Supplemental material, part 5

"Speaking About Science: A Student-Led Training Program Improves Graduate Students' Skills in Public Communication"

Statistical analysis for assessment of speaking skills

Sections "b. External assessments" and "c. Comparison of self and external assessments" correspond to sections within the Results in the main paper.

All data analysis was performed with R statistical software, version 3.2.3 (R Core Team, 2015), using custom scripts.

b. External assessments

With multiple reviewers per video, it is important to evaluate the consistency of scores, and by extension, the reliability of external reviewers as a standard for assessing students' skills in speaking about their research. Calculating Fleiss' κ (Fleiss & Cohen, 1973) for intercoder reliability, we found that though the scores for each video were more consistent than expected by random chance, there was still high variability in scores per video ($\kappa = 0.191$, $p = 4.29\text{E-}11$). This lack of agreement suggests the potential that a student's skill estimation could reflect the assignment of their reviewers rather than their actual skill in each core competency. To account for this variability in reviewer evaluation, we adopted a mixed-effects regression model, which is described in the following section.

Data analysis

For analysis of external assessment data we used mixed-effects, ordinal, logistic regression models (Agresti, 2013). This choice of model accommodates the ordered nature of the rating data and allows us to analyze the influence of Engage training while accounting for the variability in assessment scores introduced by our study design. Ordinal logistic regression frames the response of a study as a sequence of odds (probability of an event occurring over not). In our study, these were the odds of being equal or below a given rating over being above that rating.

Ordinal regression for the probability $P(Y_i \leq j)$ of a sample *i* being equal or below rank *j*, is characterized by one of a set of logistic regression equations of the basic form:

$$\begin{gathered} \boldsymbol{logit}[\boldsymbol{P}(\boldsymbol{Y_i} \leq \boldsymbol{j})] = \boldsymbol{\theta_j} - \boldsymbol{\beta X_i} \\ \boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{J} - \boldsymbol{1} \end{gathered} \tag{1}$$

where Y_i is the ordinal response variable for the *i*-th sample; J is the highest attainable rank; $j = 1, \ldots, J - 1$ indicates the rank that regression equation j is evaluating; $logit\,[P\,(Y \leq j\,)]$ is the log odds of $Y \leq j$ as opposed to $Y > j$; θ_j is the intercept term associated with the threshold between rank j and rank $j + 1$; $\boldsymbol{\beta}$ is a vector of regression coefficients for fixed effects; and $\mathbf{X}_i$ is the vector of predictor variables for the *i*-th sample. Note that $P\,(Y_i \leq J\,)$ equals 1.

Consequently, the probability for the response variable to take on any particular rank *j* is:

$$\begin{aligned} \boldsymbol{P}(\boldsymbol{Y_i} = \boldsymbol{j}) &= \boldsymbol{P}(\boldsymbol{Y_i} \leq \boldsymbol{j}) - \boldsymbol{P}(\boldsymbol{Y_i} \leq \boldsymbol{j} - \boldsymbol{1}) \\ \boldsymbol{P}(\boldsymbol{Y_i} = \boldsymbol{j}) &= \boldsymbol{logit^{-1}}[\boldsymbol{\theta_j} - \boldsymbol{\beta X_i}] - \boldsymbol{logit^{-1}}[\boldsymbol{\theta_{j-1}} - \boldsymbol{\beta X_i}] \end{aligned} \tag{2}$$

To evaluate the effect of Engage training, the pre-course versus post-course status of videos was included as a (fixed-effect) predictor in our regression model. We wanted to account for potential biases in reviewers' scores, pseudo-replication in study design, and variation in students' overall ability to deliver effective presentations. Thus, we included students' identity and reviewers' identity into our model as random effects, which are typically used to account for underlying group variation and biases (Moulton, 1986). Thus, our final model was:

$$\begin{gathered} \boldsymbol{logit}[\boldsymbol{P}(\boldsymbol{Y_i} \leq \boldsymbol{j})] = \boldsymbol{\theta_j} - \boldsymbol{\beta} \cdot \boldsymbol{PrePost_i} - \boldsymbol{u}(\boldsymbol{StudentID_i}) - \boldsymbol{u}(\boldsymbol{ReviewerID_i}) \\ \boldsymbol{j} = \boldsymbol{1}, \boldsymbol{2}, \boldsymbol{3} \quad \boldsymbol{u}(\boldsymbol{x_i}) \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_x^2}) \end{gathered} \tag{3}$$

where $PrePost_i$ is the status of a student's Engage training, $u(StudentID_i)$ is the random effect of student identity, and $u(ReviewerID_i)$ is the random effect of reviewer identity. Both random effects are assumed to be normally distributed with mean zero and variance to be estimated via regression. Like the self-assessments, we evaluated the impact of Engage training on each core competency separately with an alpha level of 0.01.

Results

For each of the five assessment items, we found that the science communication training had a significant, positive influence on the odds that a student's post-course video was scored higher than his or her pre-course video, except for the self-confidence metric (**Table S5-1**, α = 0.01). This is consistent with the results from the self-assessment. External reviewers saw the most improvement in students' ability to take their audience into consideration.

This analysis allows us to quantify the impact of the Engage course. For example, the value of 3.71 associated with the competency of audience consideration means that, all else being equal, the log odds of a student with Engage training scoring well on this competency is 3.71 higher than a student who has not had the training. This is equivalent to a 41-fold increase in the student's odds of scoring well.

Table S5-1. Ordinal regression coefficient estimates from external assessment data

Core competency	$\sigma^2_{StudentID}$	$\sigma^2_{ReviewerID}$	Engage Impact		*p*
			Mean	Standard Error	
Audience consideration	0.57	0.90	3.71	0.70	< 0.001
Distillation	0.52	0.65	3.10	0.63	< 0.001
So what	1.18	0.00	2.60	0.53	< 0.001
Storytelling	1.50	0.51	2.91	0.60	< 0.001
Self-confidence	4.06	0.64	1.33	0.52	0.011

Impact estimates are presented on the log odds scale
σ^2 values are the variances in student ability and reviewer bias

c. Comparison of self and external assessments

Data analysis

Students may be biased (e.g. overly self-critical) in assessing their own communication skills. To evaluate these biases, we performed a post-hoc comparison of students' self-assessment scores versus the scores predicted by ordinal regression models given an average reviewer with minimal bias. To obtain the predicted scores for each student, we determined the ranking j that maximized the model-predicted probability $P\,(Y_i = j\,)$ for a given student (Equation 2), omitting the effect of reviewer identity on video scores. We used sign tests to examine differences between pairs of self-assessment scores and predicted external-assessment scores for each assessment item of pre-course and post-course videos. To mitigate the chance of Type 1 errors when making these 10 comparisons of self- and external-assessment scores, we used the Bonferroni correction and set our alpha-level at 0.05/10 = 0.005.

Results

For the competencies of audience consideration, distillation, and storytelling, there was no significant difference between self-assessment scores for pre-course videos and the predicted external-assessment scores, but post-course self-assessment scores were significantly lower than that of external assessments (**Figure S5-1**, $\alpha = 0.005$). When assessing the competency of self-confidence, students' evaluations of their pre-course videos were significantly more critical than that of their external reviewers, but there were no significant differences between self- and external-assessments for post-course videos. Finally, we found no significant differences between self- and external-assessments of students' abilities to convey the "so what" of their research, either before or after the Engage course.

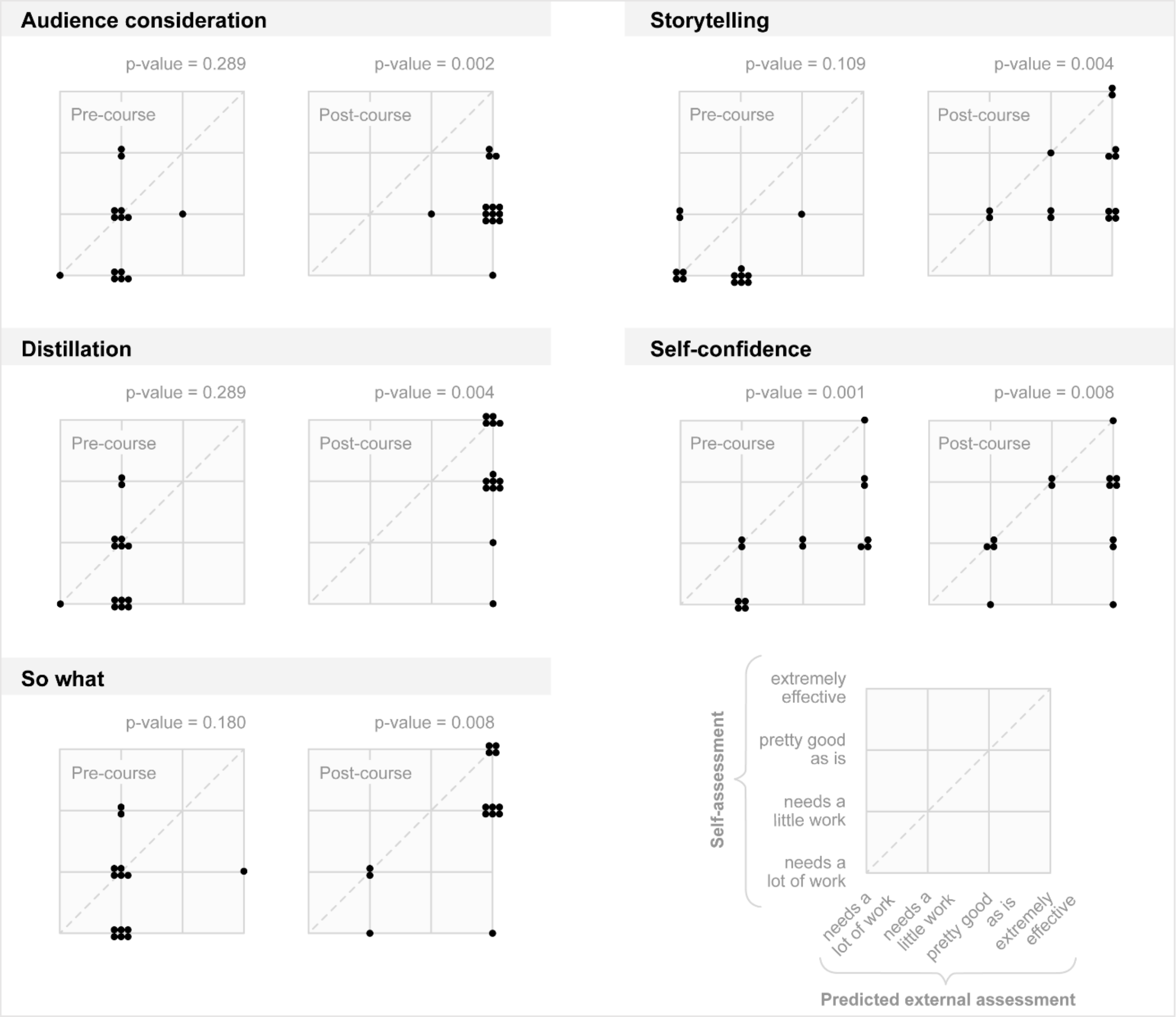

Figure S5-1: Comparison of self-assessment scores with predicted external-assessment scores. For each of the five assessment items, plots are shown for pre-course and post-course videos. Each dot represents a single student's video, with the self-assessment score plotted on the vertical axis and the score predicted based on our modeling of the external-assessment scores on the horizontal axis. If the scores are the same, the dot will appear on the diagonal. Videos plotted above the diagonal were rated higher by the student than our model predicted, while those below the diagonal were rated lower by the student than predicted. p-values are from sign tests comparing self-assessment scores and predicted external-assessment scores.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed). Hoboken, NJ: Wiley.

Fleiss, J., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures fo reliability. *Educational and Psychological Measurement*, *33*(3), 613–619.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, *32*(3), 385–397. https://doi.org/10.1016/0304-4076(86)90021-7

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing.