Evaluation of an Online Mentoring Program

By Sharon Sherman & Gregory Camilli

In this article, the evaluation of an online mentoring program for preparing pre-service elementary teachers at a small liberal arts college is described. An intervention was created to investigate the effects of online mentoring with pre-service teachers, where mentoring is defined as a reciprocal relationship formed between an experienced teacher and a novice. This relationship is designed to provide ongoing support, advice and feedback during transition into the teaching profession (Andrews & Martin, 2003; Haney, 1997). According to Lloyd, Wood & Moreno (2000), policymakers in many states mandate or recommend mentoring for novice teachers during the first year of service. Such programs can potentially have a positive effect on both novice and experienced teachers and lead to greater retention (Boreen, Johnson, Niday & Potts, 2000), especially if a mentor is selected based on a set of competencies and trained to develop specific skills needed to provide student support (Brown & Kysilka, 2005; Haney, 1997).

A secondary purpose of the article is to describe an efficient procedure for collecting and scoring rubric-based instruments because scoring performance outcomes is labor-intensive, time-wise and financially, even with small-scale studies. Observational or rating-scale data are required in many educational settings. Most of the instruments used in teacher evaluation systems require rating of observa-

Sharon Sherman is dean of the School of Education at Rider University, Lawrenceville, New Jersey. Gregory Camilli is a professor in the School of Education at the University of Colorado at Boulder, Boulder, Colorado. ssherman@rider.edu & g.camilli@colorado.edu

tional data. The procedures described in this paper illustrate a coherent process for designing an instrument and collecting data in the framework of a comparative study; however, the same procedures could be applied to a single group of ratings. Three important aspects of this procedure are (1) how raters can be incentivized and trained, (2) how to make the most use of the limited availability of raters, (3) and how measurement information concerning the validity of a set of ratings can be obtained. In this paper, the general logistics are described; technical details are provided in a companion paper (Camilli & Sherman, 2013).

Description of the Evaluation Study

In a relatively small state college in the eastern U.S., pre-service teachers enrolled in a junior practicum were assigned, by course section, in roughly equal numbers to a treatment group and to a comparison group. Treatment group members received traditional face-to-face mentoring supplemented by expert online mentors. Control group members received only traditional mentoring face-to-face mentoring.

Students and Assignment to Treatment

In fall 2005 and spring 2006, 108 juniors enrolled in a practicum received mentoring from six university professors and 75 host teachers in seven classes. The sample consisted of 97% female and 3% males. The junior practicum consisted of three three-hour classes per week: integrating and differentiating instruction for all learners; methods of teaching social studies; reading and literacy for middle childhood plus a practice teaching experience. Teacher candidates worked in partnership with another teacher candidate and a host teacher in an elementary school. The practice teaching practicum component included a one and a half day weekly field experience for 12 weeks plus two weeks of full time teaching at the end of the semester.

Seven junior practicum sections took part in the study. Order of registration for junior practicum was determined by the number of credit hours a student earned and then by alphabetical order of last name (from A to Z). Some teacher candidates selected particular college professors, some selected particular host schools, some had no preference and some selected spots in sections with available seats. After registration, sections were randomly assigned to treatment (4) and control conditions (3), and an online communication platform was selected to support student-mentor interaction. Students in all seven sections agreed to participate in the study prior to group assignment.

Online Mentors

Online mentors were sought with documented expertise in mathematics or science content and pedagogical knowledge. For this purpose, a list was developed of teachers who were locally recognized as outstanding, were active in their professional organizations, or who worked in schools with high levels of achievement in science or mathematics. Those teachers were contacted and 23 were eventually recruited.

Each mentor received an honorarium of $150 per student up to a maximum of six students. A total of 23 mentors were recruited, which included elementary, middle and high school teachers.

Each online mentor's responsibilities included attending an hour-long training session on using the online platform; mentoring each student in planning and development of four lesson plans in the mentor's area of expertise; having one face-to-face session with each student; and keeping a log of the support requested and provided. Online mentors communicated with their treatment group mentees via the Internet as a way to share ideas and written documents. Though mentors were recruited from elementary, middle and high schools, pre-service teachers prepared lesson plans for elementary classes only.

The project management team worked with the Internet-based communication corporate team so that the elements of the required lesson plan format appeared on the proprietary software interface. A custom set of all the state's content standards were entered into a database, which facilitated one-click mapping of standards into mentee lesson plans.

Mentoring

All teacher candidates in the control group received face-to-face mentoring from their professors and host teachers. The mentoring focused around preparing lesson plans to foster student learning. Throughout the semester teacher candidates were given instruction in how to write sound lesson plans. They learned the elements of the lesson plan, wrote lesson plans and received feedback from their professors. Students worked in pairs on completing the lesson plans for programmatic rather than experimental purposes (that is, to benefit from collaborative learning).

Students in the treatment sections attended classes regularly and experienced blended mentoring, receiving online mentoring from experienced teachers with content and pedagogy expertise in mathematics and science in addition to face-to-face mentoring from their college professors and host teachers. All students and online mentors received an incentive of six months of access to the platform, and online mentors received an honorarium. Again, students worked in pairs on completing the lesson plans for programmatic rather than experimental purposes (that is, to benefit from collaborative learning). Mentors were assigned for up to six student pairs. Treatment and control group students were asked to submit four mathematics and four science lesson plans, complete a mathematics teaching efficacy instrument, a science teaching efficacy instrument, and participate in a focus group at the end of the semester.

Data

A rubric-scored instrument to assess the quality of lesson plans was developed containing 14 items. This instrument focused on essential aspects of lesson plan preparation such as subject matter, objectives, relevant questions, sharing student work, and grouping of students. Items on the instrument were targeted to the standard lesson plan template provided to the pre-service teachers. About 200 lesson plans were collected and processed for analysis. Of these, about 20 were withheld for training on scoring, and 180 were designated for analysis. Of the latter, 90 lesson plans were available for both the experimental and control groups. A holistic rubric was designed specifically for each item (see Appendix).

Raters and Training

A team of 20 raters was recruited, and two raters were designated as alternatives. The raters consisted of teachers with expertise in mathematics and science, one science education professor, one state department official with a science background, a retired mathematics supervisor and two retired science supervisors. Four of the raters had served as mentors. Each rater was paid $200 for the day.

The raters were contacted prior to scoring and given an agenda. On the day of scoring, raters arrived at the college by 8:30 a.m. for registration and refreshments (lunch was also provided). Before scoring, they received training for about one hour on the lesson plan rubric. The initial goal was to review the rubric with raters, and then to train them relative to a standard using pre-selected teacher candidate lesson plans of varying quality (as rated by the project team). About one hour was set aside for training raters to the standard. The first sample lesson plan was selected by the project team from the pool of submitted lesson plans as an exemplar of excellence. The lesson plan was written for the Everyday Mathematics curriculum and involved teaching tessellations to fifth graders. Ensuing lesson plans were presented that spanned a continuum of quality.

For the first hour the raters discussed the meanings of the descriptors and it was difficult to achieve clarity, so the lead trainer focused on each rubric item separately. Raters considered one criterion at a time, the group rated the lesson plan for that particular item and a group discussion followed. It was not surprising that some raters were high scorers and others were not. The training continued until every rater understood the meaning of the descriptors and all ratings were either in the bottom, middle, or upper third (i.e. exemplary/proficient; proficient/needs improvement; needs improvement/serious concern) of ratings. The training of raters took two full hours.

Design for Scoring

The program's institution donated resources and a working facility for the 20 raters. Based on project team's prior scoring of about 20 training lessons plans, it was estimated a priori that a rater could reasonably score a lesson plan in 15 minutes given the holistic nature of the judgments. An intermediate goal, in order to make this a practical application, was to complete the training of scorers and the scoring in one day. Though 18 raters fit comfortably within the project budget, not enough time was available for all raters to score all lessons. If all raters were to score all lesson plans, this would require a total of 18*180=3240 lesson plan ratings resulting in a total of 810 person hours, or 45 hours per rater. Yet only six hours were available for scoring, given that training was designed to take the first two hours of the session.

Accordingly, a balanced incomplete design was created in which a rater would score 24 lesson plans (12 experimental, 12 control), thus requiring about 6 six hours. Using 18 raters (excluding two alternative raters), this implied a total of 18*24 = 432 lesson plan ratings requiring 108 person-hours, also requiring 6 hours per rater. As shown below, this incomplete design is sufficient to obtain information for analysis. The key in the incomplete design is to distribute raters in a systematic fashion across treatment groups and lessons. This is an attractive alternative to randomly selecting 12 lesson plans for both treatment and control groups, and then having all 18 raters score each of the resulting 24 lesson plans.

In Figure 1, the staggered scoring design is shown that accommodates the constraints described above. First, 18 raters were randomly assigned to nine pairs. Second, lessons were randomly divided into 10 overlapping sets. Each lesson set contained either two lessons or eight lessons. Third, each pair was assigned three lesson sets: one set of eight (non-overlapping) and two sets of two. Each member of a pair would independently read and rate those lessons. For example, in Figure 1, Rater Pair 1 received Lesson Set 1 (2 lessons), 2 (8 lessons), and 3 (2 lessons).

Figure 1

Scoring Design (Repeated for Experimental and Control Groups)

Lesson	*Rater Pair*									*Cum #*
Set	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*of Lessons*
1	2								2	
2	8									
3	2	2								12
4		8								
5		2	2							22
6			8							
7			2	2						32
8				8						
9				2	2					42
10					8					
11					2	2				52
12						8				
13						2	2			62
14							8			
15							2	2		72
16								8		
17								2	2	82
18									8	90

Note that each lesson set of 2 plans overlapped with one other pair of raters. Thus, Lesson Set 1 was scored by Pairs 1 and 9, and Lesson Set 3 was scored by Pairs 1 and 2. Each Rater Pair was assigned three Lesson Sets for scoring from both the experimental and control groups.

In terms of logistics, only the shared lesson sets needed to be duplicated twice; unique sets (seen by only one rater pair) required only one copy. This provided for a minimum amount of time associated with shuffling hard-copy materials in a room filled to capacity. This design could be modified by changing the number of unique or common Lesson Sets for each Rater Pair. For example, the number of unique and common sets could have both been set to four. This would have created greater overlap, but would have resulted in fewer than 180 scored lesson plans as well as slightly more duplication for scoring materials.

Analysis

A statistical model as developed for analyses of these data, and the full technical report is available (Camilli & Sherman, 2013). Below, we consider the effectiveness of the program and the importance of validating the instrument.

Results

The analysis was carried out in three phases. The first phase provides both traditional descriptive statistics and reliability analysis of the instrument. In the second phase, an analysis was carried to obtain information about the reliability of the instrument and simultaneously to estimate what effect the benefit of the treatment had over control (if any). In the third phase, another statistical procedure was carried out to adjust for the initial nonequivalence of the treatment and control groups. Usually random assignment creates equivalent groups at the outset of an evaluation. This is important because the effect of a treatment may obscure one group if it initially has more highly achieving students. A fair comparison requires ability to be about the same. Note that we couldn't determine why the assignment method didn't work as expected, but it may have been due to the small number of sections. Randomization works best with larger numbers.

Descriptive Analyses

In Table 1, the means, standard deviations, and item-total correlations are given for the instrument. Across the 14 items, an average of 2.53 rating was given to lesson plans on a 4-point scale (4=Exemplary, 3=Proficient, 2=Needs Improvement, and 1=Serious Concern). Overall, it did not seem that raters were lenient because the average rating was somewhat less than Proficient. Average item ratings ranged from 3.08 (Developmentally Appropriate Activities) to 1.46 (need to intervene). Based on the latter item, it is clear that many of the raters were satisfied with the lessons.

The rubric-score responses for lesson plans were first analyzed with traditional reliability in mind using the 432 lesson plans as the units of analysis. A Cronbach's α of .873 was obtained. The item-total correlations (or item discriminations) are given in the last column of Table 1. It can be seen that the items in the instrument tend to be moderately to highly cohesive with the total score. Problematic items typically exhibit low or negative item-total correlations, but all observed discriminations were in an acceptable range.

Reliability

As a byproduct of the analysis, information was obtained for estimating two reliability (also called generalizability) coefficients. We estimated the reliability of the lesson plans for four raters as , which is similar to the value for Cronbach's alpha above. Thus, the reliability lesson plans scores averaged across four raters based on this particular set of 14 evaluation items is moderately high. A different perspective is given with inter-rater reliability, computed to be .35, which can be understood as the correlation between two raters on a single item. For comparison, consider the study by Hill, Charalambous, and Kraft (2012). Different dimensions of teacher performance on the Mathematical Quality of Instruction (MQI) observational instrument were examined. With a single rater, an "inter-rater" reliability for one item ranged from .35-.45, and across four lessons, ranged from about .65-.75.

The MQI is a formally developed set of rubrics, as opposed to the locally developed instrument used as an illustration in the present study. Each of the 6-8 MQI "items" in a domain was a 7½ minute segment of a video tape, and two-day training was provided to raters. So it is not surprising that the inter-rater reliabilities in the Hill

Table 1
Item Statistics for Lesson Plan Instrument

Item	*Mean*	*SD*	*Item-Total Correlation*
Q1	2.75	.805	.492
Q2	2.54	.890	.542
Q3	3.01	.622	.615
Q4	2.51	.856	.532
Q5	2.76	.823	.482
Q6	2.89	.744	.633
Q7	3.08	.741	.629
Q8	2.16	.991	.382
Q9	3.02	.883	.562
Q10	2.38	.846	.555
Q11	2.47	.895	.492
Q12	1.79	.900	.365
Q13	2.63	.734	.793
Q14	1.46	.499	.636

et al. study are higher. It should be added that the MQI concerns much more complex behaviors than the instrument in this study, and the level of reliability obtained by Hill et al. is therefore more impressive than the level obtained with the current data.

Effect of Online Mentoring

We obtained an effect size of d=.41 for the treatment (p=.04). This can be interpreted as follows: about 66% of the teacher candidates in the online mentoring group scored higher than the average score in the control group. We noticed, however, the treatment group initially had a mean SAT Verbal 67 points higher than the comparison group, and 35 points higher on mean SAT Quantitative. Moreover, the covariate GPA was moderately correlated with SAT scores in the control group, but not the treatment groups. This indicates that randomization did not work as well as expected (see Kenny, 1975). To control for possible bias resulting from this nonequivalence, we reran the analysis with a subset of the control group that was much more similar to the treatment group initially. The treatment effect increase to d=.70 (p=.02), where about 76% of teachers candidate in the treatment group had higher scores than the average in the control group.

With rubric score rating, some raters are tougher than others. It is thus unfair if the lesson plan for one student is rated leniently and another is rated more strictly. An important aspect of the scoring design and statistical procedure presented in this paper is that because raters are staggered across the lessons plans, rater influences (and potential biases) can be removed from scores given to the lesson plans. Thus, the effect of the treatment is not compromised by different standards being applied in the treatment and control groups.

Discussion

We found that online mentoring did have a moderately strong impact, but a number of factors may have contributed to this outcome. First, students had access to an online platform for organizing their materials and accessing instructional information. Second, pre-service teachers in the blended group experienced online support by mentors that differed from traditionally-recruited mentors for the pre-service program. Online mentors were recognized experts in the content areas of science and mathematics, while traditional mentors varied in expertise. Thus, the effect of online mentoring cannot be disentangled from either the expertise of the online mentors or for that matter, the online software platform. However, the online format intentionally allows flexible access to information and to expert mentors that would not be available in the traditional approach. From a pragmatic point of view, is not clear whether the two distinct effects should be disentangled. At the same time, it should not be expected that the results of this study can be replicated with ineffective online platforms or under-qualified mentors.

The results of this study are important in informing the design of online support in teacher education. The current study is an existence proof that such programs can be effective, and that such programs can be evaluated at a reasonable cost. The study described here may help to shape future comparative studies methodologically. In addition, this methodology need not be restricted to mentoring programs, or even educational experiments. Rather, any scoring-intensive comparative investigation would be a candidate for the strategies offered in this paper.

For example, suppose there are 100 teachers to be evaluated, there are 20 raters total, and each teacher requires two raters. The design offered in this paper suggests how those raters can be distributed across teachers in order to collect both evaluation data, to obtain basic reliability information for validating an instrument, and to make sure that observation scores are fair by removing rater influences. Under many new state accountability systems, teachers are evaluated with both student achievement scores and observational measures. The current study suggests a practical and coherent approach for establishing "reliable and valid classroom observation instruments" (Crowe, 2011) as well as a method for obtaining comparable ratings for different teachers.

References

Andrews, S. P. & Martin, E. (2003). No teacher left behind: Mentoring and supporting novice teachers. Paper presented at the Annual Meeting of the Georgia Association of Colleges for Teacher Education/Georgia Association of Teacher Educators. St. Simons Island, GA. (ERIC Document Reproduction Service No. ED481998).

Boreen, J., Johnson, M. K., Niday, S., & Potts, J. (2000). *Mentoring the beginning teacher: Guiding, reflecting, coaching*. York, ME: Stenhouse.

Brown, S. C., & Kysilka, M. L. (2005). Investigating telementoring with preservice and professional teachers. In F. K. Kochan & J. T. Pascarelli (Eds.). *Successful telementoring* (pp. 185-204). Greenwich, CT: Information Age Publishing.

Camilli, G. & Sherman, S. (2013). A balanced incomplete design for experiments with rubric-scored outcomes (unpublished technical report).

Crowe, E. (2011, March). *Race to the Top and teacher preparation*. Washington, DC: Center for American Progress.

Haney, A. (1997). The role of mentorship in the workplace. In M. C. Taylor (Ed.), *Workplace education* (pp. 211-228). Toronto, Ontario, Canada: Culture Concepts. (ERIC Document 404 573).

Hill, H. C., Charalambos, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*, 56-64.

Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. *Psychological Bulletin, 82*, 345-362.

Lloyd, S. R., Wood, T. A., & Moreno, G. (2000). What's a mentor to do? *Teaching Exceptional Children, 33*(1), 38-42.

Appendix

Instrument with Rubric

	Exemplary	*Proficient*	*Needs Improvement*	*Serious Concern*
	4	3	2	1
1. Description	The description is complete and detailed and includes the big ideas of the lesson.	The description is complete but not detailed. It includes the big ideas of the lesson.	The description is incomplete and lacks detail. It may not include the big ideas of the lesson.	The description is incomplete and does not include the big ideas of the lesson.
2. Hook	The hook includes a discrepant event, demonstration, or activity that is ikely to increase motivation. It is clear that the pre-service teacher has researched possibilities and made a good	The hook includes a discrepant event, demonstration, or activity that is likely to increase motivation.	The hook is included but is not likely to increase motivation.	The hook is missing.
3. Subject Matter Knowledge	The lesson plan shows that the pre-service teacher possesses solid content knowledge and has well-researched the topic.	The lesson plan is free of inaccurate content.	The lesson plan shows little evidence of solid content knowledge or adequate understanding on the part of the pre-service teacher.	The lesson plan shows significant errors in content knowledge.
4. Objectives	The objectives are not too narrow or too broad. They address cognitive, psychomotor, and affective domains.	The objectives may be too narrow or too broad. They address cognitive, psychomotor, and affective domains.	The objectives may be too narrow or too broad. They do not address all three domains.	The objectives are too narrow or too broad. They do not address any of the domains.

	Exemplary	*Proficient*	*Needs Improvement*	*Serious Concern*
	4	3	2	1
5. Challenging Activities	The activities are challenging and may not be solved during the class period. They cause students to leave class thinking about possible strategies and solutions.	The activities are challenging and may not be solved during the class period. They cause students to think.	The activities may require students to use a formula or memorized definition to arrive at a solution but don't challenge students to think.	The activities are not meaningful.
6. Developmentally Appropriate Activities	Activities are well thought out and developmentally appropriate. Problems and activities relate to students' lives and are interesting	Activities are developmentally appropriate. Problems and activities relate to students' lives.	Activities are not appropriate for the age level of the students. Problems and activities don't relate to students lives and may not require much thinking or be interesting	Activities are totally inappropriate based on the developmental level of the students. Problems and activities are inappropriate and don't challenge students.
7. Questions Related to the Topic	The pre-service teacher has listed essential questions that are related to the topic and make students think.	The pre-service teacher has listed essential questions that are not necessarily related to the topic but make students think.	The pre-service teacher has listed questions that have little substance.	The pre-service teacher has listed questions that have no substance. Questions may be missing from the lesson plan.

	Exemplary	*Proficient*	*Needs Improvement*	*Serious Concern*
	4	3	2	1
8. A Series of Questions Leading to Deep Understanding	The pre-service teacher has predicted numerous possible student responses and formulated a series of questions that lead students to generate knowledge and develop deeper understanding. Some questions are at or above the analysis level of Bloom's Taxonomy.	The pre-service teacher has predicted some student responses and formulated a series of questions that lead students to generate knowledge and develop understanding. Some questions are at or above the analysis level of Bloom's Taxonomy.	The teacher has not predicted any student responses or formulated questions that lead students to generate knowledge or develop deeper understanding. The questions are not in a series. Few questions are at or above the analysis level of Bloom's Taxonomy.	The questions do not relate to one another. Questions are not in a series. None of the questions are at or above the analysis level of Bloom's Taxonomy.
9. Assessment Plan	The assessment plan reflects a variety of evaluation strategies including methods of formal, informal, traditional, performance, diagnostic, formative, and/or summative assessment. The pre-service teacher selects the most appropriate assessment strategies for the lesson.	The assessment plan reflects a variety of evaluation strategies including methods of formal, informal, traditional, performance, diagnostic, formative, or summative assessment. The pre-service teacher selects at least one good strategy but there is room for others.	The assessment plan is not well thought out. No appropriate assessment strategies are included. Obvious assessment strategies are missing.	The assessment plan is missing.

	Exemplary	*Proficient*	*Needs Improvement*	*Serious Concern*
	4	3	2	1
10. Using Assessment Information to Plan Future Lessons	There is a sophisticated plan which explains how the pre-service teacher will use assessment information to plan future lessons. It is well thought out.	The plan explains how the pre-service teacher will use assessment information to plan future lessons, but it is not detailed.	The plan states that the pre-service teacher will use assessment information to plan future lessons but doesn't explain how this will be done.	The plan does not include a way to glean information about what the student has learned and what s/he needs to learn in the future.
11. Sharing Student Work	In the closure session students share their work, justify their thinking, and engage in discussion.	In the closure session students share their work.	In the closure session, students review the lesson.	The closure session is missing from the lesson plan. There is no opportunity to share work or review the lesson.
12. Grouping of Students	Specific details for student grouping are provided.	General details for student grouping are provided.	Grouping is mentioned but not clarified.	Grouping is not mentioned in the plan.
13. This Lesson Plan Represents Quality Work from a Pre-Service Teacher.	The pre-service teacher has a complete and detailed understanding of lesson planning.	The pre-service teacher has a complete but not detailed understanding of lesson planning.	The pre-service teacher has an incomplete and/or misconception of lesson planning, however s/he maintains a basic understanding of the process.	The pre-service teacher's understanding of lesson planning is so incomplete or has so many misconceptions that s/he cannot be said to understand lesson planning.
14. Based on This Lesson Plan as Written, I Would Ffeel the Need to Intervene with This Pre-Service Teacher.	No, intervention is not needed.			Yes, intervention is needed.