Research Article

Science Communication
2021, Vol. 43(2) 139–169

DOI: 10.1177/1075547020971639
journals.sagepub.com/home/scx
SAGE

Assessment by Audiences Shows Little Effect of Science Communication Training

Margaret A. Rubega[1], Kevin R. Burgio[2], A. Andrew M. MacDonald[3], Anne Oeldorf-Hirsch[1], Robert S. Capers[1], and Robert Wyss[1]

Abstract

As the science community has recognized the vital role of communicating to the public, science communication training has proliferated. The development of rigorous, comparable approaches to assessment of training has not kept pace. We conducted a fully controlled experiment using a semester-long science communication course, and audience assessment of communicator performance. Evaluators scored the communication competence of trainees and their matched, untrained controls, before and after training. Bayesian analysis of the data showed very small gains in communication skills of trainees, and no difference from untrained controls. High variance in scores suggests little agreement on what constitutes "good" communication.

Keywords

science communication, graduate training, assessment, evaluation, evidence

[1]University of Connecticut, Storrs, CT, USA
[2]Cary Institute for Ecosystem Studies, Millbrook, NY, USA
[3]Université de Montréal, Montreal, Quebec, Canada

Corresponding Author:
Margaret A. Rubega, Department of Ecology and Evolutionary Biology, University of Connecticut, 75 N. Eagleville Road, U-3043, Storrs, CT 06269, USA.
Email: margaret.rubega@uconn.edu

> The single biggest problem in communication is the illusion that it has taken place.
>
> —George Bernard Shaw

Scientists spend upward of a decade learning to communicate in the specialized language of their disciplines and subdisciplines. The science community is unified behind the idea that it is also vitally important that scientists communicate the results of their work to the public, with federal funding agencies increasingly focused on formal and informal outreach as a component of research activities, and with communication training as a component of STEM (science, technology, engineering, mathematics) graduate education. Rigorous assessment of such training has lagged behind.

There is broad agreement that to communicate to the public successfully, scientists must use different language and approaches than those used in the scientific arena itself (Fischhoff & Scheufele, 2013, 2014). The belief that those skills can be taught has led to the proliferation of programs to provide training in science communication both in and out of academic institutions.

Programs are aimed at undergraduate and graduate students as well as working scientists in academia, government, or nongovernmental organizations. Training programs vary widely and include 1 or 2 hours over 1 day; full-day or week-long workshops once or repeated over several months; and full degree programs (Baram-Tsabari & Lowenstein 2017).

Most programs are aimed at oral communication, and can include media training with journalists, or storytelling exercises (see, e.g., StoryCollider.org, StoryCirclesTraining.com). However, written elements intended to improve science communication, such as message distillation (e.g., message boxing; Baron, 2010; COMPASS, 2017) are often included, and there are formal programs aimed at writing about science for the public (e.g., Druschke et al., 2018). One well-known training program incorporates acting improvisation (AldaCenter.org), while others include exercises in using dance, visual arts, and poetry to communicate scientific information.

An important element of most training programs involves identifying the audience of a message, whether other scientists, public officials, journalists, or other nonscientists. Increasingly, consideration of audience values, goals, and identity (Besley, 2015; Besley et al., 2015; Dudo & Besley, 2016; Peterman et al., 2017; Smith et al., 2013), sometimes referred to as "engagement" (see Rowe et al., 2016 for a review of various and other uses of the term), have become a feature of well-recognized training programs, such as COMPASS.

This drive to provide science communication training is necessary and welcome; cognitive awareness of the barriers to communication is an essential first step that trainings contribute to. However, to date, there is very little research establishing standards of evidence by which we can judge whether these training activities work to produce effective science communicators *in practice* (but see Rodgers et al., 2018 for a recent exception): How do we know that the training actually increases communication skills? Furthermore, there is no scale along which the relative effectiveness of one training approach can be compared with another. If graduate students are going to spend time away from the bench or field sites to learn to communicate with public audiences, should that time be invested in a full-semester course, a 3-Minute Thesis competition, or a day-long improvisation workshop? Is one training sufficient, or should training be ongoing throughout a graduate program (or, indeed, a career)? While trainees may gain different but equally valuable skills from different trainings, when federal funding and graduate training time are being invested, the ability to identify the most time- and cost-effective approach is fundamental.

Ideally, the skills taught in science communication training are based in communication theory about how audiences seek, receive, assimilate, and use scientific information. In addition, the training should draw on educational theory about skill development. Science communication, as a discipline, is influenced by many other fields, making it a loosely connected patchwork of concepts and theory (Kuehne et al., 2014).

In addition, communication is a multistep process, and each step must be executed successfully if the goal of the communicator with respect to the audience is to be achieved. Although some change in the behavior of the audience may be the ultimate goal of communication, achieving this change depends on mastery, and integration, of all the steps. In the context of science communication training, the change in behavior of the trainee is the subject of interest, although achieving such a change in the audience certainly counts as evidence of successful training. Communication research has focused on many elements that comprise effective communication, particularly in terms of credibility or trust. This concept alone has been conceptualized many ways, such as a mix of ability, benevolence, and integrity (Mayer et al., 1995, 2007); of believability, accuracy, trustworthiness, bias, and completeness (Flanagin & Metzger, 2000); competence and warmth (Fiske & Dupree, 2014); or accuracy, authenticity, and believability (Appelman & Sundar, 2016); among many others. Yet there is no agreed-upon standard for this measure, which varies across contexts.

Thus, in this context we focus on the communicator's ability to provide information clearly and understandably (clarity), on the communicator's ability to appear knowledgeable and trustworthy (credibility), and on the communicator's ability to make the audience interested in the subject (engagement). We hold that, while communication is a complicated, multistep process and communication experts disagree about the meaning of "effectiveness," it cannot be achieved, whatever the ultimate goal, unless each of these conditions exists. As in any other branch of science, communication theory requires validation, and measurement that is comparable across different situations (Schemer et al., 2014). A carefully constructed training, assuming that it leads to communication with a public audience, can provide a test of both the theory and of the approach to training. Viewed in this framework, all science communication trainings are experiments, albeit usually uncontrolled ones, and the results should indicate whether communication theory works in the field, producing effective communication and successful science communicators. Thus, every science communication training should be accompanied by rigorous assessment of the ability of trainees to communicate science effectively, and that assessment needs to be transferable among training styles.

Frequently, the assessment of science communication training is based on trainees' self-assessment via survey instruments (e.g., Rodgers et al., 2020); true external assessment of their skills (as opposed to, e.g., their sense of self-efficacy) is almost unknown (Baram-Tsabari & Lewenstein, 2013, 2017). While it may be useful to assess whether trainees believe that they have learned something, there are serious, well-known shortcomings with this approach (Dunning et al., 2004; Falchikov & Boud, 1989; Hansford & Hattie, 1982). First, the reason training is attempted in the first place is that scientists consistently, predictably, make mistakes in judging what audiences will find clear and interesting, much less what will move them to some desired action. Moreover, trainees assessed this way are rarely asked to compare the value of the training they are assessing with a *different* form of training; in most cases, trainees have been exposed to only one form of training, and are therefore unable to provide a comparison. More fundamentally, self-assessors are likely to be resistant to ranking their own performance as low, either as learners (Dunning et al., 2004) or as active communicators (Mort & Hansen, 2010). To the extent that they find their training interesting or thought-provoking, trainees may be inclined to provide the trainer(s) with positive feedback and rank the training itself as useful, even if they have gained no practicable skills as communicators.

Most important, *a belief in self-efficacy is not itself a measure of effectiveness.* Research has shown, repeatedly and across disciplines, that self-assessments inflate communication competence relative to external evaluation (e.g., Duran & Zakahi, 1987; Eva et al., 2004; Gruppen et al., 1997; McCroskey & McCroskey, 1988; Mort & Hansen, 2010). The key measure of the effectiveness of any form of communication training is not only evidence that a target *audience* judges the trainee effective (Bray, 2012; Rodgers et al., 2018) but also that the target audience finds the trainee a more effective communicator after training than before. This is a crucial point when the explicit goal of so much of science communication is not merely to inform but to influence public opinion and policy on matters of profound civic importance, such as climate change, and to engage public audiences in science as a tool for decision making.

In order to develop a rigorous approach to science communication training assessment that would be comparable across varied training approaches and would provide a direct measure of audience reaction to a communicator, we conducted a fully controlled experiment in science communication training. As the treatment, we used a semester-long graduate science communication course, which was carefully designed to teach best practices according to theory about the communication of science (Fischhoff & Scheufele, 2013, 2014), and we used a large undergraduate class as a test audience. Audience members provided fully independent scores of the effectiveness of the standardized communication of both the trainees and their matched, untrained controls, both before and after the training period. Our aim was three-fold: (a) we wished to explore the usefulness of audience members' responses in assessing communication effectiveness, in the interests of developing a rigorous, scalable, transferable assessment method that could be used to evaluate the effectiveness of individual training programs, and to compare different programs; (b) we wished to determine whether self-assessment aligned with the assessment provided by external evaluators; and (c) we wanted to assess whether science communication training, including our own course, results in measurable improvement in science communication skills, as assessed by an audience.

Method

Science Communication Course

With the assistance of an expert in educational theory, we created a graded, three-credit, semester-long science communication course that was designed to engage both STEM graduate students and journalism undergraduate students in the theory and process of communicating science to public audiences. Three of us were involved in both the design process and in teaching the course (MR, an active science communicator who had been teaching science communication to STEM graduate students for the previous decade; RW, a journalist with 30 years of experience before becoming a journalism professor; and RC, a former journalist with a Pulitzer Prize in investigative reporting and a PhD in Evolutionary Biology). Journalism students were present in the class as a training aid to the subjects; although their own learning was facilitated by the class, they were not themselves experimental subjects, and our data collection activities did not include them, or their work. Although a training approach that took less time (e.g., day- or week-long workshops) would have yielded a larger sample size, we chose to work more intensively with fewer students in order to maximize the likelihood of a training effect large enough to be measurable.

We taught the course every fall semester for 3 years (2016-2018). In order to attract students from a wide range of STEM disciplines, each year we advertised the course to every STEM department on the University of Connecticut campus via e-mail to departmental e-mail lists and campus-wide news digests. Journalism students were recruited via announcements to journalism classes and the departmental e-mail list. In order to ensure a consistent and high level of active interaction with the journalists and practice for trainees, we limited the course to 10 graduate STEM students each year and aimed for at least half as many undergraduate journalism students; in 2 of 3 years we exceeded that mark (Fall 2016, 4 students; Fall 2017, 8 students; Fall 2018, 7 students).

The course consisted of a 4-week introductory phase in which readings from the science communication literature, lectures, and discussions highlighted the role of scientists and journalists in public communication of science. We also identified known barriers to effective science communication and introduced various approaches to overcoming those barriers (e.g., Message Boxing, COMPASS Science Communication, Inc., 2017; framing, Davis, 1995; Morton et al., 2011; narrative structure, Dahlstrom, 2014; intellectual humility, that is, openness to audience expertise and viewpoint: Lynch, 2017; Lynch et al., 2016). Active learning exercises during this phase were designed to make science and journalism students comfortable with collaboration, and to make theory concrete (see Supplemental Materials, available online, for our syllabus with further detail). All 11 subsequent weeks of the semester were devoted to active practice and postpractice reflection on science communication skills. We required each STEM student to be interviewed by a journalism student; the 20-minute interviews were conducted outside of class and were video recorded. Both the STEM student and the journalism student were required to complete and submit forms detailing their process of preparation for the interview. The journalism student then produced a short (500-word) news story based on the interview, which served, in part, to make manifest the ways in which the STEM student had failed to help the journalist understand the material. The whole class in a subsequent course meeting reviewed both the written piece and the video. Every student was required to produce and hand in written peer analysis/feedback forms completed while watching the video. We discussed and critiqued with the students the level of success the scientist had in communicating a technical research issue, and explicitly drew connections between the communication behavior of the scientist in each video with the conceptual material covered earlier in the course. We also reviewed and discussed the level of understanding the journalist gained in interpreting that message, as displayed in the news story. We required that each STEM graduate student do two interview sessions, resulting in 20 interviews displayed and discussed in each semester's course, for a total of 60 over the entire study.

In addition to our other data collection (see Data Collection), all STEM students completed both standard university Student Evaluations of Teaching and our own end-of-course evaluation survey, in which STEM students addressed their own perceived self-efficacy in greater detail.

Data Collection

Subject Selection. True randomization of students enrolled in a treatment class is, of course, not possible since students who did not wish to take the class could not be compelled to do so. Given that, we focused on controlling factors other than training that might influence results. We selected a total of 30 STEM trainees during the fall semesters of 2016 to 2018. In the first (Fall 2016) iteration of the course, a first-year postdoctoral researcher was allowed to take the class when an admitted student failed to register; the admitted postdoc completed all course requirements and participated in all research-related activities and is treated in our data set as any other trainee. In the fall of 2017, one student dropped out of the course too late to be replaced, leaving us with a total pool of 29 trainees. Course advertisements generated requests for permission numbers for the class from STEM graduate students across a wide range of disciplines, degree programs, and stages of graduate career; there was a waiting list every semester we taught the course, which by the third iteration had more students on it than there were seats in the class.

We sent STEM students who asked for permission to take the class an information sheet that stated that the course was the subject of research on the effectiveness of science communication teaching methods, and as such, would require complete attendance (i.e., would not allow skipping class for research activities or conferences out of town) even from students who chose not to give their consent to being study subjects; this policy reduced variance in communication competence that may have arisen due to missing class exercises, discussions, or active practice. Prospective students were also asked to fill out a questionnaire affirming that they had no barriers to consistent, complete attendance, and providing information in their discipline, degree program (MS or PhD), year of their program, stage of their research project (e.g., project design, data collection, analysis, writing), gender, status as an English as a first- or second-language speaker, and previous experience with science communication and science communication trainings (e.g., independent reading; hour-long, day-long, or week-long workshops; or semester-long classes).

Exact composition of the classes depended on the pool of applicants for entry to the class, but in choosing STEM students to admit to the class, we applied a hierarchy of goals to be met for the study; in descending order of importance they were: Discipline (maximizing the range of disciplines represented in the classroom), Stage (preferring late-stage students over early-stage), Gender (balancing in a given class), and ESL (English as second language) status (non-ESL students were preferred, all else being equal). We excluded those with scheduling conflicts (e.g., students who declared they were already committed to fieldwork or a conference presentation that would cause them to miss classes), those who were at too early a stage in their graduate careers to have any data they could communicate about, and those with more than a single hour-long science communication workshop training in their background.

Recruitment for the course resulted in the enrollment of students from a wide variety of STEM disciplines: Animal Science, Chemistry, Ecology and Evolutionary Biology, Environmental Engineering, Genetics and Genomics, Geological Sciences, Molecular and Cell Biology, Natural Resources, Physiology and Neurobiology, and Statistics. Factors higher in our hierarchy of goals resulted in the selection of at least one ESL speaker in every class.

Control Selection. Many factors can affect an individual's ability to communicate science well, including experience, prior training, and scientific discipline. We wished to isolate the effect of training, specifically, in our course. Therefore we analyzed subjects in pairs: For each STEM trainee, we recruited (via campus-wide ads that offered payments for participation) and selected a control from a pool of volunteer graduate students across STEM departments at the University of Connecticut. Graduate students who volunteered as controls filled out an online survey in Qualtrics XM (Qualtrics, Provo, UT, USA) that asked for demographic, first language, and education information, along with information about the level of previous science communication training (none; short workshop [hour-long, day-long], longer [week to semester-long training]); the latter information helped us control for the fact that students who registered for the course were a self-selected sample with declared interest, and perhaps greater-than-average experience, in science communication concepts and practice. From the pool, we selected the individuals who matched most closely with each trainee taking the course, taking into account (in order of importance): gender, first language, department, number of years in graduate school, and prior science communication training, if any. All 29 students were matched with controls with the same gender and first language (i.e., English vs. ESL). We were able to match 18 of the 29 students to controls in their same academic department; where limitations of the volunteer pool of controls did not allow controls to be drawn from the same department as their trainee, we matched as closely as possible within general discipline (e.g., a Statistics trainee matched to a Mathematics control). Twenty of the 29 trainee/control pairs were matched in having had no previous formal communication training. The remaining 10 trainee control pairs were matched as closely as possible, given the volunteer pool; none of either the trainees or controls in the imperfectly matched pairs had more than a short workshop aimed at science communication, and in all but two cases, the trainees had the greater training exposure.

Video Recording. At the beginning and end of the semester, we asked both trainees and controls to respond to the prompt: *How does the scientific process work*? while we recorded them with a video camera. (Journalism students did not make videos, and their performance is not analyzed here.) The prompt, by design, had no relation to any specific communication tasks that trainees were assigned in class; the aim of the training was to prepare them to apply what they had learned, and successfully communicate about science, in any context. Using a prompt not encountered in the class also avoided confounding results by preventing the instructors from "teaching to the test," and thereby incorporating instructor feedback (that controls had no access to) into the performance measure. We selected this prompt because the scientific process is often mis- or incompletely understood by the target audience (undergraduates; see the "Video Ratings" section), it is a question that any graduate STEM student should be able to answer, regardless of scientific discipline, and it removed the potential for audience bias that could be introduced by controversial subjects (e.g., climate change, evolution). While standardizing what the students communicated about prevented them from making judgments about what might interest the audience that might have improved audience engagement in some cases, it also eliminated such judgments as a source of variance in performance.

Via consent documents that subjects read and signed in agreeing to participate in the research, trainees and controls were informed about the pre-and post-semester video-recording requirement and about the prompt they would be expected to address during the recordings, before the class began. The consent form also explained to students that the videos were being used as part of our experimental procedure to measure the effectiveness of the training program. Subjects received the information a minimum of 1 week before the first recording, and were aware for the entire 15-week semester that they would be repeating the recordings, with the exact same prompt, at the end of the course. Subjects were also provided with an additional written copy of the prompt immediately before every recording.

During the recording, we allowed subjects to talk for a maximum of 3 minutes and allowed them to stop as early as they felt appropriate. All recordings were made in the same university studio, using the same cameras, positioning, and lighting, with the same uniform, featureless background, under the direction of a university staff member. Videos showed only the head and shoulders of the trainee or control who was speaking.

Video Ratings. To assess the effectiveness of the trainees' and controls' communication, videos were rated by undergraduate students in a research participation pool (*evaluators*) that is part of a general education introductory communication course in which students receive course credit for participating in research. We uploaded both the current semester's "before" videos and the previous semester's "after" videos for trainees and controls to an online Qualtrics portal, totaling approximately 40 videos per semester. Students in the research participation pool could choose to participate in our study by evaluating a video. Each evaluator was randomly assigned by Qualtrics to view just one video and complete a set of ratings about it, after confirming that she or he could see and hear the video. Once students had participated, they could no longer evaluate videos in our project, ensuring that we avoided any evaluation bias resulting from an evaluator's seeing, for example, an After video before evaluating a Before video, or a Trainee's video before evaluating a Control. We included a "speed bump" question ("Please click the value for '3'") to eliminate the evaluations of students who clicked either at random or on just a single Likert-type rating throughout the whole scoring tool to complete the task for credit without actually evaluating the video. We also eliminated evaluations that were not completed. Overall, 400 to 700 evaluators (M = 550) participated each semester, providing, after data quality control eliminations, a minimum of eight ratings per video, with most having 10 or more.

The video rating survey focused on the evaluator's assessment of the communicator *as* a communicator, rather than testing for content understanding in the evaluator after the communication. The survey tool included 16 items using 7-point Likert-type scales (1 = *strongly disagree*; 7 = *strongly agree*) about the clarity (six items, e.g., "The presentation was clear"), engagement (six items, e.g., "The speaker seems enthusiastic about the subject"), and credibility (four items, e.g., "The speaker seems knowledgeable about the topic") of the presenter. These items were developed specifically for this

Table 1. Video Rating Items.

Question category	Question number	Item
Clarity	1	The presentation was clear
	2	The presentation was easy to follow
	3	The speaker used confusing terms
	4	I felt confused at one or more points during the presentation
	5	The speaker used examples and/or analogies to improve my understanding of the information
	6	I was distracted by the speaker's lack of fluency (e.g., pause, stuttering, repetitions, etc.)
Engagement	7	The speaker seems enthusiastic about the subject
	8	The speaker kept my attention
	9	I am more interested in this subject after watching this talk
	10	I want to know more about this subject
	11	The speaker used nonverbal communication (e.g., facial expressions, gestures, body language) that enhanced the presentation
	12	I was distracted from the presentation by the speaker's nonverbal communication (e.g., facial expressions, gestures, body language)
Credibility	13	The speaker seems knowledgeable about the topic
	14	The speaker is likable
	15	The speaker made the subject seem important
	16	The subject is relevant to my interests

Note. All items rated on a 7-point Likert-type scale.

study, based on evaluations of effective speech communication used in public-speaking courses at the university, and with reference to the National Communication Association's competent speaker speech evaluation guidelines (National Communication Association, n.d.). See Table 1 for the rating questions we used. Students were also asked to state in open-ended items what they did and did not like about a presentation.

Self-Assessment

Pre-and post-training self-assessments by trainees are often used to assess change in trainee belief in their own ability ("self-efficacy"). Since our interest was solely in whether self-assessments accurately reflected performance, as judged by audiences, we did not ask trainees to complete pretraining self-assessments. In order to assess whether self-assessments align with those of outside evaluators, we asked trainees to complete self-assessments of their skills in communicating scientific information to a public audience at the end of each semester. In the Fall 2017 and Fall 2018 semesters, students ($N = 19$) completed 12 items asking them to rate their confidence in successfully accomplishing a variety of communication tasks on a scale of 0 (*cannot do at all*) to 10 (*highly certain can do*). We aligned these items with rating items given to evaluators of videos where applicable. Two items ($r = .64$) correspond to the video rating items about clarity: "I can avoid barriers to communication (e.g., jargon, incorrect framing)," and "I can have a respectful conversation with a non-scientist who disagrees with me." Three items correspond to the video rating items about engagement ($r = .69$): "I value the opinions of public audiences," "I can engage a public audience," and "I can engage a public audience via social media." Two items ($r = .73$) corresponded to the video rating items about credibility: "I can describe scientific research results for public audiences," and "I can adjust my communication to the proper level for my audience." The remaining five items were self-reflection items about expectations and satisfaction (e.g., "What were your expectations for this course?" "Were your expectations fulfilled?"), which had no equivalents in the items for the video ratings.

We converted the 10-point scales used for the self-assessment items to 7-point scales for comparison to the evaluators' video rating scores on trainee "after" videos. Because trainee responses on self-assessments were completely anonymous, direct comparisons of the evaluators' scores for a particular trainee to that trainee's own self-evaluation was not possible. We, therefore, calculated median ratings for the self-assessments of all trainees of the clarity, engagement, and credibility items, and compared them with the median ratings by evaluators of these sets of items for the "after" videos of trainees in the Fall 2017 and 2018 courses ($N = 18$ trainees).

Data Formatting

We downloaded raw survey data from the Qualtrics portal. We removed all responses that had answered the "speed bump" question incorrectly to ensure that we only included data from evaluators who were paying close attention to the survey. We also removed all incomplete surveys, and any in which the evaluator responded "no" to either of the postvideo questions "Could you see the presentation?" and "Could you hear the presentation"? To visualize and analyze the scores as consistently ranked from positive to negative for each question, we reversed the order of Likert-type scores on questions in which high scores represented more negative evaluation: Clarity-related Questions 4, 5, and 6, and engagement-related Question 12 (Supplemental Table S1). To ensure the anonymity of the trainees and controls, we assigned a unique identifier for each individual that encoded whether the student was in the experimental group or control, the semester, and the year.

Analysis

Ordinal data, such as those measured on a Likert-type scale, can be misleading when analyzed as if they are metric (Liddell & Kruschke, 2018); the data are not continuous, since participants cannot choose values on the scale between whole numbers, and evidence suggests that participants do not necessarily perceive (or use) the difference between score values as equivalent along the length of the whole scale (e.g., the difference between a score of 2 and 3 vs. the difference between a score of 6 and 7; Liddell & Kruschke, 2018). Additionally, we have many nested observations in this data set (e.g., multiple answers per question, multiple evaluators per video). Both of these features are best represented by a hierarchical generalized linear mixed-effects model in a Bayesian statistical framework (as described in Bürkner & Vuorre, 2019).

Our model (see Item S1 for the complete model equation) assumes that the Likert-type scale measures an unobserved, continuous variable (i.e., the degree of the agreement an evaluator felt for a particular question about a particular video). This "latent variable" is assumed to be normal, and broken into discrete Likert-type values at specific points. The precise values of these breakpoints *to the evaluators* are estimated from the data during the modeling process, rather than treated as an a priori assumption. The hierarchical structure of the data set is captured with random effects. This means that we model the average response and then allow individual members of the different groups to vary around it. For example, we estimate an average response for all questions and then allow every specific question to depart from this average by some amount. These departures (sometimes called "offsets") are assumed to come from their own normal distribution, centered on 0 and with an estimated standard error. These standard errors are also estimated from the data; the smaller they are the more consistent are individuals within groups (i.e., the more closely they follow the group average). We fit this model using R and the package "brms" (Bürkner 2017) to model the scorers' assessments of subject videos and visualized model results using ggplot2 (Wickham, 2016), tidybayes (Kay, 2019), and used the colorblind accessible color palette from colourblindR (Flores et al., 2019).

Our model follows recommendations for analyzing ordinal response variables as described in Bürkner and Vuorre (2019). Specifically, the model estimates six breakpoint values among the different response categories. This allows the model to reflect the nonmetric nature of Likert-type responses, that is, a response of "6" is not necessarily twice as high as a response of "3." We also measure the average effects of two variables and their interaction: time of year (start or end of the semester), stage of training (before or after the course), and finally their interaction. The interaction term represents our hypothesis test: How much does training improve students, beyond the effects of the mere passage of time?

We also add a combination of random effects (see Supplemental Material for the complete model equation), and this allowed us to test various kinds of nonindependence in our model. Specifically, we included a random intercept for every question category (clarity, engagement, and credibility), allowing each category to differ in average evaluation, and for each semester, to account for nonindependence in time. We also used a random intercept for every evaluator (allowing evaluators to vary between those that mostly disliked or mostly liked the video they viewed). Most important, we fit varying effects (varying intercept, and correlated effects of time, training, and their interaction) for every question and every trainee/control pair. This allows individual questions to respond to time and training independently: For example, it may be possible that only some of the questions we asked accurately measured student learning. The varying effects for pairs are important because, depending on their background, members of a pair may have on average higher or lower average evaluations, or the trainee in a pair may respond to training to lesser or greater degrees. Pairs were chosen to be homogeneous based on training, ESL status, gender, and other external factors; thus, this random effect conditions our estimate of the overall effect on all these factors.

Additionally, as a check on our methods, and to examine whether a more conventional approach to the analysis would yield different results, we analyzed the same data using a generalized linear mixed model, using SAS software (Version 9.4 for Windows, Copyright © 2013 SAS Institute Inc.) for each question individually. We designated each individual scorer as a random effect in the model, with an interaction between "Before & After" and "Trainees vs. Controls" giving an estimate of the average amount of change of the trainees and controls over the course of the semester. We present the results of these analyses in Supplemental Table S1.

Results

Our results show that science communication training had virtually no effect, on the time scale of the training itself, on communication skills; that trainees overestimate the degree of improvement training makes on their own communication skills; and that rigorous assessment of science communication training will require grappling with enormous variation in what audiences consider “good” communication.

Our intensive, semester-long training in scientific communication resulted in no greater improvement of trainees’ communication skills than that of controls who received no training at all (Figure 1). While the average scores of trainees did improve compared with themselves before training, controls also had improved scores at the end of the semester, as compared with themselves at the beginning (Figure 2). Therefore, the difference in improvement between trainees after the course and controls at the same point at the end of the semester (i.e., *the improvement attributable to the training itself*) was not only slight but too slight to conclude that trainees improved more than controls did (Figure 3). The result is robust to analytical approach; when we repeated the analysis of the same data on a question-by-question basis, using more typically employed univariate generalized linear mixed models, instead of our hierarchical Bayesian model described in the “Method” section, we still failed to find any significant difference between the improvement of trainees and controls for any rating question (Supplemental Table S2). Whether the improvement in trainees and controls is simply the result of time (and associated professional growth) or the repetition of the task itself cannot be addressed within our experimental framework, but the actual improvement of trainees, itself, was slight; on average, scores improved only about the equivalent of a one fifth to one quarter of a Likert-type response value across all questions, for all trainees and across all evaluators (Figure 2).

Variance in the scores given by evaluators was very high (Figure 4), with no obvious pattern in the data with respect to trainees versus controls, and with variance in responses to most rating questions spanning most, or all, of the Likert-type scale. Variance was not only high with respect to how evaluators rated trainees versus controls; variance in the scores for individual subjects (trainees or controls) was similarly high; even in the cases of the individuals with the highest and lowest median scores, respectively, evaluators did not agree on a question-by-question basis on the scores (Figure 5). We also found no relationship in the average scores, or the variance in scores, to either the subjects’ (trainees or controls) or the evaluators’ genders or ESL status.

In an outcome consistent with other research on self-evaluation of communication competence, on the other hand, trainees rated their own communication effectiveness more highly than did the evaluators (Figure 6). Evaluators rated trainees in terms of their clarity at *Median* = 4.76 (M = 4.83,

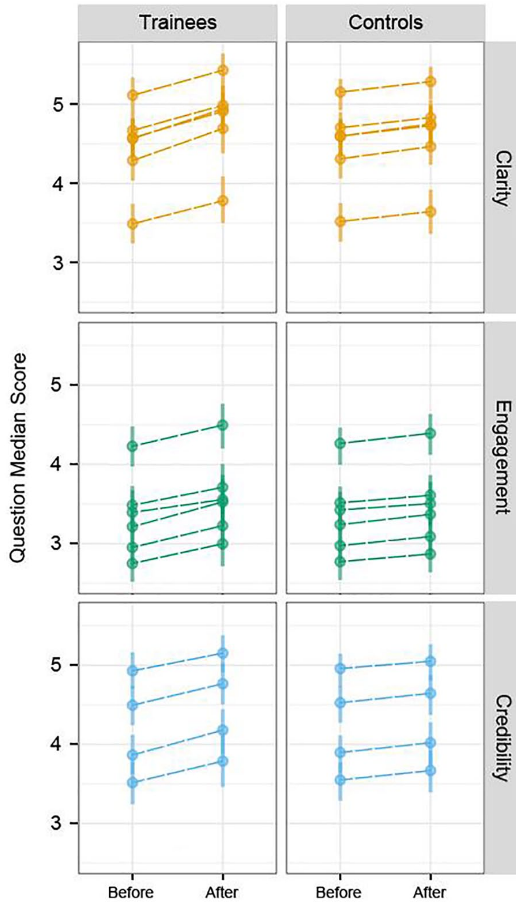

Figure 1. Communication performance as a function of training, or time. Posterior median scores given by evaluators, in response to questions about videos of communicators, grouped by area of assessment (clarity, six questions; engagement, six questions; and credibility of the presenter, four questions; see Supplemental Table S1 for questions) for all trainees (left-hand panels) and controls (right-hand panels). Dots show the posterior median for each question, and the vertical bar around each dot shows the 89% posterior density. Dotted lines connect the median response for the same question before and after training in science communication; controls were scored without any training, after the training period. All question scores reflect the improvement in the performance of both trainees and controls after the training period; trainees in the class exhibited only a slightly greater increase in scores.

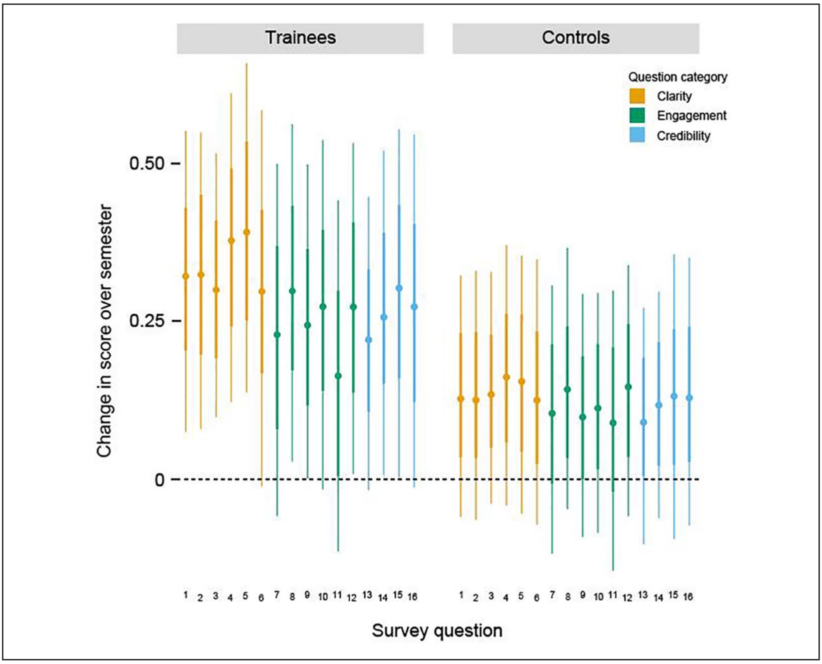

Figure 2. Effect of training or time on communication scores. The *x*-axis locations correspond to questions assessing communication videos (see Supplemental Table S1 for specific questions), and the *y*-axis shows the magnitude of the change in scores after training (trainees) or time (controls), on the same scale as the scores themselves (Likert-type scale of 1 to 7). Points are (posterior) median values of scores; thin lines show 95% posterior density, and thicker lines show 67% posterior density. While on average, scores of both trainees and controls increased, and trainee scores increased slightly more, note that improvements, and differences in improvement, are measured in only fractions of a single Likert-type scale value. Lines overlapping the zero line are statistically equivalent to no change. Questions are colored according to which category of scoring question they cover.

SD = 0.81). However, the trainees in these videos rated themselves on clarity at *Median* = 5.60 (*M* = 5.53, *SD* = 0.97). Similarly, for ability to engage an audience, the evaluators' ratings came to a *Median* = 3.40 (*M* = 3.56, *SD* = 0.65), whereas the trainees rated themselves on engagement at *Median* = 5.37 (*M* = 5.32, *SD* = 0.94). Finally, credibility showed the same pattern, with evaluators' ratings at *Median* = 4.30 (*M* = 4.41, *SD* = 0.70), while trainees rated themselves at *Median* = 5.95 (*M* = 5.91, *SD* = 0.78).

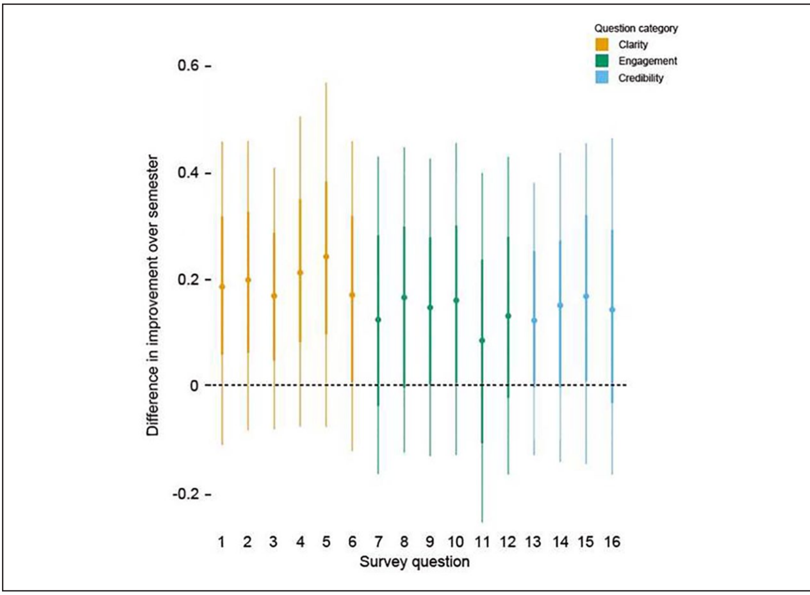

Figure 3. Improvement of scores in trainees, relative to controls. The *y*-axis is the difference in the magnitude of change in scores of trainees versus controls (increase in trainee scores—increase in control scores), on the scale of the scores themselves (Likert-type scale of 1 to 7). Each *x*-axis location is a question for assessing communication videos (see Supplemental Table S1 for the questions), and the questions are colored according to which broad category they cover. Points are posterior median values; thin lines show 95% posterior density, and thicker lines show 67% posterior density. Note that differences in improvement in communication scores between trainees and controls are, on average, equivalent to less than one fifth of a single Likert-type scale value; lines overlapping the zero line are considered statistically equivalent to no difference between trainees and controls.

Discussion

The critical question in evaluating the effectiveness of any form of training is not whether trainees learn, but *whether they learn more than they would have learned on their own* without training. This question is particularly salient when significant time and money are being expended to provide and take training courses.

Whatever content knowledge may be gained during training, science communication is a practice, and the ultimate arbiters of success are audiences. Our study is the only one of which we are aware in which the effect of science

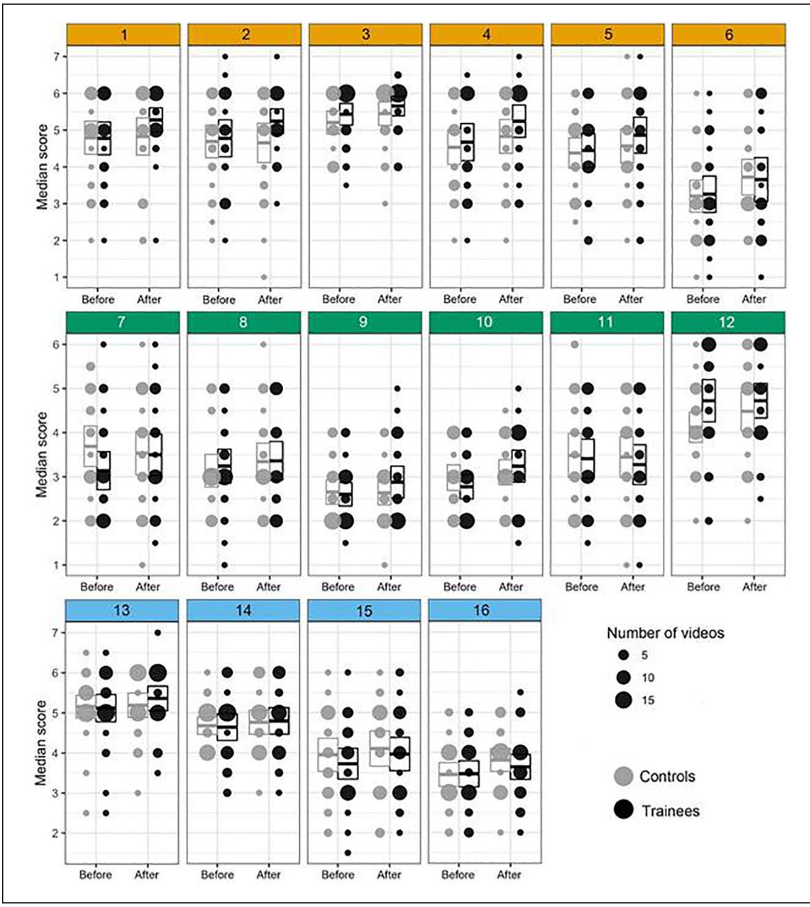

Figure 4. Variation in scores among communicators. Panels are numbered to correspond to questions for assessing communication videos (see Supplemental Table S1 for questions). *y*-Axis values represent the score values on a 1- to 7-point Likert-type scale. Dots represent the median score a video received for a given question during the 3-year period of the study. The size of each dot is in proportion to the number of videos with that score for that question. Colors show whether videos were made by trainees (black) or controls (gray). Boxes are bootstrapped confidence intervals of the median of the medians for each question. Within each panel, we show scores for videos made before training, on the left side, and after training (or after the training period, for controls), on the right. The questions are organized into three categories, asking about the clarity of the presentation (top row, yellow header), the engagement of the presenter (middle row, green header), and the credibility of the presenter (bottom row, blue header). Variation in scores is high, with no obvious pattern with respect to training versus control. A color version of this figure is available online at: https://journals.sagepub.com/doi/figure/10.1177/1075547020971639

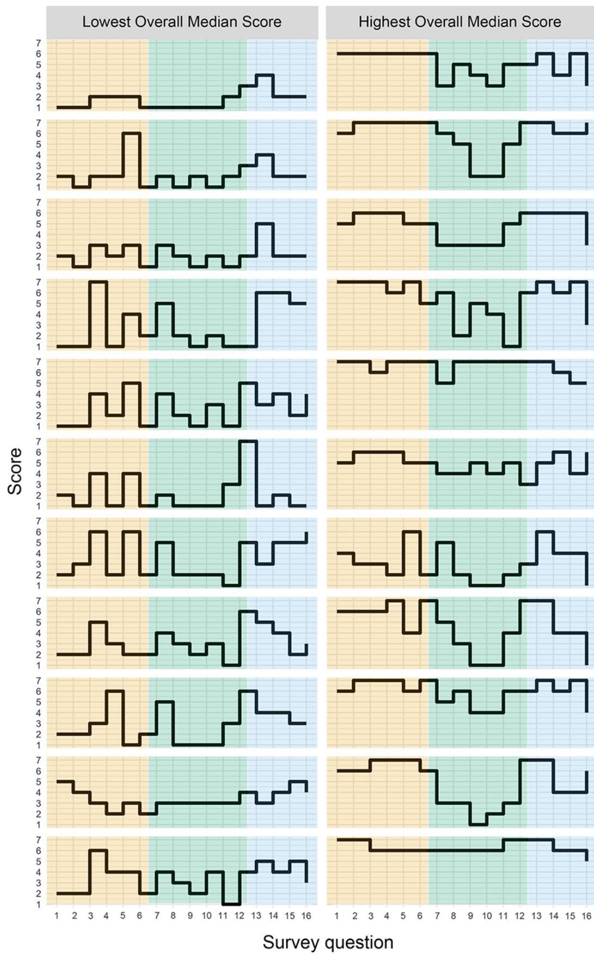

Figure 5. Evaluators do not agree about the skill of communicators. Variation in the scores given by different evaluators to the communication video with the lowest median score (left) and the highest median score (right). Each panel (top to bottom) shows a different evaluator's scores on those videos; the *y*-axis on each panel corresponds to the score given by that evaluator, on a 7-point Likert-type scale; and the *x*-axis hatches correspond to the 16 evaluation questions (see Supplemental Table S1 for questions). Variation is very high, both for a given question (e.g., the scores for Question 3 range from a low of 2 to a high of 7, with scores for every value in between represented) and across questions (no question exhibits low variation).

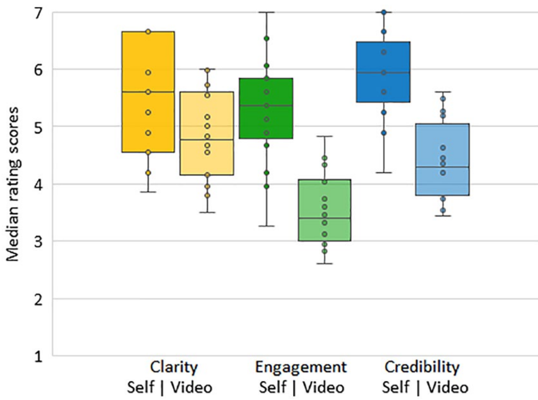

Figure 6. Trainees self-evaluate themselves more highly than evaluators do. Median scores for clarity (left side of each panel; yellow), engagement (center of each panel; green), and credibility (right side of each panel; blue); trainee self-evaluation scores on left, paler evaluators' scores of communication videos on right. *y*-Axis values are on a 1- to 7-point Likert-type scale. Central lines represent medians of ratings. Trainees self-evaluate themselves more highly than do evaluators across all areas of evaluation, and by magnitudes greater than the change in evaluators' scores before and after training. A color version of this figure is available online at: https://journals.sagepub.com/doi/figure/10.1177/1075547020971639

communication training on the ability of trainees to communicate with an audience, as judged by that audience alone, is measured directly while rigorously controlling for factors other than the training itself. Our results strongly suggest that even an intensive, semester-long, active-learning training program using what are widely viewed as best educational practices has little effect, in real-time, on improving science communication skills. The skills of students who took our course did improve over the course of the semester; specifically, evaluators rated students' ability to present information with clarity more highly after training than before (Figure 2). However, the average degree of improvement in trainees was small (about the equivalent of one quarter of a single Likert-type score value), and the overall improvement in other communication skills did not differ from zero. Perhaps more important, the few gains in skills that trained students made were only slightly greater than those made by students who were not trained at all, suggesting that the training itself had little effect (Figure 3).

On the face of it, this result is hard to believe. Our trainees were advanced graduate students who invested months of time (estimated at a total of 75 hours of in- and out-of-class), attention, and committed effort to learn to identify the problems in their own, and others', communication styles, planning for communication with journalists and other public audiences, and practicing actually communicating complicated technical information in a clear and engaging way. Further, as individuals who had to request permission to enter the class, they were a self-motivated sample of those whose science communication skills we might wish to improve. It is difficult to accept that such training had little effect on their practical skills, as far as an audience might be concerned; all of the course instructors would have rated most trainees as significantly improved by the end of the semester. It is even more difficult to accept that control students, who were not trained at all, exhibited nearly as much improvement in scores from the evaluators as our trainees did.

What is the possibility that these results are simply wrong, and that the effect of training is somehow obscured? Our sample size of trainees is limited, as an inevitable corollary of an intensive training; if the variation in skill gains among trainees is large, or factors other than performance are influencing scores, then a few performers with little improvement in scores could have a large impact on the apparent mean of improvement of the group as a whole. Using controls, not only drawn from the same graduate student population but also matched to the trainees for discipline, year of the program, gender, and ESL status, allowed us to reduce the possibility that factors other than performance would obscure real gains in skill. While there was indeed variation among the most improved and least improved trainees in our sample (Figure 2) for most scoring questions, the range of that variation was no more than about the equivalent of half a Likert-type score value, a small degree of variation on a 7-point scale. We are confident that if our sample size obscures a real training effect, it is likely so small as to be of little practical difference with respect to the impact of the training on trainees. We found no effect of gender or ESL status on the likelihood of improved scores of either trainees or controls, and no effect of the gender of the evaluator on the scores they gave (Supplemental Figure S1).

What about the possibility that we were teaching the "wrong things"? The course we built for this experiment was informed by the most widely used science communication books, and the most recent literature addressing the communication of science by scientists at the time the course was designed (including Baron, 2010; Dean, 2009; Menninger & Gropp, 2008; Olson, 2009; and multiple authors in National Academy of Sciences [Fischhoff & Scheufele, 2013, 2014]; see Supplemental Information for our syllabus). Students were assigned readings from the above, and engaged in active learning exercises on identifying and removing jargon from their speech; identifying, and identifying with, audiences; message refinement; the use of metaphors and analogies; the use of narratives (storytelling) instead of explanation; and the nature and constraints on the work of journalists, in particular.

The only well-known training technique we did not use was improvisation, which we viewed as outside our collective formal expertise and experience. However, every practice interview our trainees participated in was an exercise in uncontrolled exchange (i.e., not a lecture) with a partner whose expertise and outlook was very different.

One possibility that requires consideration is that the trainers, themselves, were ineffective. As in every other endeavor, there is variation in the performance of those who teach, and if we are less skilled than we believe we are, then we might expect our trainees to fail to improve. What evidence do we have that the trainers, themselves, were competent to train students to communicate science? To the extent that experience matters, all three of the course instructors were experienced with both the content and teaching pedagogy. One instructor (MR) is herself an alumnus of the widely respected COMPASS training associated with the Leopold Leadership Fellowships, has been teaching at the university level since 1998, science communication to graduate students formally since 2006, and won a university-wide teaching award in 2016. She is also an active researcher in avian biomechanics whose work has received considerable press coverage, and as the CT State Ornithologist speaks frequently to reporters and public audiences. The other two instructors were former newspaper reporters who have been teaching, part- or full-time, at the university level for a combined total of more than 45 years; one (RW) is the author of the most widely used Environmental Reporting textbook, and the other (RC) is the winner of a Pulitzer Prize in explanatory journalism, who subsequently obtained a PhD in ecology and evolutionary biology. All three are highly rated in student evaluations of teaching, both in general and in the courses run for this experiment (although we acknowledge that student evaluations have been demonstrated to have little relationship to measures of actual student learning; Uttl et al., 2017).

Finally, the course itself was created through a rigorous, months-long process in consultation with an Harvard-trained education specialist, who ensured that we identified and worked backward from course goals to create structure, active learning, and formative and summative feedback mechanisms, and who monitored all but a few of the class meetings in person in order to provide adaptive teaching feedback to the instructors. While it is still possible, despite all of the above, that the instruction in our course itself was somehow lacking, we think it is unlikely that poor instruction is a plausible explanation for the overwhelming lack of differentiation between trainees and controls. Perhaps the more important question to ask is: If the qualifications and preparation of our instructors for this course were insufficient to produce greater skill development in trainees over a 15-week course, how likely is it that shorter trainings administered by trainers who have no formal education in teaching will produce better results?

Given that our stated goal at the outset of this project was to develop an assessment method, it could be that our training succeeded, but the assessment metrics we developed did not. Our assessment method is predicated on the idea that audience response to a communication is the only metric that matters; nonetheless, if the audience is asked the wrong questions, the response may not be reflective of whether a communicator succeeded or failed with the audience. The questions used in the survey tool that evaluators responded to are provided in Supplemental Table S1. We designed the content and form of the questions in an iterative process with the communication and education specialists on our team, both of whom are experienced in the use of surveys in research. The questions were designed to assess major conceptual areas considered fundamental in science communication, and in communication generally (clarity, engagement, and credibility). We could have asked additional, and more specific questions, but considered a longer more detailed question set more likely to go unfinished by evaluators, and possibly more likely to be leading or ambiguous (e.g., the response to "Did the speaker use jargon?" would have depended on whether the evaluator responding was familiar with the jargon, as an evaluator who was a STEM major might have been). While it is possible that we failed to ask a question or questions that would have better differentiated trainees from controls, we think it unlikely that if they were, in fact, significantly different in their communication performance that all of the questions we asked would have failed to reflect that difference also.

If these questions were insufficient to detect a difference in "good" communication practices between trainees and controls, is it possible that widely held ideas about what constitutes "good" communication are simply wrong, and therefore, we are measuring the wrong things? A striking result of our work is the lack of agreement among evaluators; variation in scoring was very high (Figure 4) across both trainees and controls. We might expect that if "good communication" were universally recognizable—if we know it when we see it—then evaluators of any single video would tend to agree—to give similar scores—even if the variation among videos was high. Even if Likert-type scoring is a difficult tool with which to repeatably measure the performance of a mediocre communicator (is middling performance equivalent to a score of 3 or 4 or 5?), we would expect variance to be low when the performance of the communicator was either particularly good or bad. If there is agreement about what constitutes effective communication, any particular group of evaluators should tend to give a good communicator high scores, and a bad communicator low scores, even if performance among the communicators varies widely overall. Thus, we would expect scores for individual videos to vary less than the scores among videos. Instead, variation at the level of the individual videos is as high as the variation among videos, and even the subjects with the lowest and highest scores exhibit wide score variation (Figure 5). While undergraduates at a public university are not a homogenous audience, they are also not "the general public." They have similar ages, level of education, concerns, and a shared vernacular. We might reasonably expect less variance than we found in their response to "good" and "bad" communicators. This suggests that even a relatively narrow audience does not agree on what constitutes "good communication," a significant problem for the goal of establishing a rigorous assessment framework that allows us to compare the relative value of different training approaches.

If this result is real, what does it mean? Is it impossible to train scientists to communicate successfully, or to assess those training attempts? Is the apparent success of training programs the result of the facilitation of those who already have an affinity and talent for communication? Some of the most pointed-to examples of successful science communicators (e.g., Carl Sagan, Neil deGrasse Tyson) never had training, per se, other than what came from repeated self-directed attempts at communication, and the resulting successes and failures.

We believe that the latter point is likely to be an important locus for future research into what makes science communication training effective. Even in an intensive course like ours, between our "before" and "after" tests, each trainee had no more than two opportunities to practice a complete sequence of planning a communication, delivering it, reviewing it, and reflecting on their own strengths and weaknesses in order to improve the next attempt, however informed by reading, review of communication attempts of others, discussion and exercises (i.e., content knowledge) those attempts may have been.

Successful science communication would seem to be a complex integration of a number of skills; along with the skills typically taught during trainings, a successful science communicator, in practice, has to attend and respond to the particular circumstances and feedback from an audience in real-time. Every encounter provides information about what works and does not work, to be drawn on in future communication attempts ("deliberate practice," Ericsson et al., 1993). New communication tasks require the ability to apply what is already known in a different context. One possible explanation of our results is, simply, that trainees need more repetition putting what they have learned cognitively into practice than even an entire semester affords them. Another is that trainees require more opportunities to apply their conceptual knowledge to a greater diversity of communication tasks before they can perform well outside the structure of a class. It was beyond the scope of this research to investigate whether trainees show greater gains in communication skills than controls over longer periods of time, posttraining; a fruitful area of future research would be to directly measure performance gains as a function of the number of attempts at communication, and as a function of the number of novel communication tasks they have experienced.

Our results make a strong case for the importance of direct, external assessment of science communication training models through measurement of the impact on an audience, rather than self-assessments by trainees, or by personnel administering the training. They also demonstrate, as have numerous other studies, how misleading it can be to rely only on trainee self-assessment to assess the value of a particular training approach or course. If we are serious about helping scientists succeed at communicating information that is crucial to informed policy and public welfare, we will need to reconsider how training is assessed, and quite possibly the nature of the training itself. Given that both our time to make a crucial difference in the public sphere on subjects like climate change and our resources are limited, the programs and agencies providing the funding for lectures, workshops, and longer trainings—not to mention the scientists devoting time to those trainings—should carefully weigh the evidence about the nature and size of the impact resulting from their investments.

Acknowledgments

We thank Jae Eun Joo for her extensive and expert assistance with course and survey design; Paul Lyzun for his assistance with video production; Stephen Stifano for giving permission and facilitating our use of the University of Connecticut communication research pool for this study; Scott Wallace for graciously learning our class design, and filling in; Todd Newman for his work in program support in the early stages of the project; and the Departments of Ecology and Evolutionary Biology and Journalism at the University of Connecticut for teaching releases to MR and RW in support of this project. The Dean of the College of Liberal Arts and Sciences and the Office of the Vice Provost for Research at the University of Connecticut contributed funding that made possible the assistance of Dr. Joo and Mr. Lyzun.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded by National

Science Foundation NRT-IGE award 1545458 to MR, RW, and RC. Permission for human subjects research was granted by the University of Connecticut Institutional Review Board, Protocol No. 016-026.

ORCID iDs

Margaret A. Rubega https://orcid.org/0000-0002-2747-8114
Kevin R. Burgio https://orcid.org/0000-0002-8375-2501
A. Andrew M. MacDonald https://orcid.org/0000-0003-1162-169X
Anne Oeldorf-Hirsch https://orcid.org/0000-0002-3961-3766
Robert S. Capers https://orcid.org/0000-0002-9563-4625

Supplemental Material

Supplemental material for this article is available online at http://journals.sagepub.com/doi/suppl/10.1177/1075547020971639

References

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility. *Journalism & Mass Communication Quarterly*, *93*(1), 59-79. https://doi.org/10.1177/1077699015606057

Baram-Tsabari, A., & Lewenstein, B. V. (2013). An instrument for assessing scientists' written skills in public communication of science. *Science Communication*, *35*(1), 56-85. https://doi.org/10.1177/1075547012440634

Baram-Tsabari, A., & Lewenstein, B. V. (2017). Science communication training: What are we trying to teach? *International Journal of Science Education, Part B*, *7*, 285-300. https://doi.org/10.1080/21548455.2017.1303756

Baron, N. (2010). *Escape from the ivory tower, a guide to making your science matter*. Island Press.

Besley, J. C. (2015). What do scientists think about the public and does it matter to their online engagement? *Science and Public Policy*, *42*(2), 201-214. https://doi.org/10.1093/scipol/scu042

Besley, J. C., Dudo, A., & Storksdieck, M. (2015). Scientists' views about communication training. *Journal of Research in Science Teaching*, *52* (2), 199–220. https://doi.org/10.1002/tea.21186

Bray, B., France, B., & Gilbert, J. K. (2012). Identifying the essential elements of effective science communication: What do the experts say? *International Journal of Science Education, Part B*, *2*(1), 23-41. https://doi.org/10.1080/21548455.2011.611627

Bürkner, P.-C. (2017). brms : An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77-101. https://doi.org/10.1177/2515245918823199

COMPASS Science Communication, Inc. (2017). *The message box workbook*. https://www.compassscicomm.org/

Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, *111*(Suppl. 4), 13614-13620. https://doi.org/10.1073/pnas.1320645111

Davis, J. J. (1995). The effects of message framing on response to environmental communications. *Journalism & Mass Communication Quarterly*, *72*(2), 285-299. https://doi.org/10.1177/107769909507200203

Dean, C. (2009). *Am I making myself clear? A scientist's guide to talking to the public*. Harvard University Press.

Druschke, C. G., Reynolds, N., Morton-Aiken, J., Lofgren, I. E., Karraker, N. E., & McWilliams, S. R. (2018). Better science through rhetoric: A new model and pilot program for training graduate student science writers. *Technical Communication Quarterly*, *27*(2), 175-190. https://doi.org/10.1080/10572252.2018.1425735

Dudo, A. & Besley, J. C. (2016). Scientists' prioritization of communication objectives for public engagement. *PLoS ONE*, *11*(2), 1–18. https://doi:10.1371/journal.pone.0148867

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*(3), 69-106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Duran, R. L., & Zakahi, W. R. (1987). Communication performance and communication satisfaction: What do we teach our students? *Communication Education*, *36*(1), 13-22. https://doi.org/10.1080/03634528709378637

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363

Eva, K. W., Cunnington, J. P. W., Reiter, H. I., Keane, D. R., & Norman, G. R. (2004). How can I know what I don't know? Poor self assessment in a well-defined domain. *Advances in Health Sciences Education: Theory and Practice*, *9*, 211–224.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*(4), 395-430. https://doi.org/10.3102/00346543059004395

Fischhoff, B., & Scheufele, D. A. (2013). The science of science communication. *Proceedings of the National Academy of Sciences of the U S A*, *110*(Suppl. 3), 14031-14032. https://doi.org/10.1073/pnas.1312080110

Fischhoff, B., & Scheufele, D. A. (2014). The science of science communication II. *Proceedings of the National Academy of Sciences of the U S A*, *111*(Suppl. 4), 13583-13584. https://doi.org/10.1073/pnas.1414635111

Fiske, S. T., & Dupree, C. (2014). Gaining trust as well as respect in communicating to motivated audiences about science topics. *Proceedings of the National Academy of Sciences of the U S A*, *111*(Suppl. 4), 13593-13597. https://doi.org/10.1073/pnas.1317505111

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, *77*(3), 515-540. https://doi.org/10.1177/107769900007700304

Flores, I., Tse, S., & Boyce, H. (2019). *ColourblindR: An R package that creates themes that make plots accessible for people with colour blindness*. https://ubc-mds.github.io/ColourblindR/

Gruppen, L. D., Garcia, J., Grum, C. M., Fitzgerald, J. T., White, C. A., Dicken, L., Sisson, J. C., & Zweifler, A. (1997). Medical students' self-assessment accuracy in communication skills. *Academic Medicine*, *72*(10, Suppl. 1), S57-S59. https://doi.org/10.1097/00001888-199710001-00020 (Published erratum appears in *Academic Medicine*, 1997, 72(12), 1126)

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research*, *52*(1), 123-142. https://doi.org/10.3102/00346543052001123

Kay, M. (2019). *tidybayes: tidy data and geoms for Bayesian models*. https://zenodo.org/record/3238563

Kuehne, L. M., Twardochleb, L. A., Fritschie, K. J., Mims, M. C., Lawrence, D. J., Gibson, P. P., Stewart-Coaster, B., & Olden, J. D. (2014). Practical science communication strategies for graduate students. *Conservation Biology*, *28*(5), 1225-1235. https://doi.org/10.1111/cobi.12305

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328-348. https://doi.org/10.1016/j.jesp.2018.08.009

Lynch, M. P. (2017). Teaching humility in an age of arrogance. *Chronicle of Higher Education*, *64*. https://www.chronicle.com/article/Teaching-Humility-in-an-Age-of/240266

Lynch, M. P., Johnson, C. R., Sheff, N., & Gunn, H. (2016). Intellectual humility in public discourse. *IHPD Literature Review*. https://humilityandconviction.uconn.edu/wp-content/uploads/sites/1877/2016/09/IHPD-Literature-Review-revised.pdf

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709-734. https://doi.org/10.2307/258792

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, *32*(2), 344-354. https://doi.org/10.5465/amr.2007.24348410

McCroskey, J. C., & McCroskey, L. L. (1988). Self-report as an approach to measuring communication competence. *Communication Research Reports*, *5*(2), 108-113. https://doi.org/10.1080/08824098809359810

Menninger, H., & Gropp, R. (2008). *Communicating Science: A primer for working with the media*. American Institute of Biological Sciences.

Mort, J. R., & Hansen, D. J. (2010). First-year pharmacy students' self-assessment of communication skills and the impact of video review. *American Journal of Pharmaceutical Education*, *74*(5), 78. https://doi.org/10.5688/aj740578

Morton, T. A., Rabinovich, A., Marshall, D., & Bretschneider, P. (2011). The future that may (or may not) come: How framing changes responses to uncertainty in climate change communications. *Global Environmental Change*, *21*(1), 103-109. https://doi.org/10.1016/j.gloenvcha.2010.09.013

National Communication Association. (n.d). *Learning outcomes & assessment*. https://www.natcom.org/academic-professional-resources/teaching-and-learning/learning-outcomes-assessment

Olson, R. (2009). *Don't be such a scientist: Talking substance in an age of style*. Island Press.

Peterman, K., Evia, J. R., Cloyd, E., & Besley, J. C. (2017). Assessing public engagement outcomes by the use of an outcome expecations scale for scientists. *Science Communication*, *39*(6), 782–797. https://doi:10.1177/1075547017738018

Rodgers, S., Wang, Z., & Maras, M. (2018). Decoding science: Development and evaluation of a science communication training program using a triangulated framework. *Science Communication*, *40*(1), 3-32. https://journals.sagepub.com/doi/pdf/10.1177/1075547017747285

Rodgers, S., Wang, Z., & Schultz, J. C. (2020). A scale to measure science communication training effectiveness. *Science Communication*, *42*(1), 90-111. https://doi.org/10.1177/1075547020903057

Rowe, G. & Frewer, L. J. (2005). A typology of public engagement mechanisms. *Science*, *Technology and Human Values*, *30*(2), 251-290. https://doi:10.1177/0162243904271724

Schemer, C., Kühne, R., & Matthes, J. (2014). The role of measurement invariance in comparative communication research. In M. J. Canel & K. Voltmer (Eds.), *Comparing political communication across time and space* (pp. 31-46). https://doi.org/10.1057/9781137366474_3

Smith B, Baron N, English C, Galindo H, Goldman E, McLeod K, et al. (2013). COMPASS: Navigating the rules of scientific engagement. *PLoS Biol*, *11*(4): e1001552. https://doi.org/10.1371/journal.pbio.1001552

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.

Author Biographies

Margaret A. Rubega is a professor in the Department of Ecology and Evolutionary Biology at the University of Connecticut, where she teaches graduate courses on Science Communication. She was principal investigator on a National Science Foundation NRT-IGE grant funding research on quantitative assessment of communications training approaches, which funded this project, and is a Co-PI on recently funded NSF NRT grant. Her other research uses mechanistic, quantitative approaches and tools from functional morphology, biomechanics, physiology, and animal behavior to explain why birds look, live, and act as they do.

Kevin R. Burgio is broadly interested in the processes that form, and limit, where species are distributed, and their roles are in communities and ecosystems. He uses an integrative approach to examine historical ecology, community assembly, the effects of climate change on communities, and extinction. Additionally, he conducts research in the fields of science communication and education (webpage: https://kevinburgio.com/).

Robert S. Capers is a botanist who has retired from the University of Connecticut herbarium, where he was the plant collections manager. He continues to do research on the dynamics of alpine plant communities and the distribution of bryophyte species in northeastern North America. Before obtaining a PhD in botany, he worked for more than 20 years as a newspaper editor and writer.

A. Andrew M. MacDonald has worked on plant-insect interactions, invasive species, and metacommunity ecology. He is currently a postdoctoral researcher at Université de Montreal, Canada. He works on experimental community ecology, statistics, and data synthesis.

Anne Oeldorf-Hirsch is an associate professor in the Department of Communication at the University of Connecticut, where she conducts research in the Human-Computer Interaction lab. Her research focuses broadly on the use of social media for information sharing as a form of communication. Specifically, her work investigates the effects of using social and mobile media to engage with news, health, and science content. She teaches courses on media effects, communication technology, and social media.

Robert Wyss is a retired professor of journalism at the University of Connecticut. He is the author of three books and has written extensively about environmental issues both at his time at the university and as a newspaper writer and editor. He has taught courses in news writing and editing, the history of journalism, environmental journalism, and science communication.