle and secondary students in the digital age: A status report and future prospects.* Commissioned Paper for the National Academies of Sciences, Engineering, and Medicine, Board on Science Education, Committee on Science Investigations and Engineering Design for Grades 6-12. Washington, D.C.

McNamara, A. (2019). Key attributes of a modern statistical computing tool. *The American Statistician, 73*(4), 375-384.

Rosenberg, J. M., Lawson, M. A., Anderson, D. J., & Rutherford, T. (2020). Making data science count in and for education. In E. Romero-Hall (Ed.), *Research methods in learning design & technology* (pp. 94-110). Routledge: New York, NY.

Wise, A. F. (2020). Educating data scientists and data literate citizens for a new generation of data. *Journal of the Learning Sciences, 29*(1), 165-181.