Validating a scale that measures scientists' self-efficacy for public engagement with science

Jane Robertson Evia[a]*, Karen Peterman[b], Emily Cloyd[c], and John Besley[d]

[a]*Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA, 540-231-1131, jane.robertsonevia@vt.edu, ORCid: 0000-0003-4196-9814*

[b]*Karen Peterman Consulting, Co., Durham, NC 27707, USA, 919-627-2676, karenpetermanphd@gmail.com, ORCid: 0000-0003-4388-9412*

[c]*American Association for the Advancement of Science, Washington, DC 20005, USA, 202-326-6506, ecloyd@aaas.org, ORCid: 0000-0003-1277-1702*

[d] *Department of Advertising and Public Relations, Michigan State University, East Lansing, MI 48824, USA, 517-884-4411, jbesley@msu.edu, ORCid: 0000-0002-8778-4973*

* Corresponding author: Jane Robertson Evia, Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA, 540-231-1131, jane.robertsonevia@vt.edu

Validating a scale that measures scientists' self-efficacy for public engagement with science

Self-efficacy, or the beliefs people hold about their ability to succeed in certain pursuits, is a long-established construct. Self-efficacy for science communication distinguishes scientists who engage with the public and relates to scientists' attitudes about the public. As such, self-efficacy for public engagement has the potential to serve as a key indicator in the evaluation of scientist training and public outreach programs. To date, most evaluation scales have been designed for public audiences, rather than scientists. This study used think-aloud methods and Item Response Theory to develop a scale to measure scientists' Self-Efficacy for Public Engagement with Science. The results from this study support the use of a 13-item self-efficacy scale, and provide initial validation evidence to support its use with scientists who engage with the public. The findings are presented in relation to the continued study of public engagement through both research and evaluation.

Keywords: public engagement; science communication, Rasch process

Introduction

Self-efficacy is an important phenomenon to measure and understand for myriad reasons: increased self-efficacy leads to better performance (Bandura, 1997; Pajares, 1996; Pajares & Urdan, 2006), motivates initial and persistent engagement in various activities (Dudo, 2012; Poliakoff & Webb, 2007), and provides a sense of high personal satisfaction (Chemers, Li-tze, & Garcia, 2001). Within the context of science communication, self-efficacy is also a primary motivator among scientists who participate in public engagement with science (PES) activities (Dudo & Besley, 2016). Perhaps not surprisingly, increased self-efficacy is a common goal of many science communication training programs and related evaluation efforts. Currently, there are no scales that measure the self-efficacy of scientists for PES activities, which limits the field's ability to study and evaluate the influence of this construct within the context of public engagement. A validated measure of this phenomenon has the potential to enhance our understanding of effective science communication training and the impact of public engagement on scientists and the public alike.

Literature Review

Self-efficacy, or the beliefs people hold about their ability to succeed in certain pursuits, is a long-established phenomenon that has been studied through various lenses and in a multitude of settings. Based on his social cognitive theory (Bandura, 1986), Bandura (1997) defined self-efficacy as, 'people's beliefs in their capabilities to produce desired effects by their actions' (p. vii). A tenet of social cognitive theory is that self-reflection allows individuals to assess their knowledge, experiences, and thoughts as a means of determining their likelihood of success. Bandura (1977) maintained that, through self-reflection, people appropriately alter their thoughts, behaviors, and motivation to engage in or continue with various undertakings. The construct of self-efficacy has been studied in many contexts to confirm the relation between ideas about self, behavior, and success. It has been explored in a general academic sense, at the discipline-specific level, and in relation to specific tasks. Increased self-efficacy has a positive impact on achievement (Bandura, 1997; Pajares, 1996; Pajares & Urdan, 2006) and on the motivation to engage in and/or persist with various activities (Dudo, 2012; Poliakoff & Webb, 2007). The positive outcomes related to increased self-efficacy beliefs hold true for diverse and international groups (Hsu, Ju, Meng, & Chang, 2007; Morony, Kleitman, Lee, & Stankov, 2013; Skaalvik, Federici, & Klassen, 2015; Wirawan & Bandu, 2016).

These studies span multiple academic contexts and audiences. Pajares and Urdan (2006), for example, explored general academic self-efficacy among adolescent students. They found that academic achievement is positively related to academic self-efficacy (i.e., the belief that one has their ability to do well in school). Chemers and colleagues (2001) examined the academic self-efficacy of first-year university students, and found that general academic self-efficacy is strongly, and positively related to academic achievement, adjustment to college, and both personal satisfaction with and persistence in school.

Discipline-specific examples also exist, confirming an association between self-efficacy and both academic and professional behavior. Science self-efficacy predicts whether students choose to take part in science-related activities, their engagement with those activities, and their science achievement (Britner & Pajares, 2001; Pajares, Britner, & Valiante, 2000; Zeldin & Pajares, 2000). Finney and Schraw (2003) found that students with a higher sense of discipline-specific self-efficacy in statistics tended to perform better when solving specific statistics problems and that they attained higher academic achievement at the overall course level than students with lower statistics self-efficacy. Skaalvik, Federici, and Klassen (2015) found that self-efficacy beliefs have a positive relationship with mathematics achievement for Norwegian middle school students, replicating research involving middle school math students in the U.S (Fast, et al., 2010; Pajares & Miller, 1997; Schunk, 1989). Anderson et al. (2016) developed an instrument to measure self-efficacy to communicate science with other scientists. They found that the self-efficacy of early career biomedical scientists toward writing, presenting, and conversing about science were all significantly and positively related to specific tasks, such as preparing manuscripts and making oral presentations at national meetings. Self-efficacy for science communication has also been related to scientists' level of engagement with the public. Dudo (2012) found that biomedical students with higher self-efficacy to communicate science were more likely to engage with the public about science than those with lower communication self-efficacy. Dudo and Besley

(2016) expanded on these findings with a wider pool of Ph.D. scientists from across the U.S. Their study revealed that scientists with high self-efficacy for engaging with the public tend to engage more frequently and felt more positive about these public interactions than those with lower self-efficacy.

Dudo (2012) posits that one method of helping to increase scientists' self-efficacy to engage with the public is science communication training. He notes that scientists who have high communication self-efficacy feel empowered to engage with the public, and that those who have been trained tend to engage in more public science events than those who have not. Indeed, a number of programs have been created with the intent of improving scientists' self-efficacy for communicating with the public. One of the most extensive is Portal to the Public (Pacific Science Center, 2017), a project that has trained scientists to communicate with the public in their local communities since 2007, providing a scaleable model for integrating scientists into museum-based learning environments for the purposes of public engagement (Selvakumar & Storksdieck, 2013). The Alan Alda Center for Communicating Science (n.d.) was established in 2009 to improve communication training related to both science and medicine. The Beacons for Public Engagement project was initiated in 2008 and resulted in the creation of the National Co-ordinating Centre for Public Engagement (NCCPE) (n.d.) to increase capacity for public engagement. NCCPE offers their own scientist training, and feature public engagement training programs for 14 additional organizations across England. One of the most recent additions is the American Association for the Advancement of Science's (AAAS) Leshner Leadership Institute (LLI) for Public Engagement with Science (n.d.), which was founded in 2015 to train mid-career scientists in an effort to bolster public engagement at their academic institutions.

The LLI is guided by a logic model and theory of change that underscore the importance of public engagement, and the kinds of outcomes that can be expected for both scientists and publics who are part of public engagement with science (PES) activities (AAAS, 2016). Scientists' self-efficacy is one such outcome that should develop as scientists gain engagement experience, both through training and practice. The current study was conducted as a first step toward developing a series of scales that could be used to document the short-term outcomes from the logic model for scientists who participate in PES activities. It was the hope that these scales could serve as common measures for the field, with the potential to enhance understanding of the impact of the LLI, in particular, and public engagement more broadly.

The LLI exemplifies a recent and explicit focus within science communication training on the need to shift the 'deficit model' that has motivated some scientists to conduct public outreach in the past. Deficit model thinking assumes that the public is ignorant when it comes to science, that this ignorance results in negative attitudes toward science, and that one-way communication from scientists to the public has the potential to improve public knowledge and attitudes (Ahteensuu, 2011; Miller 2001). This is a problem, as substantial research suggests that there is little evidence to support that increasing knowledge will have any impact on attitudes toward science or support for science (Allum, Sturgis, Tabourazi, & Brunton-Smith, 2008; NASEM 2016a; NASEM 2016b). Research has nevertheless demonstrated that deficit-model thinking has been and continues to be prevalent within the scientific community, both among those who choose to interact with the public and those who do not (Besley, Oh, & Nisbet, 2013; Cortassa, 2016; Horst, 2013; Sims, Madden, Cacciatore, & Yeo, 2016). There is strong support in the contemporary literature to step away from this model in an effort to move toward a reciprocal relationship between scientists and the public

(Cortassa, 2016; Mayhew and Hall, 2012). The expectation is that improved dialogue between scientists and their communities may lead to non-science literacy outcomes related to both positive attitudes toward science and scientists (Besley et al. 2016; Dudo and Besley 2017). While the evidence is limited, Mayhew and Hall (2012), for example, offer that a reciprocal exchange between scientists and non-scientists is imperative in order to create interest and buy-in from the public, and to demonstrate the personal impact of scientific research results. They go onto say that effective science communication requires full engagement from both parties, a skill not innate to many scientists.

As noted previously, self-reflection is a necessary step in developing one's sense of self-efficacy (Bandura, 1977). A key component within this process is what Bandura (1995) calls mastery experience, or historical experiences that he believes are the primary influence upon one's self-efficacy. These mastery experiences are crucial for scientists to develop a good understanding of their PES self-efficacy. Scientists need experience sharing and discussing research with non-scientists to begin to understand public thought and, in turn, to develop an understanding of their own PES self-efficacy. In the absence of this reciprocal exchange of information and ideas it is unlikely that scientists would fully conceptualize their self-efficacy to communicate science. Tools to support post-engagement reflection may be helpful in this process. Such tools can help both those charged with training and evaluating scientists' PES efforts, as well as helping scientists personally reflect upon their efforts. A validated measure of self-efficacy for PES therefore has the potential to play a critical role in this self-reflection process. Validated items have the potential to reinforce the use of best practices within the field, and to hone scientists' reflection around those topics. Even so, to date there are no known scales that measure scientist self-efficacy within the context of public engagement.

Purpose of the Current Study

The current study was designed to use a multi-method approach to develop and validate a scale to measure scientists' self-efficacy for PES activities, with an explicit focus on reciprocal exchanges between scientists and the public. Many existing scales are designed to measure the impact of science communication on the public. Scant options are available for those who are interested in measuring the impact of public engagement on scientists. The Self-Efficacy for PES Scale was developed to fill this void. A validated measure of scientists' self-efficacy has the potential to contribute to both our growing understanding of public engagement by providing a new evaluation and research tool for those who study the impact of science training and public engagement programs. This article presents the methods used to develop and refine the scale and initial reliability and validation results to support the scale's use as a measure of scientists' self-efficacy for PES activities. Although the scale was designed in the United States, the items are meant to be applicable in any context where there is a desire to have scientists engage in two-way dialogue with others.

Method

Instrument

Development of the Self-Efficacy for PES scale was guided heavily by Bandura (2006), particularly in relation to the content of the items. Bandura states that scales to measure self-efficacy must be tailored to the particular area of interest by including items that focus on the many ways that skill can be perceived in that area. Ideal items also focus on the range of challenges that are likely to be experienced within the context of interest. They are worded to focus on the here and now, rather than the future, and should be phrased to focus on perceived rather than demonstrated capability.

For the purposes of the current scale, the domain of interest was defined as public engagement activities that involved direct interaction between scientists and non-scientists who scientists are seeking to engage (i.e., ‘the public’, recognizing that many publics exist). The items chosen focused on several topics that reflected a range of task demands within the context of PES, including self-efficacy for: communicating science, interacting with the public, and preparing for and conducting PES activities and events. These task demands were identified based on the defining factors of PES according to the AAAS Theory of Change for Public Engagement with Science (AAAS, 2016), and with particular attention to reciprocal rather than deficit models of science communication (Ahteensuu, 2011; Cortassa, 2016; Mayhew and Hall, 2012). The AAAS Theory of Change was developed by the organization through an informal but iterative consultation with subject science communication scholars and practitioners. All items focused on perceived judgement of capability, rather than the outcomes that would be expected to result from performance of those capabilities.

A total of 30 items were developed for the initial item pool. The structure and testing of the items followed best practices for scale development (Bandura, 2006; DeVellis, 2016). Several existing scales of self-efficacy for adults were reviewed to serve as models for the final structure of the items. Three proved particularly useful, guiding the framing of some individual items (Midlgey et al., 2000; Poynton, Carlson, Hopper & Carey, 2006, and the DEVISE Self-Efficacy for Science Scale (n.d.)). A six-point Likert scale was used; all responses were either on the positive or negative side of the scale: *strongly disagree*, *moderately disagree, mildly disagree, mildly agree,* *moderately agree, strongly agree*. Positively and negatively worded items were included in an attempt to reflect a comprehensive range of perceived ability and challenge across items; negatively worded items were reverse-coded for all analyses.

Response process validity evidence was then gathered through think-aloud interviews to understand how scientists interpreted and responded to the items (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). A total of 25 scientists were interviewed. Scientists were recruited from networks known to conduct PES activities, and through snowball sampling of those who had already participated in the research. All had been involved in PES within the past year. Most had recently engaged in some sort of public dialogue event (n=16), while others were involved with a university or cooperative extension activity (n=5), policy deliberation (n=2), or knowledge co-production activity (n=2).

Results from the interviews indicated that scientists were likely to choose a range of responses for 19 of the items and that their rationale for those choices reinforced the intent of the questions themselves. These data indicated the potential for these items to capture the full array of beliefs about self-efficacy (DeVellis, 2016). The remaining 11 items were eliminated from the scale; all scientists chose ratings at only the positive or negative end of the scale for these items, indicating that they were not likely to provide meaningful differentiation related to self-efficacy for doing PES. For 17 of the 19 items that remained, scientists' rationale for choosing responses seemed to align with the item's intent. As a result, these items were left unchanged. The wording of two items was edited slightly based on scientists' feedback. The first of these two items was changed from asking about *thinking critically with non-scientists* to *engaging in critical discussion with non-scientists*. For the other, unnecessary verbiage was removed to produce a shorter and more targeted item. The 19 items were then used to collect data from the test sample of scientists. See Table 1 for a full list of items along with descriptive statistics.

Table 1. Descriptive statistics by item

	N	**Range**	**Mean**	**SD**
1. I am good at listening to participants during PES activities.	296	1 - 6	5.30	.85
2. I am good at leaving time for discussion during PES activities.	296	1 - 6	5.08	.94
3. I find it difficult to remove scientific jargon when talking with non-scientists.*	295	1 - 6	2.76	1.31
4. I am able to create props/activities/demonstrations that participants find engaging.	294	1 - 6	4.77	1.14
5. I have a hard time finding PES topics that people connect with.*	296	1 - 6	2.15	1.30
6. I am good at helping people think about the ways that science applies to them.	292	1 - 6	4.96	.98
7. I find it difficult to leave time for people to share their perspectives during PES activities.*	290	1 - 6	2.74	1.25
8. I have a hard time finding the right words to convey my message during PES activities.*	291	1 - 6	2.47	1.22
9. I am good at thinking together with PES attendees about science topics.	289	1 - 6	4.74	.95
10. I am good at knowing when to inform and when to listen during my PES activities.	292	1 - 6	4.57	.99
11. I have a hard time communicating about scientific results with non-scientists.*	287	1 - 6	2.19	1.22
12. I am able to figure out how to improve PES activities based on the kinds of questions the public asks.	288	1 - 6	4.71	1.01
13. am able to engage in critical discussion about science topics with non-scientists.	287	1 - 6	4.89	1.00
14. am able to moderate discussions with participants, even when they include a wide range of perspectives.	286	1 - 6	4.48	1.03
15. am good at reading the audience during PES activities, and making adjustments as needed.	288	1 - 6	4.60	1.09
16. I am good at finding ways to approach difficult topics.	289	1 - 6	4.62	.99
17. I have a hard time answering questions from non-scientists in ways they understand.*	288	1 - 6	2.28	1.13
18. I am able to moderate discussions that allow participants to engage with me and with each other.	288	1 - 6	4.51	1.01
19. I am able to explain a scientific idea in many different ways.	289	2 - 6	5.01	.90

*These items are negatively worded, therefore lower scores indicate higher self-efficacy.

Sample Recruitment

Data for this study were collected via email invitations to networks of scientists who are known to conduct PES activities, and by posting the survey link to scientist listservs that include members who are known to conduct outreach. The survey invitation also included a link that scientists could share with others in their network who might want to participate. At least 12 organizations and listservs shared the survey link. Many participants also learned of the survey through communication from a colleague. Data were collected from September 2016 to February 2017.

The survey invitation specified that the scale was intended for scientists who had conducted at least one PES activity within the past year. Survey questions were then used to confirm and identify the following: whether the participant's current job was that of a scientist or a PES practitioner (as defined by the AAAS PES Logic Model), the total number of years the participant had conducted PES activities, and whether they had conducted a PES activity in the past year. Two final items focused specifically on scientists' most recent PES activity; these documented the topic of their most recent PES activity and the type of PES activity (as defined by the AAAS PES Logic Model).

Data were collected from a total of 361 participants. Of those, 297 defined themselves as scientists, and thus were eligible to participate in the study. Table 2 provides a summary of scientists' background with PES. Almost all participants had been conducting PES for at least one year at the time of the survey. The largest group had between one and five years of experience with PES. The majority of the sample was made up of scientists who had recently conducted PES activities defined as either public dialogue or university/cooperative extension. PES topics were coded by type; the most common PES topics were those from the natural sciences and life sciences disciplines.

Table 2. PES experiences of scientists in the test sample

	Scientists (N=297)
PES Experience	
< 1 year	9%
1-5 years	37%
6-10 years	22%
11-20 years	14%
20+ years	17%
Missing	1%
Type of Most Recent PES	
Knowledge co-production	11%
Policy deliberation	4%
Public dialogue	52%
University/cooperative extension	23%
Missing	10%
Topic of Most Recent PES	
Natural sciences	46%
Life sciences	28%
Computer science and engineering	2%
General STEM topics	3%
Other	9%
Missing	12%

Results

Results are presented below to describe the reliability and validity of the Self-Efficacy for PES scale. Both classical test theory and item response theory methods were used. Classical test theory methods have a long history of being used to develop scales, though these methods have been replaced by Item Response Theory (IRT) in recent years (see Embretson & Reise, 2000 for a review). It has been posited that researchers have favored classical test theory methods because they are more convenient (i.e., model assumptions are often easily attained) and easier to understand than IRT models (Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985; Speer, Robie, & Christiansen, 2016). Several recent articles have taken initial steps toward introducing IRT to psychological scale development, by beginning with classical test theory models to explore the data before moving toward the more appropriate analysis in IRT (see Costa, Asghari & Nicholas, 2017; Edlen & Reeve, 2007; Neumann, Neumann, & Nehn, 2011; Zannon et al., 2016), or by using IRT methods with data produced from scales where, historically, classical test theory methods had been used (see Oon & Subramaniam, 2013; Romine & Walter, 2014). We have modeled our analysis on this approach.

Exploring Internal Structure

Two initial analyses were conducted to examine the internal structure of the 19-item scale as a measure of self-efficacy. Cronbach's coefficient alpha was conducted as a basic measure of reliability, and indicated high internal consistency of the items (α=.90) (Creswell, 2008). Next, an exploratory factor analysis was conducted, using an oblimin rotation. Given that the literature indicates that self-efficacy is a unidimensional trait, the analysis was limited to a one-factor solution. This solution explained 33.62% of the variance in self-efficacy. All 19 items had positive factor loadings, with a range of .39 and .70. Four items had factor loadings below .50. The range of factor loadings found and those with the lowest factor items, in particular, provide initial evidence to suggest variability in the quality of individual items as a measure of self-efficacy. IRT was then used to determine the most useful items among the 19 with regard to creating a parsimonious scale of Self-Efficacy for PES.

Item Response Theory Modeling

Item response theory (IRT) describes the relation between an unobserved (latent) variable, such as self-efficacy, and responses to individual items designed to assess that variable. One benefit of IRT is that it can be used with categorical data such as the Likert-style ratings used on the Self-Efficacy for PES scale. For the purposes of the current study, graded response models were used (see Samejima, 1969; 1996). Graded response models assume that the order of response options is known (for example, that

Strongly Disagree is a lower rating than Disagree) and that the slope of each item will vary based on the item itself, such that some items are stronger measures of different levels of the latent trait than others. The latter of these assumptions was tested as part of the analysis. Two versions of the graded response model were calculated and compared using MPlus 7 (Muthen & Muthen, 1998-2012). The first model generated estimates for both the slope and location of each item in the model. The second estimated the location only, and held the slope constant (and thus was akin to a Rasch model). The two models were then compared using a likelihood test; results indicated that the default graded response model was the better fit for the data compared to the model that constrained the slope of each item and thus confirmed that items in the model were differently discriminating with regard to self-efficacy (df=18, p<.001).

For the remaining analysis, the graded response model was used to create estimates for two parameters: (a) item-level estimates were created to determine the *difficulty* of each item; and (b) person-level estimates were created to determine each item's ability to *discriminate* between people with high and low scores on the latent trait. These estimates were then used to determine the items that provided the best measures of self-efficacy. Two criteria were used to eliminate items. The first focused on the discrimination parameters, or each item's ability to differentiate between scientists who had different levels of self-efficacy (i.e., the person-level estimate). In practice, a high discrimination parameter value means that the probability of a correct response increases more rapidly as the ability (latent trait) increases (An and Yung, 2014). Acceptable discrimination values are greater than 1.0; the discrimination values for items on the Self-Efficacy for PES scale ranged from .87 to 2.04. Two items (#2 and #3) were eliminated due to low discrimination levels.

The Item Characteristic Curves (ICCs) were then explored for the remaining items. ICCs present the extent to which an item is reliable across the range of all response options provided. An ideal ICC for a polytomous item is distributed such that each response option has its own peak when graphed against the latent variable. The peaks indicate that the item discriminates at that level of the latent variable (i.e., the item can detect people who are at a level of two on the six-point scale). Variations of this curve can also reflect appropriate items, depending on the intended audience and use for the scale. The intended audience for the current scale is scientists with a range of experience conducting PES activities. Though the current sample consisted of a self-selected sample of scientists, the scale is intended to capture the full range of self-efficacy performance. Given this intended use, a pair of related criteria were used to determine suitable discrimination for the purposes of the current study. Items were retained if at least five of the six rating responses were used by scientists to rate their self-efficacy *and* if at least five of the six had the highest curve for some portion of the distribution across self-efficacy ratings. Four items were eliminated because they did not meet these criteria (items 1, 6, 11, and 19).

The results from the initial graded response model reduced the scale to 13 items. Next, the graded response model was calculated again to include only these items. The test information curves from both the 19-item and 13-item model are presented in Figure 1 to demonstrate the reliability of the entire scale across the latent variable in relation to the conventional cut-off of .80. The results for both scales indicate high reliability across a wide range of self-efficacy levels. Both models yielded the highest reliability for those with low levels of self-efficacy. Lower reliability was found for those with high levels of self-efficacy, though reliability for these individuals was also well above the conventional cut-off for all but those who provided the highest ratings to describe their self-efficacy for PES (i.e., those who provided ratings that were greater than 2.5 standard deviations above the mean). The difference between the two scales was greater for those with low compared to high self-efficacy. Both scales demonstrated reliable scores from the lowest end of the range up through approximately two standard deviations above the mean. Given our interest in creating the most parsimonious scale possible, the remaining analyses focused on the 13-items scale.

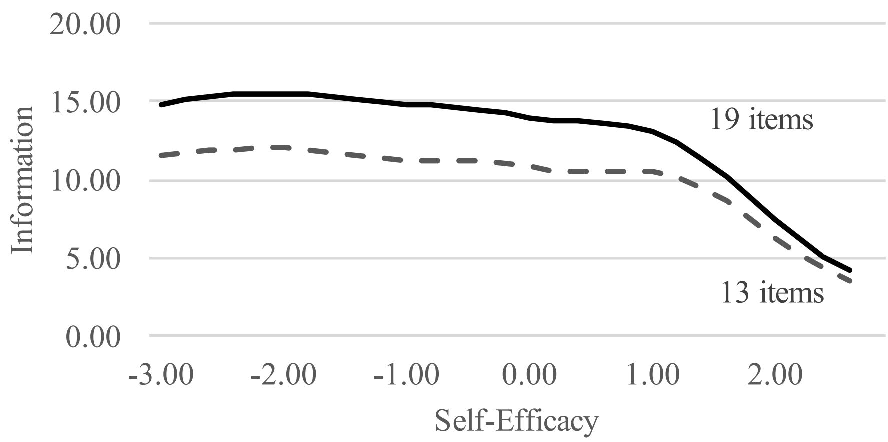

Figure 1. Information curves for the self-efficacy for PES scale.

The means and standard deviations for each individual item are shown in Table 3, along with their threshold and discrimination parameters. Item 15 had the highest discrimination parameter, followed by Items 18 and 13, indicating that these items had the greatest ability to detect differences between perceived levels of self-efficacy among scientists in the sample. Item 7 had the lowest discrimination parameter. The mean Self-Efficacy for PES score for the sample was 4.64 (*SD*=.69), with a range of 2.46 to 6. These results indicate that scientists had moderately positive self-efficacy for PES overall. The distribution of scores indicates that the scale detected a broad range of self-efficacy among scientists. The distribution of scores overall is presented in Figure 2.

Compared to the ideal normal distribution curve on the graph, scores on the Self-Efficacy for PES were slightly low. This graph reiterates the results presented earlier that indicated that the scale may not yet include the kinds of difficult items that measure those with the highest levels of self-efficacy.

Table 3. Means, standard deviations, and estimated parameters for Self-Efficacy for PES items

			Thresholds					
	Mean	**SD**	1	2	3	4	5	**Discrimination**
4. I am able to create props/activities/demonstrations that participants find engaging.	4.77	1.14	-3.97	-2.98	-2.16	-0.80	0.95	1.17
5. I have a hard time finding PES topics that people connect with.*	2.15	1.30	-4.14	-2.52	-1.55	-0.83	0.24	1.28
7. I find it difficult to leave time for people to share their perspectives during PES activities.*	2.74	1.25	-5.37	-2.57	-1.15	0.01	1.86	1.01
8. I have a hard time finding the right words to convey my message during PES activities.*	2.47	1.22	-4.76	-2.78	-1.29	-0.50	1.25	1.16
9. I am good at thinking together with PES attendees about science topics.	4.74	.95	-3.99	-2.79	-1.89	-0.54	1.15	1.78
10. I am good at knowing when to inform and when to listen during my PES activities.	4.57	.99	-3.47	-2.61	-1.76	-0.32	1.52	1.65
12. I am able to figure out how to improve PES activities based on the kinds of questions the public asks.	4.71	1.01	-3.29	-2.50	-1.81	-0.49	1.11	1.83
13. I am able to engage in critical discussion about science topics with non-scientists.	4.89	1.00	-3.35	-2.49	-1.88	-0.78	0.78	1.95
14. I am able to moderate discussions with participants, even when they include a wide range of perspectives.	4.48	1.03	-3.20	-2.50	-1.41	0.01	1.30	1.91
15. I am good at reading the audience during PES activities, and making adjustments as needed.	4.60	1.09	-3.00	-2.06	-1.48	-0.33	1.07	2.14
16. I am good at finding ways to approach difficult topics.	4.62	.99	-4.14	-2.81	-1.71	-0.26	1.27	1.70
17. I have a hard time answering questions from non-scientists in ways they understand.*	2.28	1.13	-3.67	-2.57	-1.51	-0.58	0.84	1.68
18. I am able to moderate discussions that allow participants to engage with me and with each other.	4.51	1.01	-3.73	-2.26	-1.57	-0.11	1.36	1.99

*These items are negatively worded, therefore lower scores indicate higher self-efficacy.

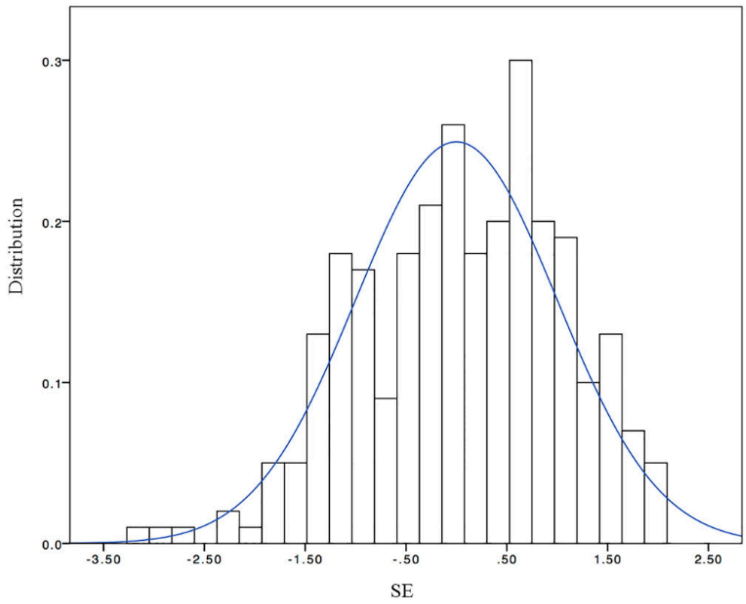

Figure 2. Distribution of self-efficacy for PES scores within the graded response model.

Discussion

This study used a multimethod approach to develop a scale to measure scientists' self-efficacy for public engagement, and to provide initial validation evidence to support the scale's use. Two versions of the scale were tested, and the results indicated that both provided reliable information to document most scientists' self-efficacy with PES, including those who believed their ability to conduct PES was well below and well above the mean. Though the 13-item scale was slightly less reliable than the 19-item scale, the reliability for the shorter version was still well above the conventional cut-of .80 for the vast range of self-efficacy levels, and likely to be the preferred version for future research and evaluation efforts. This version of the scale resulted in a range of scores for individual scientists in the sample, and an average in the moderate positive range overall.

The Self-Efficacy for PES scale was developed with the hope that it could eventually become a common measure for the evaluation of scientists' self-efficacy. Research on self-efficacy has documented the strength of this construct in relation to scientists' public engagement choices and attitudes (Dudo, 2012; Dudo & Besley, 2016), and thus it is not surprising that self-efficacy is a short-term outcome in the AAAS logic model for PES and a common outcome for scientist training programs. Though self-efficacy is an intuitive construct of interest for those who study and evaluate public engagement, there are currently no validated measures of this construct for scientists. The Self-Efficacy for PES scale has the potential to fill this void, serving as either a reflection tool for scientists during an intervention or as part of a formal research or evaluation study. The scale's focus on reciprocal rather than deficit model exchanges between scientists and the public has the added benefit for reinforcing this key shift in science communication.

Scientists from a range of PES environments were included throughout the validation process, to collect data from the wide range of contexts in which PES occurs. The test sample includes those from a range of scientific disciplines and those who conducted PES in each of the four contexts defined by the AAAS PES Logic Model. The test sample also included scientists with a range of experience conducting PES, from those who had decades of experience to those who were relatively new to public outreach. Scientists' PES experience was positively related to overall scores on the Self-Efficacy for PES scale, providing initial evidence to support its discriminant validity.

Though the results from this study are encouraging, this research is only a first step in exploring the validity and potential of the Self-Efficacy for PES scale. Future validation efforts should investigate the use of the scale with a wider range of scientists than the self-selected sample used for the current study to explore the discriminant validity of the measure further. Predictive validity evidence would also provide additional information to support the use of this scale, and might be documented by comparing scientists' ratings on the scale with observations of their PES skills during activities. Studying the scale's ability to detect pre-post change is an area that is likely to be a great interest to those who lead communication training programs for scientists. The responses provided by the test sample indicate that the full range of ratings were selected for each item. Individual mean scores ranged from the lower middle of the scale (i.e., around 2) to the highest rating possible (a 6), with a sample average in the moderate positive range (4 out of 6. These results indicate the lack of a ceiling effect, and thus the potential for growth in scientists' ratings. Though these results do indicate the potential for upward movement, future work is needed to document whether the scale is sensitive enough to detect change over time.

A related limitation is the fact that the scale was a less reliable measure for scientists with the highest levels of self-efficacy. This characteristic might be particularly problematic for those who would choose to use the scale to measure pre-post change in response to a scientist training or intervention. The current scale provides a reliable measure for a wide range of self-efficacy. Even so, new items should be developed to target those who consider themselves to have highest self-efficacy for PES. This additional work would enhance the scale's potential to detect pre-post change for interventions that are successful at fostering perceived abilities that are more than two standard deviations about the mean.

Conclusions

The current study provides critical initial evidence to support the use of the Self-Efficacy for PES Scale to measure scientists' perceived ability to conduct public engagement. The items on the scale reinforce an intent for public engagement activities to be reciprocal rather than one-way exchanges between scientists and the public. The scale has the potential to serve as both a reflection tool for scientists who conduct PES and as a measurement tool for those who evaluate and study public engagement.

Though this study documents the potential of the scale, its true value will only be learned through the continued study of its use within the wider range of public engagement activities, science communication interventions, and research and evaluation related to those efforts. There is much to learn about public engagement, its impact on scientists, the relation between a scientist's self-efficacy and the quality of their PES activities, and the relation between a scientist's self-efficacy and their impact on the public via PES activities. It is our hope that the Self-Efficacy for PES scale can play a critical role in helping to answer these and other research and evaluation questions in the years to come.

References

American Association for the Advancement of Science (n.d.). *Leshner Leadership Institute for Public Engagement with Science.* Retrieved from https://www.aaas.org/pes/leshner-leadership-institute

American Association for the Advancement of Science (2016). *Theory of change for public engagement with science.* Retrieved from https://www.aaas.org/page/theory-change-public-engagement-science

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ahteensuu, M. (2011). Assumptions of the deficit model type of thinking: Ignorance, attitudes, and science communication in the debate of genetic engineering in agriculture. *Journal of Agricultural and Environmental Ethics, 25,* 295-313.

Alda-Kavli Learning Center for Science Communication (n.d.). *Alda Center info.* Retrieved from http://www.aldakavlilearningcenter.org/get-started/about-us

Allum, N. C., Sturgis, P., Tabourazi, D., & Brunton-Smith, I. (2008). Science knowledge and attitudes across cultures: A meta-analysis. *Public Understanding of Science, 17*(1), 35-54.

An, X. & Yung, Y. (2014). *Item response theory: What it is and how you can use the IRT procedure to apply it* [White paper]. Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Anderson, C. B., Lee, H. Y., Byars-Winston, A., Baldwin, C. D., Cameron, C., & Chang, S .(2016). Assessment of scientific communication self-efficacy, interest, and outcome expectations for career development in academic medicine. *Journal of Career Assessment, 24*(1), 182-196.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191-215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Upper Saddle River, N. J.: Prentice Hall.

Bandura, A. (1995). *Self-efficacy in changing societies*. New York: Cambridge University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman and Company.

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents*, *5*(307-337).

Besley, J. C., Dudo, A., Yuan, S., & AbiGhannam, N. (2016). Qualitative interviews with science communication trainers about communication objectives and goals. *Science Communication, 38*(3), 356-381.

Besley, J. C., Oh, S. H., & Nisbet, M. (2013). Predicting scientists' participation in public life. *Public Understanding of Science, 22*(8), 971-987.

Britner, S. L., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering, 7*(4), 1-15.

Chemers, M. M., Li-tze Hu, & Garcia, B. F. (2001). Academic self-efficacy and first-year college student performance and adjustment. *Journal of Educational Psychology, 93*(1), 55-64.

Costa, D. S. J., Asghari, A., & Nicholas, M. K. (2017). Item response theory analysis of the pain self-efficacy questionnaire. *Scandinavian Journal of Pain, 14*, 113-117.

Cortassa, C. (2016). In science communication, why does the idea of a public deficit always return? The eternal recurrence of the public deficit. *Public Understanding of Science, 25*(4), 447-459.

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, N. J.: Pearson.

DeVellis, R. F. (2016). Scale development: Theory and applications (Vol. 26). Sage publications.

Dudo, A. (2012). Toward a model of scientists' public communication activity: The case of biomedical researchers. *Science Communication, 35*(4), 476-501.

Dudo, A. & Besley, J.C. (2016). Scientists' prioritization of communication objectives for public engagement. *PLOS One, 11*(2), 1-18.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.

Fast, L. A., Lewis, J. L., Bryant, M. J., Bocian, K. A., Cardullo, R. A., Rettig, M., Hammond, K. A. (2010). Does math self-efficacy mediate the effect of the perceived classroom environment on standardized math test performance? *Journal of Educational Psychology, 102*(3), 729-740.

Finney, S. J. & Schraw, G. (2003). Self-efficacy beliefs in college statistics courses. *Contemporary Educational Psychology, 28*(2), 161-186.

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

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* New York: Springer.

Horst, M. (2013). A field of expertise, the organization, or science itself? Scientists' perception or representing research in public communication. *Science Communication, 35*(6), 758-779.

Hsu, M-H, Ju, T. L., Yen, C-H, & Chang, C-M (2007). Knowledge sharing in virtual communities: The relationship between trust, self-efficacy, and outcome expectations. *International Journal of Human-Computer Studies, 65*(2), 153-169.

Mayhew, M. A. & Hall, M. K. (2012). Science communication in a café scientifique for high school teens. *Science Communication, 34*(4), 546-554.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., & Urdan, T. (2000). *Manual for the patterns of adaptive learning scales.* Ann Arbor, 1001, 48109-1259.

Miller, S. (2001). Public understanding of science at the crossroads. *Public Understanding of Science, 10*(1), 115-120.

Morony, S., Kleitman, S., Lee, Y. P., and Stankov, L. (2013). Predicting achievement: Confidence vs. self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research, 58,* 79-96.

Muthén, L.K. and Muthén, B.O. (1998-2012). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

National Academies of Sciences, Engineering and Medicine. (2016a). *Communicating Science Effectively: A Research Agenda*. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering and Medicine. (2016b). *Science Literacy: Concepts, Contexts, and Consequences*. Washington, DC: The National Academies Press.

National Co-ordinating Centre for Public Engagement (n.d.) Retrieved from https://www.publicengagement.ac.uk/

Neumann, I., Neumann, K., & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education*, *33*(10), 1373-1405.

Oon, P. & Subramaniam, R. (2013). Factors influencing Singapore students' choice of physics as a tertiary field of study: A Rasch analysis. *International Journal of Science Education, 35*(1), 86-118.

Pacific Science Center (2017). *About Portal to the Public.* Retrieved from https://popnet.pacificsciencecenter.org/about/

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-577.

Pajares, F., Britner, S. L., & Valiante, G. (2000). Relation between achievement goals and self-beliefs of middle school students in writing and science. *Contemporary Educational Psychology, 25*(4), 406-422.

Pajares, F., & Miller, M. D. (1997). Mathematics self-efficacy and mathematical problem solving: Implications of using different forms of assessment. *Journal of Experimental Education, 65*(3), 213-28.

Pajares, F., & Urdan, T. C. (2006). *Self-efficacy beliefs of adolescents.* Greenwich, Conn.: IAP – Information Age Publishing, Inc.

Poliakoff, E. & Webb, T. L. (2007). What factors predict scientists' interactions to participate in public engagement of science activities? *Science Communication, 29*(2), 242-263.

Poynton, T. A., Carlson, M. W., Hopper, J. A., & Carey, J. C. (2006). Evaluation of an innovative approach to improving middle school students' academic achievement. *Professional School Counseling, 9*(3), 190-196.

Romine, W. L. & Walter, E. M. (2014). Assessing the efficacy of the measure of understanding of macroevolution as valid tool for undergraduate non-science majors. *International Journal of Science Education, 36*(17), 2872-2891.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory.* New York: Springer.

Schunk, D. H. (1989). Self-efficacy and achievement behaviors. *Educational Psychology Review, 1*(3), 173-208.

Selvakumar, M. & Storksdieck, M. (2013). Portal to the Public: Museum educators collaborating with scientists to engage museum visitors with current science. *Curator: The Museum Journal, 56*(1), 69-78.

Sims, M. J., Madden, H., Cacciatore, M. A., & Yeo, S. K. (2016). The lure of rationality: Why does the deficit model persist in science communication? *Public Understanding of Science, 25*(4), 400-414.

Skaalvik, E. M., Federici, R. A., & Klassen, R. M. (2015). Mathematics achievement and self-efficacy: Relations with motivation for mathematics. *International Journal of Education Research, 72,* 129-136.

Speer, A. B., Robie, C., & Christiansen, N. D. (2016). Effects of item type and estimation method on the accuracy of estimated personality trait scores: Polytomous item response theory models versus summated scoring. *Personality and Individual Differences, 102*, 41-45.

Wirawan, H. & Bandu, M.T. (2016). A review of self-efficacy training for international student. *The International Journal of Information and Learning Technology, 33*(2), 115-128.

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 18.

Zeldin, A. L. & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal, 37*(1), 215-246.