Abstract Title Page
Not included in page count.

Title: Instrumental Iteration Toward FOI Evaluation of Pedagogical Methods

Authors and Affiliations:

Manuel S. Gonzalez Canche, The University of Georgia
Pedro R. Portes, The University of Georgia
Paula J. Mellon, The University of Georgia
Robert A. Stollberg, The University of Georgia
Jonathan M. Turk, The University of Georgia

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

The current study is part of the core research agenda of the five-year, *Goal 3* grant funded by the Institute of Education Sciences, U. S. Department of Education *Improving the Teaching and Learning of English Language Learners: The Instructional Conversation Model.*

In video review of teaching, bias in evaluation or rating is evident in many studies. Some of it is inevitable given the use of video review for professional development purposes and the reflexive nature of that rating process (Carolan & Wang, 2012; Cheng-Chih & Hue-Ching, 2008; Collins, Cook-Cottone, Robinson, & Sullivan, 2004; Ostrosky, Mouzourou, Danner, & Zaghlawan, 2013; Tripp & Rich, 2012). Other sources of bias are driven by competition for resources (pay raises in a district) or stakes of reputation. A recent report from the Bill and Melinda Gates foundation (Ho & Kane, 2013) details recent US trends toward video review of teachers and takes a step forward in determining reliability and revealing transparency of biases. The discovered bias inherent was present on two levels: teachers were permitted to select which video lessons were reviewed, and administrators and peers were often acquainted with those reviewed. The authors listed several impressions garnered from the ratings data including: peer teachers tended to rate lessons more favorably than administrators, same-school administrators generated higher teacher ratings, and ratings tended toward the 'middle' in that high and low scores were infrequently cited. Bias in selection of participants and judges can be difficult to avoid in the arena of K-12 education. Unfortunately, attempts to mediate this bias further via instrument selection and revision have been poorly documented.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

This study utilizes an iterative process to refine an existing evaluation rubric for rating pedagogical intervention in recorded 3rd and 5th grade lessons across subject types. The rubric in question was originally designed to determine instructor adherence to the Instructional Conversation (IC) model of pedagogy (Dalton, 2007) in control and experimental groups as part of IES Grant # R305A100670. Fortunately, selection bias referenced by Ho & Kane is minimized by enlisting raters which have no connection to the teachers or districts participating. Upward bias is diminished by a 'no stakes' review consisting of only scoring IC components and by not issuing an assessment of the quality of teaching present. Yet, the bias inherent in repeat viewings of rated media remains intact (Tinsley & Weiss 1975; 2000). The goal of our reliability methodology is to meet research standards via a three-stage analysis - leading to instrument refinement and reliability measures.

Setting: > N/A

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features, or characteristics.

While the original grant is clear in terms of population dispersed throughout the three districts, it is more apt to suggest that the participants in this secondary study are the video raters themselves, a team of University of Georgia faculty and graduate students dedicated to reaching interrater reliability.

Significance / Novelty of study:
Description of what is missing in previous work and the contribution the study makes.

There is a dearth of literature regarding instrument development and establishment of reliability in the assessment of in-class teacher performance. The end result of this study is a nuanced attempt to develop a process of instrument development with the anticipated outcome of greater interrater reliability in determining fidelity of implementation (FOI) of pedagogical interventions. Reliability measures and factor analysis determine the efficacy of the instrument and result in a refined evaluation tool. While the final instrument is of note only in the context of the IC intervention, the process of instrument refinement offers a template for future FOI assessments as well as future use in determination of reliability in evaluation of teaching from recorded video.

Statistical, Measurement, or Econometric Model:
Description of the proposed new methods or novel applications of existing methods.

The Tetrachoric Correlation
The tetrachoric correlation (Pearson, 1900), for binary data, and the polychoric correlation, for ordered-category data, are excellent ways to measure rater agreement (Bonnet & Price, 2005). The data upon which the analysis is conducted consisted on 20 items. If raters considered that an item attribute was present they marked the item and this item was given a value of one. Items that were not marked were coded as zero. The resulting 20 x (number of raters) binary matrix allowed conducting tetrachoric correlation coefficient which expresses rater association.
The tetrachoric correlation is the inferred Pearson Correlation from a n x n table with the assumption of bivariate normality (Brown, 1977). Although these statistics make certain assumptions that cannot be tested with the tetrachoric correlation if there are only two raters (Digby, 1983). This limitation is overcome in our study given that there are more than two raters.

Factor Analysis with Tetrachoric Correlation
An important part of the purpose of this study is to provide a list of items (or scale) that can be reliably implemented by other researchers using video rating. This scale is particularly useful for the analysis of fidelity of implementation of instructional conversation instruction methods. To achieve this goal, we conducted factor analysis using the raw raters' responses. This dimension reduction mechanism will allow us to potentially decrease the number of items that may be used in future research an in the second and third waves of implementation of our current grant. Tetrachoric and polychoric correlations can be factor-analyzed in the same way as Pearson correlations (Lieberman, 1970). The first step if to construct a matrix of tetrachoric correlation coefficients. We use this matrix as the input for our factor analysis. As with any other factor analyses models the use of matrix rotation to better identify/differentiate factors is recommended. In this study we rely on Varimax rotation.

Cohen's Kappa

In addition to actual ratings per each rater, we have an aggregate measure that adds their total number of marked items, thus constitution a scale that ranges from 0 to 20. Assuming that raters consistently assessed teachers' performance, we would expect there to be a high level of reliability between any given numbers of rater scores that assessed a teacher. To this end we performed a Cohen's Kappa test for interrater reliability. Cohen's Kappa is a coefficient designed to measure the degree of agreement in nominal scales and is an improvement over measures of average percent agreement between reviewers because it also considers when agreement occurs by chance. This coefficient provides a means for testing hypotheses and setting confidence limits for this coefficient. This method of testing interrater reliability compares observed agreement between raters with the level of agreement that is expected to occur randomly by analyzing the following two quantities The coefficient *k* is the proportion of agreement after both chance agreement and chance disagreement were removed from consideration (Cohen, 1960, p. 40). From the above equation we can infer that $k = 0$ when obtained agreement equals chance agreement. This result is particularly interesting because it indicates that raters' decisions were no more consistent than we would expect based on chance (Stemler, 2001). Greater than chance agreement leads to positive values of *k* and rare cases in which less than chance agreement occurs leads to negative values. In this regard the upper limit of $k = 1.00$ (when there is perfect agreement between the raters) and the lower limit is $k = -1.00$ (when there is perfect disagreement between the raters). When is $k = 0.00$ the level of agreement between raters is exactly equal to hypothetical chance agreement. Stemler (2001) recommends affirming strong interrater reliability when $k = 0.61$.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.
(May not be applicable for Methods submissions)

The original video rating instrument was a 20 item checklist of yes/no binary outcomes. A total of 25 videos were rated by 7 raters. Raters were instructed to check an item should compelling evidence of its existence be shown in the rated video –the greater the number of checked items, the greater the evidence that an IC had taken place during the lesson. The Tetrachoric correlation and factor analysis are based on these 25 videos rated. For interrater reliability the number of videos analyzed is seven given that the same subject (video) was required to be rated by a group of judges. In this case 4 raters scored 7 videos in common. The main goal of including this coefficient is to test for rater reliability of the scale analyzed with an innovative method (tetrachoric correlation and factor analysis).

Findings / Results:

Description of the main findings with specific details.
(May not be applicable for Methods submissions)

The tetrachoric correlation found is 0.5086 (p. <. 001) expressing overall rater association. The importance of this analysis, however, relies in observing which items are highly correlated. For example, we observed (See Table 1) that raters' identification of teacher encouraging student conversation is highly correlated with raters' identification of actual student dialogue (rho= .80). However, the former is uncorrelated (rho= .003) with raters' identification of teacher listening to

assess student levels of understanding. This implies that the promotion of dialogue does not necessarily impact teachers' implementation of the listening component of the IC. Yet, the identification of listening is correlated with (rho= .43) the identification of teachers' connecting lesson content to student experiences, which is an important component of the IC model.

The factor analysis with tetrachoric correlation rendered nine factors which potentially among 14 items (Table 2). It is worth highlighting that five items were dropped from the model due to collinearity and one was not included as it showed not variation (see notes in Table 3). For space constrains we limit this analysis to the interpretation of the first factor. This factor was composed of three items, teachers: (1) allow students to articulate for themselves, (2) encourage students to speak in the IC, (3) encourage students to build on their comments and questions and is expressing the creation of single item that captures dialogue in IC. It is important to notice that dialogue was one of the items dropped from the model. This may imply future revisions of the scale may build a single item that captures the four identified with tetrachoric factor analysis.

The Kappa coefficient of four raters analyzing seven videos was 0.586 ($p < 0.001$). This result indicates that raters responses were correlated above chance and that the evaluation of the IC implementation cab be reliably assessed with the current version of the scale.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

The results indicate that raters' responses were correlated above chance occurrence and that the evaluation of the IC implementation can be reliably assessed with the current version of the scale. However, modification to the instrument based on the findings from the tetrachoric correlation and factor analyses may improve interrater reliability and provide a better assessment of fidelity of implementation.

The conclusion reached is that a goal of interrater reliability for teaching assessment purposes can be reached through advanced statistical methods to refine earlier processes. The primary limitation is not in the use of these methods but in the inability of schools, districts, etc. to minimize bias in the rater selection process. The UGA experiment benefitted greatly from rater participants with no vested interest in the individual outcomes or pressure to inflate scores.

The next video ratings will be assessed using the modified scale resulting from this study. This scale will be validated and submitted to peer review in order to assess its contribution to the evaluation of fidelity of implementation using video ratings.

Appendices

Not included in page count.

Appendix A. References

Bonett, D. G., and Price, R. M. (2005). Inferential methods for the tetrachoric correlation coefficient. Journal of Educational and Behavioral Statistics, 30(2), 213-225.

Brown, M. B. (1977). Algorithm AS 116: The tetrachoric correlation and its asymptotic standard error. Journal of the Royal Statistical Society. Series C (Applied Statistics 26,3,343-351.

Carolan, L., & Wang, L. (2012). Reflections on a transnational peer review of teaching. ELT Journal, 66(1), 71-80.

Cheng-Chih, W., & Hue-Ching, K. (2008). Streaming videos in peer assessment to support training pre-service teachers. Journal of Educational Technology & Society, 11(1), 45-55.

Collins, J.L., Cook-Cottone, C.P., Robinson, J.S., & Sullivan, R.R. (2004). Technology and new directions in professional development: Applications of digital video, peer review, and self-reflection. Journal of Educational Technology Systems, 33(2), 131-146.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Dalton, S.S. (2007). Five standards for effective teaching: How to succeed with all learners, grades k-8: John Wiley & Sons.

Digby, P. G. (1983). Approximating the tetrachoric correlation coefficient. Biometrics, 753-757.

Ho, A., & Kane, T. (2013). The reliability of classroom observations by school personnel The MET Project: The Bill and Melinda Gates Foundation.

Lieberman, M. (1970). Fitting a response curve model for dichotomously scored items. Psychometrika, 35, 179-198.

Ostrosky, M.M., Mouzourou, C., Danner, N., & Zaghlawan, H.Y. (2013). Improving teacher practices using microteaching: Planful video recording and constructive feedback. Young Exceptional Children, 16(1), 16-29.

Pearson, K. (1900) Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Philosophical Transactions of the Royal Society of London, Series A, vol. 195, pp. 1-47.

Stemler, S. (2001). An overview of content analysis. Practical Assessment, Research & Evaluation, 7(17).

Tinsley, H.E., & Weiss, D.J. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology, 22(4), 358.

Tinsley, H.E., & Weiss, D.J. (2000). Interrater reliability and agreement. Handbook of applied multivariate statistics and mathematical modeling, 95-124.

Tripp, T., & Rich, P. (2012). Using video to analyze one's own teaching. British Journal of Educational Technology, 43(4), 678-704.

Appendix B. Tables
Table 1. Tetrachoric correlation

adj-corr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. arrange	1																		
2 articulate	-0.49	1.00																	
3 acadgoal	0.43	-0.02	1.00																
4 cohesive	0.57*	0.11	0.55*	1.00															
5 tallow	0.61*	0.05	0.49*	0.55*	1.00														
6 tencourage	0.08	0.36	0.35	0.53*	0.28	1.00													
7 dialog	0.26	0.36	0.62*	0.55*	0.60*	0.80*	1.00												
8 build	0.06	0.76*	0.36	0.65*	0.45	0.54*	0.59*	1.00											
9 assist	0.39	-0.06	0.66*	0.72*	0.52*	0.41*	0.61*	0.35*	1.00										
10 listen	0.31	0.05	0.40*	0.48*	0.27	0.00	0.05	0.37*	0.58*	1.00									
11 tsorts	0.18	0.19	-0.02	0.34*	0.34	0.08	0.06	0.43*	0.23	0.45*	1.00								
12 connect	0.16	-0.12	0.07	0.10	0.25	-0.19	-0.05	0.07	0.06	0.31	-0.06	1.00							
13 priorexp	0.61*	-0.03	0.45*	0.45*	0.31	0.25	0.44*	0.32	0.38*	0.43*	0.38*	-0.02	1.00						
14 models	0.09	0.10	0.26	0.29	-0.14	0.15	0.09	0.32	0.05	0.27	0.22	0.17	0.39*	1.00					
15 vocab	0.18	-0.11	-0.06	-0.20	0.24	-0.09	0.11	-0.09	0.12	0.22	0.04	0.22	0.21	0.16	1.00				
16 encvocab	0.52	0.17	0.20	0.38	0.75	0.15	0.39*	0.41*	0.23	0.37*	0.22	0.45*	0.39	0.13	0.53*	1.00			
17 sclarify	-0.49	0.67*	0.31	0.09	0.10	0.21	0.31	0.56*	0.21	0.28	0.22	0.07	0.07	0.23	-0.10	0.10	1.00		
18 reflect	0.51	0.20	0.75	0.57	0.72	0.38*	0.59*	0.62*	0.48*	0.45*	0.50*	-0.03	0.59*	0.21	0.03	0.44*	0.33*	1.00	
19 hypo	0.02	0.501*	0.19	0.27	0.23	0.41*	0.55*	0.56*	0.35*	0.26	0.30	-0.21	0.56*	0.30	0.34	0.38*	0.49*	0.40*	1.00

Note *p < .05

Table 2. Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Uniqueness
articulate	0.9324									0.0643
tencourage	0.4760			0.4208						0.3826
build	0.9050									0.0000
assist				0.9749						0.0000
listen					0.6997					0.2146
tsorts					0.5677					0.5481
connect								1.0270		0.0000
priorexp		1.0453								0.0000
models									0.9743	0.0000
vocab						0.9633				0.1756
encvocab						0.5195				0.2429
sclarify							0.9350			0.0000
reflect			0.9395							0.0000
hypo										0.1620

(blanks represent abs(loading)<.4)

Table 3. Video Rating Instrument
Video Name: _____ Date Rated: _____ Score: __/20
Rater's Name: _____

	Notes
1. The students participating in the lesson are in a small group consisting	Dropped because non-variant
2. The arrangement of students and teacher is conducive to conversation.	Dropped because of collinearity
3. The lesson or activity allows students to articulate for themselves.	
4. The conversation is guided by a clear academic goal.	Dropped because of collinearity
5. The students are working as a cohesive group.	Dropped because of collinearity
6. The teacher allows students to speak in the conversation. (Doesn't shut	Dropped because of collinearity
7. The teacher encourages students to speak to each other.	
8. The students are engaged in dialogue.	Dropped because of collinearity
9. The teacher encourages students to build on the questions and	
10. The teacher assists students in staying focused on the lesson topic and	
11. The teacher listens to assess the level of understanding on the part of	
12. The teacher sorts through student misconceptions as they arise.	
13. The teacher helps students connect the content of the lesson to their	
14. The teacher probes into students' prior academic knowledge and	
15. The teacher models academic language that relates to the lesson. (Uses appropriately)	
16. The teacher introduces new vocabulary and language skills.	
17. The teacher encourages students to use their newly acquired	
18. The teacher asks students to clarify and explain their thinking.	
19. The teacher asks students to reflect on their own misunderstandings	
20. The teacher asks the students to make hypotheses, inferences, and	