Comparing item difficulty ratings and distributions of item mentions for pre-service secondary teacher ratings on the Self-Efficacy to Teach Statistics (SETS) Instrument

Comparando el grado de dificultad de ítems y la distribución de sus menciones en el instrumento Autoeficacia para Enseñar Estadística (SETS) de docentes en formación a nivel secundario

Rebecca L. Pierce[1], Leigh M. Harrell-Williams[2], Lawrence M. Lesser[3], Shelby G. Roberts[4] and M. Alejandra Sorto[5]

[1]Ball State University, [2]University of Memphis, [3]The University of Texas at El Paso, [4]University of Memphis, [5]Texas State University, USA

Abstract

In recent years, there has been an increased interest in grade 6-12 mathematics teacher efficacy to teach statistics in the US. However, knowing teachers' overall efficacy level may not be enough. In this study, 47 pre-service secondary mathematics teachers completed the Self-Efficacy to Teach Statistics – high school (SETS-HS) instrument. Additionally, the participants answered open-ended questions in which they identified the items that were easiest and most difficult to rate with higher efficacy, as well as explained the reasons for their ratings of these items. We explore the connection between the items identified by item difficulty estimates as hardest (or easiest) and the frequencies of the specific items those pre-service teachers identified most often in the open-ended questions.

Keywords: self-efficacy, statistics teaching efficacy, SETS-HS

Resumen

En los últimos años, ha habido un mayor interés en la eficacia de profesores de matemáticas de grado 6-12 para enseñar estadísticas en los Estados Unidos. Sin embargo, conocer el nivel general de eficacia de los docentes no es suficiente. En este estudio, 47 docentes de matemática en formación completaron el instrumento Autoeficacia para Enseñar la Estadística – nivel secundario (SETS-HS). Además, los participantes respondieron a preguntas abiertas en las que identificaron los ítems más fáciles y difíciles de valorar con mayor eficacia, así como explicaciones sobre las razones de su clasificación de estos ítems. Exploramos la conexión entre los ítems identificados por las estimaciones de la dificultad del ítem como más difíciles (o más fáciles) y las frecuencias de los ítems específicos que los profesores en formación identificaron más a menudo en las preguntas abiertas.

Palabras clave: autoeficacia, eficacia para enseñar estadística, SETS-HS

1. Introduction

Teaching efficacy, defined as teachers' beliefs about their ability to teach, impacts instructional activities in the classroom, including depth of coverage and use of technology, and teacher effectiveness (Ross, 1998; Wheatley, 2005; Zee & Koomen, 2016). Measures of teaching self-efficacy should be task-specific and focus on the teachers' judgments of capability of teaching, not the outcomes of teaching (Bandura, 2006). However, it's not just enough to know an overall level about how efficacious a pre-service teacher feels regarding the teaching of statistical tasks. It is important to determine the reasons why pre-service teachers rate these tasks in the way that they do and identify which topics the teachers feel most and least efficacious about teaching.

Pierce, R.L., Harrell-Williams, L.M., Lesser, L.M., Roberts, S.G. y Sorto, M.A. (2019). Comparing item difficulty ratings and distributions of item mentions for pre-service secondary teacher ratings on the self-efficacy to teach statistics (SETS) instrument. En J. M. Contreras, M. M. Gea, M. M. López-Martín y E. Molina-Portillo (Eds.), *Actas del Tercer Congreso Internacional Virtual de Educación Estadística.* Disponible en www.ugr.es/local/fqm126/civeest.html

The middle grades and high school Self-Efficacy to Teach Statistics (SETS) instruments were developed to advance the assessment field in mathematics and statistics education by providing measures of statistics teaching efficacy for grades 6–12 teachers. Several publications report on the development of the instruments and the psychometric properties of the middle grades and high school versions of the instruments (Harrell-Williams, Sorto, Pierce, Lesser & Murphy, 2014; Harrell-Williams, Lovett, Lee, Pierce, Lesser & Sorto, 2017; Harrell-Williams, Lovett, Pierce, Sorto, Lee & Lesser, 2017).

In particular, Harrell-Williams et al. (2015) rank ordered Rasch model-based item difficulty estimates of the items from the Self-Efficacy to Teach Statistics - Middle Grades (SETS-MS) instrument to determine which tasks pre-service teachers were most and least confident about teaching their future students. This analysis raised questions: If pre-service teachers were asked to focus on one to two specific items from each subscale, which particular items do they identify as the most and least confident to teach and why? Hence, a set of open-ended questions were created to supplement the Likert items on the SETS instruments.

This study seeks to compare classical test theory estimates of difficulty (i.e., item means) with the frequency with which items are mentioned in the free response questions in order to make connections between the Likert ratings and open-ended questions. Data was collected at two public institutions of higher education in the US using the high school version of the Self-Efficacy to Teach Statistics (SETS-HS) instrument. The SETS-HS measures pre-service teachers' self-efficacy to teach 44 specific statistical tasks that span all three levels of the GAISE PreK-12 Report (GAISE) (Franklin et al., 2007) as well as the two high school data analysis strands of the Common Core State Standards for Mathematics (CCSSM; National Governors Association and Council of Chief State School Officers, 2010). Additionally, open-ended questions ask the respondent to identify the items that were easiest and most difficult to rate with higher efficacy and explain the reasons for their ratings of these items. Specifically, we try to answer if knowing only the responses to the open-ended questions provides the same information as the responses to the Likert scale questions. That is, would we be painting the same picture no matter which responses we considered to enlighten professional development?

2. Methods

2.1. Participants

At two authors' home institutions, pre-service secondary mathematics students were asked by their course instructor (either a statistics or mathematics/statistics pedagogy course) to complete a paper version of the high school version of the Self-Efficacy to Teach Statistics instrument during a class session at the end of the semester. One four-year university is located in the south and is classified as a Higher Research Activity institution by the Carnegie Classification system. The university ranks in the top 15 for awarding bachelor's degree to Hispanic students and has a diverse student body with 52% identifying as minority students. The other institution is located in the Mid-West and is also classified as a Higher Research Activity institution by the Carnegie Classification system. The university serves a small percentage of minority students (13 – 15%).

A total of 47 pre-service secondary mathematics teachers across the two universities participated in the study. About three quarters of the participants were female (72.3%).

Participants identified their ethnicity as caucasian (74.5%), hispanic (14.9%) or black (2.1%). Only 6.4% of the participants identified themselves as non-native English speakers. Of the 24 students who responded to the demographic question regarding their credentialing status, 87.5% had preliminary credentials, 8.3% had full credentials and 4.2% had intern credentials. While the percentages of female and Caucasian pre-service teachers may seem high, they follow national trends (Taie & Goldring, 2017).

2.2. SETS – High School Instrument

The High School version of the Self-Efficacy to Teach Statistics instrument (SETS-HS) contains 44 items based on a common stem ("Rate your confidence in teaching high school students the skills necessary to successfully complete the task …"). Respondents use a 6-point rating scale (1 = "Not at all confident", 2 = "Only a little confident", 3 = "Somewhat confident", 4 = "Confident", 5 = "Very confident", 6 = "Completely confident") to indicate their confidence to teach the task listed in each item, such as creating a histogram for summarizing data or identifying and interpreting a slope and intercept for a line. The first 26 items of the SETS-HS version are identical to those in the Middle Grades version of the SETS instrument (Harrell-Williams et al., 2014), which were based on Level A and B of the GAISE (Franklin et al., 2007). The last 18 items are drawn from Level C of the GAISE and the "interpreting categorical & quantitative data" and "making inferences & justifying conclusions" strands of the high school Common Core State Standards for Mathematical Practice (National Governors Association, 2010). Using the language of Friel, Curcio, and Bright (2001), these subscales can be referred to as reading the data (Level A), reading between the data (Level B), and Reading Beyond the Data (Level C). The task items are provided in appendix.

The task items on the SETS-HS are divided to provide three subscale scores corresponding to Levels A, B, and C of the PreK-12 GAISE. Factor analysis supported the use of three subscale scores (Harrell-Williams, Lovett, Lee, et al., 2017). The estimated reliability coefficient for the High School total score was 0.984. The subscale scores' reliabilities were all above the 0.80 threshold, with the Level A subscale score having a reliability of 0.841, Level B was 0.964, and Level C was 0.969.

To assess the source of variation in self-efficacy, the research team designed a set of open-ended questions which are embedded at the end of each subscale in the instrument. The questions ask the participants to reflect on their responses to specific items in each section of the SETS-HS instrument and explain why they feel that way about teaching the statistical concepts in those items. Figure 1 shows an example of an open-ended question for items related to GAISE Level A. Parallel items were written to correspond to Level B and to Level C.

2.3. Analysis

Psychometric evidence for the SETS instruments were mostly based on analyses in a multidimensional Rasch framework (Harrell-Williams et al., 2014; Harrell-Williams et al., 2015; Harrell-Williams, Lovett, Lee, et al., 2017; Harrell-Williams, Lovett, Pierce, et al., 2017). However, due to the smaller sample size of the data used in this study, we employed classical test theory techniques for the item analyses for the Likert items. Specifically, we used the item mean and standard deviation output from the SCALE menu options in SPSS v25 (IBM, 2017). The means for each item act as an indicator of

item difficulty. Non-zero standard deviations indicate that the participants are not responding all in the same way, which contributes to the instrument's discrimination.

Open-Ended Question A:
Please review your responses to **items 1 – 11.**

a) Looking at the one or two items from **items 1 – 11** that you indicated feeling **LEAST** confident about teaching high school students, think about the reason(s) you feel this way. Use the space below (and the back of this paper, if necessary) to explain your reason(s), identifying which reason goes with which item number. If you have more than one reason, please explain as many as you can.

b) Looking at the one or two items from **items 1 – 11** that you indicated feeling **MOST** confident about teaching high school students, think about the reason(s) you feel this way. Use the space below (and the back of this paper, if necessary) to explain your reason(s), identifying which reason

Figure 1. Example of the open-ended questions

For the analysis of the open-ended SETS items, the item that each student specifically mentioned (by item number or wording) was recorded. The analysis involved creating a frequency distribution for each item and identifying if it was mentioned as the hardest item or the easiest item for a pre-service teacher to endorse with a response "completely confident".

3. Results

3.1. Item Analysis

Table 1 presents the means and standard deviations for each item. The SETS items use a 6-point Likert scale, with higher categories indicating more teaching efficacy. Items that are more difficult to answer as "completely confident" have lower means. In most cases, the means for items on subscale A are higher than the means for the items on subscales B and C. The means for items on subscale A range from 4.51 to 5.09. The subscale B items have means that range from 3.40 to 5.13, with twelve out of fifteen items having means that are less than the lowest item mean in subscale A. Items on subscale C have means that range from 3.11 to 4.85, with only one item having a mean greater than the lowest mean in subscale A. The patterns of the means follow the intended design of the instrument with items on subscale A representing the pre-requisite knowledge and skills for items on subscales B and C. The items in subscales B and C that have means greater than the lowest subscale A item mean (4.51) cover such topics as using histograms to compare groups, discussing the representativeness of a sample, and identifying the slope and intercept. These items represent easier topics in statistics coursework than other items in subscale B and C, such as sampling variability (from B) or assessing model fit using residuals (from C).

Item standard deviations range between 0.86 and 1.62. Variability in item responses indicates that the items are discriminating between those with different levels of self-efficacy. Subscale C has the smallest range of standard deviations from 1.17 to 1.56. Subscale A has standard deviations from 0.86 to 1.42, while Subscale B has standard

deviations from 0.86 to 1.62. This may indicate more agreement in the ratings for subscale C items than for the items on the other two subscales.

3.2. Triangulation of item means with item mentions in open-ended questions

The open-ended questions asked the respondent to think only about the items in the last section of Likert-scale items. This makes the responses relative to the other items in that subscale instead of the instrument as a whole. Thus, results of the frequencies of mention and the agreement with item analysis are presented by subscale.

There was considerable variability in how all of the items were viewed by pre-service teachers as indicated in the open-ended responses for subscale A. All 11 items on the subscale were identified as both the item that the pre-service teachers found to be the easiest item and the one they found to be the hardest item to respond "completely confident" (see Figure 1 and Table 1). Items 1 and 5 were mentioned with the highest frequency as the "least confident" items (17.1% each) and most frequently across either of the two open-ended items (9 times and 10 times, respectively). Item 2 was mentioned most frequently as the "most confident" (11.4%). These results do not follow the trend exhibited in the item analysis results. Item 1 had the third lowest item mean (one of the least confident/most difficult item) and item 5 had the fourth highest item mean. The mean for item 5 was slightly lower than item 2, confirming that it was rated with lower confidence than item 2.

There was more agreement between the item mean results and the frequency of mention in the open-ended responses for subscale B. Seven of the 15 items (approximately 47%) on subscale B were identified in the responses only for the open-ended question about "least confident" items. Items 23 and 12 were mentioned most frequently in that question (17.3% and 13.5%, respectively) and were never mentioned in the "most confident" question. These two items were also the items with the lowest item means (3.47 and 3.40, respectively). Item 15, the item mentioned most in the "most confident" question (25.6%) and the item with the highest item mean (5.13), was the only item that was not mentioned in the question about the "least confident' item. Item 16 was the item mentioned with the second-highest frequency regarding the "most confident" item (20.5%) and the item with the second-highest mean. However, item 16 was also mentioned in 3.8% of responses to the question about "least confident" items. Almost half of the 15 items were not mentioned at all in the "most confident" questions.

For subscale C, there was a clear distinction about the "least" and "most confident" items. The 3 items with the highest items means were never mentioned in the "least confident" question and the 7 items with the lowest item means were never mentioned in the "most confident" question. Item 32 was mentioned most in the "least confident" question (18.2%), and it had the lowest item mean (3.11). Item 35 was mentioned the most in the "most confident" question (25.0%) and had the highest item mean (4.85) on this subscale. Items 27 and 38 were mentioned 14.3% and 10.7% of the time, and had the next two highest item means (4.52 and 4.37, respectively).

Table 1. Item Statistics for Likert Scale Responses and Relative Frequencies of Item Mentions in Open-Ended Reponses within Each Subscale

Subscale	Item number	Item mean	Item standard deviation	Mentions within least confident (within subscale)	Mentions within most confident (within subscale)
A	1	4.62	1.24	6 (17.1%)	3 (6.8%)
	2	4.93	1.10	3 (8.6%)	5 (11.4%)
	3	4.66	1.07	1 (2.9%)	4 (9.1%)
	4	4.89	1.31	3 (8.6%)	4 (9.1%)
	5	4.81	1.31	6 (17.1%)	4 (9.1%)
	6	4.66	1.42	2 (5.7%)	4 (9.1%)
	7	4.81	1.31	4 (11.4%)	4 (9.1%)
	8	4.51	1.12	2 (5.7%)	4 (9.1%)
	9	4.51	0.93	3 (8.6%)	4 (9.1%)
	10	5.09	0.86	1 (2.9%)	4 (9.1%)
	11	4.70	1.02	4 (11.4%)	4 (9.1%)
B	12	3.40	1.31	7 (13.5%)	NM
	13	4.15	0.93	2 (3.8%)	1 (2.6%)
	14	3.74	1.21	3 (5.8%)	NM
	15	5.13	0.88	NM	10 (25.6%)
	16	5.04	0.86	2 (3.8%)	8 (20.5%)
	17	4.36	1.61	3 (5.8%)	6 (15.4%)
	18	4.23	1.62	3 (5.8%)	7 (17.9%)
	19	4.02	1.09	3 (5.8%)	NM
	20	3.67	1.37	6 (11.5%)	NM
	21	3.98	1.42	3 (5.8%)	2 (5.1%)
	22	4.74	0.99	1 (1.9%)	2 (5.1%)
	23	3.47	1.30	9 (17.3%)	NM
	24	4.64	1.13	1 (1.9%)	3 (7.7%)
	25	3.96	1.32	6 (11.5%)	NM
	26	4.11	1.18	3 (5.8%)	NM
C	27	4.52	1.19	NM	4 (14.3%)
	28	4.17	1.32	1 (3.0%)	3 (10.7%)
	29	4.15	1.26	1 (3.0%)	2 (7.1%)
	30	3.54	1.46	4 (12.1%)	NM
	31	3.35	1.35	3 (9.1%)	NM
	32	3.11	1.29	6 (18.2%)	NM
	33	3.96	1.43	1 (3.0%)	NM
	34	3.37	1.37	4 (12.1%)	NM
	35	4.85	1.21	NM	7 (25.0%)
	36	3.77	1.56	3 (9.1%)	NM
	37	3.70	1.23	2 (6.1%)	3 (10.7%)
	38	4.37	1.24	NM	3 (10.7%)
	39	4.28	1.17	1 (3.0%)	2 (7.1%)
	40	4.04	1.29	NM	2 (7.1%)
	41	4.21	1.28	1 (3.0%)	2 (7.1%)
	42	3.40	1.33	4 (12.1%)	NM
	43	3.57	1.23	1 (3.0%)	NM
	44	3.34	1.45	1 (3.0%)	NM

Note: NM = Not mentioned in question responses.

4. Final conclusions and comments

From this small-scale study, it appears there may be additional information to be gained by including the open-ended questions when using the SETS instruments**.** For items on the Level A subscale, some of the items indicated to be associated with high efficacy according to item means received several mentions in the open-ended questions as the topics that teachers felt least confident about teaching. This indicates while the item means for Level A tended to be higher than most of those in Level C, attention in teacher preparation or professional development should not just focus on topics covered by Level B and C items or only focus on the topics in items with the lowest mean. As observed, the information provided by the forced choice of items that teachers associated with high and low levels of efficacy may be different and/or as informative as the information obtained from the Likert scale questions. At the very least, these results indicate a larger scale study regarding the format of how statistics efficacy data is collected may be warranted.

Appendix: SETS-HS Items

Level A subscale - Reading the data

1. Collect data to answer a posed statistical question in contexts of interest to high school students.
2. Recognize that there will be natural variability between observations for individuals.
3. Select appropriate graphical displays and numerical summaries to compare individuals to each other and an individual to a group.
4. Create dotplot, stem and leaf plot, and tables (using counts) for *summarizing* distributions.
5. Use dotplot, stem and leaf plot, and tables (using counts) for *describing* distributions.
6. Create boxplots for *summarizing* distributions.
7. Use boxplots, median, and range for *describing* distributions.
8. Identify the association between two variables from scatterplots.
9. Generalize a statistical result from a small group to a larger group such as the whole class.
10. Recognize that statistical results may be different in another class or group.
11. Recognize the limitation of making inference (i.e. generalization) from a classroom dataset to any population beyond the classroom.

Level B subscale - Reading between the data

12. Distinguish between a question based on data that vary and a question based on a deterministic model (for example, specific values of rate and time determines a particular value for distance in the model $d = r \times t$).
13. Identify what variables to measure and how to measure them in order to address the question posed.
14. Describe numerically the variability between individuals within the same group.
15. Create histograms for summarizing distributions.
16. Use histograms for comparing distributions.
17. Compute interquartile range and five-number summaries for summarizing distributions.

18. Use interquartile range, five-number summaries, and boxplots for comparing distributions.
19. Recognize the role of sampling error when making conclusions based on a random sample taken from a population.
20. Describe numerically the strength of association between two variables using linear models.
21. Explain the differences between two or more groups with respect to center, spread (for example, variability), and shape.
22. Recognize that a sample may or may not be representative of a larger population.
23. Interpret measures of association.
24. Distinguish between an observational study and a designed experiment.
25. Distinguish between "association" and "cause and effect."
26. Recognize sampling variability in summary statistics such as the sample mean and the sample proportion.

Level C subscale - Reading beyond the data

27. Describe characteristics of a normal distribution, such [as] general shape of distribution, symmetry, how standard deviation influences shape, and area under the curve.
28. Estimate percentages via the empirical rule (i.e., percentage of observations within 1, 2, or 3 standard deviations from the mean) using the mean and standard deviation of a dataset which has an approximately bell-shaped distribution.
29. Estimate a specified area under the normal curve using technology or a statistical table.
30. Summarize categorical data using two-way tables (i.e., contingency tables, frequency tables).
31. Calculate and interpret relative frequencies using two-way tables (i.e., contingency tables, frequency tables).
32. Find conditional and marginal frequencies from two-way tables (i.e., contingency tables, frequency tables).
33. Fit an appropriate model (e.g., linear, quadratic, or exponential) using technology for a scatterplot of two quantitative variables.
34. Assess the fit of a particular model informally by plotting and analyzing its residuals.
35. Identify the slope and y-intercept coefficients of a linear model and interpret them in the context of the data.
36. Calculate, using technology, the correlation coefficient between two quantitative variables.
37. Evaluate whether a specified model is consistent with data generated from a simulation.
38. Explain the role of randomization in surveys, experiments and observational studies.
39. Describing purposes and differences among surveys, experiments, and observational studies.
40. Evaluate how well the conclusions of a study are supported by the study design and the data collected.
41. Estimate a population mean or proportion using data from a sample survey.
42. Develop a margin of error for an estimate of a population mean or proportion using simulation models.

43. Compare two treatments from a randomized experiment by exploring numerical and graphical summaries of the data.
44. Determine if the difference between two population means or proportions is statistically significant using simulations

References

Bandura, A. (2006). Guide for constructing self-efficacy scales. In T. U. F. Pajares (Ed.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 307-337). Greenwich, CT: Information Age Publishing.

Finney, S. J., & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology*, *28*(2), 161-186. doi: 10.1016/S0361-476X(02)00015-2

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-K-12 curriculum framework*. Alexandria, VA: American Statistical Association. Available from: http://www.amstat.org/education/gaise/GAISEPreK-12_Full.pdf

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124-158. doi: 10.2307/749671.

Harrell-Williams, L. M., Lovett, J. N., Lee, H. S., Pierce, R. L., Lesser, L. M., & Sorto, M. A. (2017). Validation of scores from the high school version of the Self-Efficacy to Teach Statistics (SETS-HS) instrument using pre-service mathematics teachers. *Journal of Psychoeducational Assessment*. Advanced online publication. doi: 10.1177/0734282917735151.

Harrell-Williams, L. M., Lovett, J. N., Pierce, R. L., Sorto, M. A., Lee, H. S., & Lesser, L. M. (2017). The middle grades SETS instrument: Psychometric comparison of middle and high school pre-service mathematics teachers. In E. Galindo & J. Newton (Eds.), *Proceedings of the 39th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1064-1067). Indianapolis, IN: Hoosier Association of Mathematics Teacher Educators. Available from: http://www.pmena.org/pmenaproceedings/PMENA%2039%202017%20Proceedings.pdf

Harrell, L. M., Pierce, R. L., Sorto, A., Murphy, T. J., & Lesser, L. (2009). On the importance and measurement of pre-service teachers' efficacy to teach statistics: Results and lessons learned from the development and testing of a GAISE-based instrument. In *Proceedings of the 2009 Joint Statistical Meetings, Section on Statistical Education* (pp. 3396-3403). Alexandria, VA: American Statistical Association. Available from: http://www.statlit.org/pdf/2009HarrellEtAlASA.pdf

Harrell-Williams, L. M., Sorto, M. A., Pierce, R. L., Lesser, L. M., & Murphy, T. J. (2014). Validation of scores from a measure of teachers' self-efficacy to teach middle grades statistics. *Journal of Psychoeducational Assessment* [online], *32*(1), 40-50.

Harrell-Williams, L. M., Sorto, M. A., Pierce, R. L., Lesser, L. M., & Murphy, T. J. (2015). Identifying statistical concepts associated with high and low levels of self-efficacy to teach statistics to middle grades. *Journal of Statistics Education*, *23*(1),

1-20. Available from: http://www.amstat.org/publications/jse/v23n1/harrell-williams.pdf

IBM Corp. (2017). *IBM SPSS Statistics for Windows, Version 25.0.* Armonk, NY: IBM Corp.

National Governors Association and Council of Chief State School Officers (2010). *Common Core State Standards for Mathematics.* Washington, D.C.: Authors.

Ross, J. A. (1998). The antecedents and consequences of teacher efficacy. In J. Brophy (Ed.), *Advances in research on teaching* (pp. 49-74). Greenwich, England: JAI Press.

Taie, S., & Goldring, R. (2017). *Characteristics of public elementary and secondary school teachers in the United States: Results from the 2015–16 national teacher and principal survey first look* (NCES 2017-072). Washington, D.C.: National Center for Education Statistics.

Wheatley, K. F. (2005). The case for reconceptualizing teacher efficacy research. *Teaching and Teacher Education, 21*(7), 747-766. doi:10.1016/j.tate.2005.05.009.

Zee, M., & Koomen, H. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research, 86*(4), 981-1015. doi: 10.3102/0034654315626801.