Intern Performance In Three Supervisory Models

Author notes:

Sid T. Womack is professor of secondary education, Department of Curriculum and Instruction, Arkansas Tech University, Russellville, Arkansas

Shellie L. Hanna is assistant professor of health and physical education, Department of Health and Physical Education, Arkansas Tech University, Russellville, Arkansas

Rebecca Callaway is associate professor of curriculum and instruction, Department of Curriculum and Instruction, Arkansas Tech University, Russellville, Arkansas

Peggy Woodall is associate professor of special education, Department of Special Education, Henderson State University, Arkadelphia, Arkansas

Paper presented at the meeting of the Arkansas Association of Colleges for Teacher Education, Searcy, Arkansas, April 22, 2011.

Abstract

Differences in intern performance, as measured by a Praxis III-similar instrument were found between interns supervised in three supervisory models: Traditional triad model, cohort model, and distance supervision. Candidates in this study's particular form of distance supervision were not as effective as teachers as candidates in traditional-triad or cohort models when overall scores were analyzed. The same trend in lower performance for distance-supervised interns persisted when subscale scores on the Praxis-III similar instrument were examined. These findings prompted us to mostly discontinue distance supervision of interns, at least as distance supervision had been defined for this study.

Key words: Supervision of interns, assessment, NCATE, NCATE Standard Three, novice teachers, distance supervision, distance education.

As is probably true in most teacher education units in the United States, our college of education uses an observation form for assessing intern performance and for giving feedback. *Formative Observation and Intervention Form* for assessing intern performance and for giving feedback. When the *Formative Observation and Intervention Form* was created several years ago, it was noted that the descriptors and domains had a great resemblance to the Pathwise evaluation. According, we obtained written permission from the Educational Testing Service before beginning to use it with our candidates. The observation form is used to collect data on 21 research-based areas of teacher competency/proficiency. Those 21 areas are grouped into four domains of (A) Organizing Content for Student Learning (B) Creating an Environment for Student Learning (C) Teaching for Student Learning (D) Teacher Professionalism. Since the data obtained using the *Formative Observation and Intervention* form are used to make personnel decisions about candidates, we decided to study it in depth, using candidate data from the Spring Semester of 2010. We also compared performances of our intern candidates between three supervision models.

Definitions:

Traditional triad model—a model of intern ("student teaching") supervision in which the intern obtains support and feedback on site from a field-based supervising teacher and a campus-based supervisor. The three together form a triad.

Cohort model—a model of intern supervision in which the intern obtains support and feedback from a group, usually four, field-based supervisors. One of the four supervisors assigns the grade at the end of the internship experience. Direct involvement from the main campus of the

university is considerably less than in the traditional triad model; the cohort of field-based supervisors were hired as adjunct or as visiting professor faculty.

Distance learning model or distance supervision model—an experimental model of intern supervision in which interns obtain support and feedback from a traditional field-based supervisor and a university supervisor who reviews the candidate's lessons by DVD, VHS videotape, or other electronic media. Feedback is then sent back to the candidate, usually by email. There are no on-site visits from university faculty in the distance learning model of supervision.

Purposes of the Study

The *Formative Observation and Intervention Form*, while an important tool, was not among the eight major artifacts under the assessment expectations of the National Council for Accreditation of Teacher Education in our university. Given the findings of another study (in press), a possibility exists that we might want to replace one of our original eight artifacts with this one. We decided to collect data on it for a semester in order to (1) discover the reliability of a formative observation instrument that which, though not a principal artifact of our unit assessment plan, is used for significant decision-making about the continuation of interns; (2) find out if there were any significant differences in the full-scale scores of interns mentored on-site in a traditional triad supervision model, those mentored in cohorts, and those mentored by a new distance learning methodology; (3) determine if there were differences in domain (subscale) scores between interns mentored in the above three methodologies; (4) check for gender bias or gender differences in full-scale scores; and to (5) check for significant differences in the scores of candidates from different academic disciplines.

To help accomplish the purposes of the study, five null hypotheses were developed:

1. There will be no ($p<.05$) correlation between the scores of the odd-numbered and even-numbered competencies of the *Formative Observation and Intervention Form.*

2. There will be no ($p<.05$) difference in the full-scale scores between interns supervised in a traditional triad model, a cohort model, and a distance learning model.

3. There will be no ($p<.05$) differences in the subscale scores between interns supervised in traditional triad models, cohort models, or distance learning models.

4. There will be no ($p<.05$) difference in full scale scores between male and female interns.

5. There will be no ($p<.05$) difference in full scale scores between interns from different majors or licensure programs.

Method

Participants

Participants were 63 early childhood, 9 middle level, and 58 secondary education interns, a total of 130 senior intern candidates. They were assigned to school campuses in the Western part of Arkansas, particularly along the I-40 corridor from Morrilton westward to the Oklahoma-Arkansas state line. All were assigned to accredited public schools and in content areas appropriate to their majors and expected licensures. Placement was done through the office of Teacher Education Student Services at the university. All public school and university faculty who participated in any direct way in their evaluations were thoroughly familiar with the

Pathwise Evaluation System from the Educational Testing Service. All had been through Pathwise training.

Materials and Procedures

Before interns located to their respective placements, they were briefed about the expectations for the field experience. Early childhood majors and middle level majors enrolled in a 16 week course for 15 and 12 semester hours, respectively; secondary majors enrolled in a nine-semester hour course encompassing a 12-week internship. Secondary majors completed an on-campus course in public school law, history and philosophy of education, and content area reading before beginning their 12-week internship. The intent was to make five visits to each intern. One visit was a "hello" visit and was intended to determine if the intern and placement were off to a good start. The four succeeding visits were designated for data (evidence) collection. The *Formative Observation and Intervention Form* was completed on each of the four observational visits. The hypothetical number of observation forms expected for 130 interns would have been 520, but it was not always possible to get observational data on every visit for every intern.

Procedure

Table 1 lists the 21 research-validated items used to evaluate intern performance. Evaluators were asked to mark 1, 2, or 3 while observing a lesson being taught by each intern. Those evaluators were teacher education faculty from the university or trained cohort personnel. All personnel had been required to demonstrate adequate reliability (85 percent agreement with

an expert's ratings) in using the instrument during practice sessions before being allowed to evaluate an intern.

Table 1

Items and subscales on the Formative Observation and Intervention Form

Item		Mean
Subscale: *Domain A, Organizing Content For Student Learning*		14.91
1.	Awareness of student diversity	2.75
2.	Prepare clear learning objectives	2.84
3.	Connect past, present, future content	2.79
4.	Vary methods and materials for learning	2.86
5.	Align learning goals with assessments	2.79
6.	Total preparedness for teaching	2.83
Domain B. Creating an Environment for Student Learning		13.11
7.	Models and promotes fairness	2.84
8.	Rapport with all students	2.86
9.	Challenging learning expectations for all students	2.76
10.	Exercises consistent, appropriate management	2.70
11.	Physical environment, safety	2.90
Domain C: Teaching for Student Learning		13.05
12.	Clear goals & instructional procedures	2.79
13.	Making the content comprehensible	2.84
14.	Critical thinking	2.74

15.	Monitor & adjust, feedback	2.81
16.	Use instructional time effectively	2.79
Domain D: Professionalism		12.50
17.	Reflect on goals met	2.81
18.	Modifications, efficacy	2.79
19.	Build professional relationships	2.88
20.	Parent/guardian communication	2.68
21.	On time, professional appearance, follows school policies	2.90

A rating of 1 signified a lack of knowledge, concern, or effort about that item. A rating of 2 denoted "sufficiently motivated and knowledgeable to perform in classrooms unassisted." A rating of 3 was reserved for "very motivated, very knowledgeable about performance, and performs flexibly and capably in varied classroom situations with all learners." Interns could be recommended for licensure with ratings of "2" in all 21 areas.

In additional to the statistical comparisons that will be made, further note was made about the practical significance of a difference of one on the observation instrument. A candidate who consistently scored 1 ("insufficiently motivated") on any one of the 21 items measured by the Formative Observation would not pass the course or be recommended for licensure. A candidate who scored below 42 (21 items times a minimum acceptable rating of 2) will not usually be recommended for initial licensure. Our usual experience was that candidates who were not going to qualify for licensure will had ones on more descriptors than one.

The data collected in the field, as of the time of their collection, were considered as nominal in character. Although the teacher effectiveness data might have appeared to be nominal or ordinal, they were treated as continuous (interval) for this analysis, given the admonitions of Kerlinger (1973, pp. 159, 181, 440-441) that overly strict adherence to conventions about calculative methods might result in an unnecessary loss of variance. The data set of well over 100 participants was deemed sufficiently large to permit the assumptions inherent in stepwise regression, and SAS did not generate any error messages.

Seven interns were placed in districts some distance from the university and beyond the usual geographical bounds of traditional or cohort supervision. This was because of special needs of the interns. Communication between university supervisors and interns was done through phone calls, email, Skype, faxes, and DVDs of lessons, sent through the Postal Service. The intent was to provide the field experience while accommodating the financial and other needs of the interns who requested those placements.

Results

Data from 460 observations from 130 candidates were obtained during the spring semester of 2010. These occurred as faculty or clinical practice instructors completed four cycles of evaluations while observing interns in teaching situations. The completed observation forms were submitted to the Office of Teacher Education Student Services immediately after each observation.

Artifact Reliability The first purpose of our study was to check the reliabilities of the full scale instrument and its subscales (domains) . The split-half reliability coefficients was 0.967 with 416 pairs of data, $p<.0001$. The C domain (subscale) had the lowest correlation with the Total

scores at 0.70, n=389, p<.0001). the lowest correlation of any domain with the total scores. For validity, this artifact had been mapped to the state's licensing standards and to the Praxis III (Pathwise) assessment. These mappings were recorded on several documents that became part of the teacher education unit's electronic exhibits.

Differences in intern proficiency between supervision models:

We explored differences in full-scale scores between candidates who had been assigned to traditional triad models, those in cohort models, and those being supervised in a new distance learning format. The ANOVA in Table 2 shows what was found in proficiency between candidates in the three models of supervision.

Table 2

Full-scale cores between candidates of three supervisory conditions

Source of Variance	DF	SS	MS	F	P
Model (treatment)	2	4163.98594	2081.99297	9.86	<.0001
Error	412	89963.18091	211.07568		
Corrected Total	414	91127.16684			
R-square	0.045694				

Duncan's procedure was used followed the significant F to determine which supervisory conditions were associated with the highest ratings on teaching performance. The interns in the cohort supervision environment out-scored those in the traditional model by 53.7 total points to

48. 9, but that comparison was not significant at the .05 level. The distance learning mean at 42. 5 was significantly and, to us, *practically* below the means of the other two groups. Logistical and other difficulties with the distance supervision model as we experienced it led us to suspend the use of the distance supervision model in the foreseeable future.

Differences in Domains Between Supervisory Conditions

We explored differences in the four domains to see if interns in any of the three supervisory models seemed to perform better than those in others. In all four domains, interns in the cohort model and in the traditional model outscored those in the distance supervision model well beyond the .05 level. We rejected the third null hypothesis (Table 3).

Table 3

Subscale s cores between domains of candidates in three supervisory conditions

Domain	Mean, Traditional triad	Cohort	Distance supervision	p
A. Organizing Content For Student Learning	14.28 of 18	15.39	13.56	.0082
B. Creating an Environment for Student Learning	12.27 of 15	13.70	11.60	.0001
C: Teaching for Student Learning	12.20 of 15	13.63	11.64	.0002
D: Professionalism	12.18 of 15	13.16	7.60	.0001

The Professionalism scores (Domain D) for the distance supervision group were low enough to cause practical as well as statistical concerns.

Gender Differences

Teacher education units are also charged with the responsibility of logistically and technically insuring that their assessments are free from bias toward any demographic group (NCATE, 2008). Our principal means of detecting bias was to determine if various groups scored significantly higher than others. Gender was one variable we explored (Table 4).

Table 4

Full-scale cores between genders

Source of Variance	DF	SS	MS	F	P
Treatment	1	1106.30	1106.30381	4.95	0.0268
Error	414	92677.79	223.86		

Females significantly outscored males on the total instrument 52.44 points to 48.71. In a context where a difference of one point may have practical significance, it may have other meanings as well. The Pathwise instrument, as developed by the Educational Testing Service, is said to have no gender bias. It would appear that (a) our derivative of the instrument has some gender bias in it or (b) females do a better job of teaching than males do.

Comparisons between Academic Disciplines.

We compared the full-scale scores between the candidates of 12 academic programs. Those results are in Table 5. The entry for foreign language on the first line is most likely spurious, with four observations most likely referring to one intern. There were differences between interns from various programs, hence Hypothesis Five was rejected.

Table 5

Full-scale cores between academic disciplines

Program	Mean full-scale score	N*	Duncan Grouping**
Foreign language	58.75	4	A
Business Education	58.067	3	A
Early Childhood	54.24	212	AB
Music Education	53.5	10	AB
Mathematics Education	53.08	13	AB
Health and Physical Educ.	50.16	100	AB
Middle Level Eng/SS	47.25	14	ABC
History/Political Science	46.47	18	ABC
Art Education	44.39	16	ABC
Life Science	40.30	5	BC
English Education	34.75	8	C
Middle Level Math/Science	33.14	7	C

Notes:

*Observations, not numbers of interns. Generally each intern should have about four observations. The foreign language entry on the first line with 4 observations most likely refers to one intern.

**Majors with the same lettering are not significantly ($p<.05$) different, using Duncan's Multiple Range Test

Discussion

Determining the reliability and validity of assessments used in a teacher education program is a normal and usual part of academic life. Finding that this instrument possessed those necessary characteristics was necessary.

The finding that this particular version of distance supervision did not work very well was expected. A key limitation of watching an intern teach, either by Skype or through DVD or videotape, was that the observer could not see what was happening in the rest of the classroom. There were additional frustrations with interns either getting the tapes and DVDs to faculty in a timely manner or with getting feedback to the interns as quickly as they might have wished. There doubtless are ways that distance supervision could be done to accomplish intern supervision, but this one was not a very successful one.

Resolution of the third hypothesis only further drove home the point that this particular brand of distance learning/distance supervision was not very effective. We have discontinued it. This analysis shows that of the three supervisory models, cohort and traditional triad are

similarly effective, but distance supervision is a very distant third, even when the dependent variable is the domains (subscales) of the *Formative Observation and Intervention Form* and not the entire scale.

The fourth hypothesis was an attempt to learn if our instrument was partial to one gender or the other (non-bias). In our part of the country, there are not many other variables to investigate. Students of varied racial and ethnic backgrounds are beginning to come to this university, but too few have chosen teaching for an academic program for there to be sufficient numbers to permit quantitative analysis. The difference that was detected between genders raises the question: Does the measuring instrument discriminate against males, or at our institution, do females simply perform better on instructional skills? Resolution of the fourth hypothesis is sending us back to the literature.

Examination of the fifth hypothesis must be done carefully because the units being analyzed are observations, not interns. As has already been pointed out, the highest scoring evaluations most likely belonged to one, not four, high-achieving foreign language major. Where the numbers of observations are more substantial, more definitive conclusions might be reached.

References

Arkansas Department of Education. (2009). Schedule for novice teacher observations.

Retrieved November 13, 2009 at

http://www.arkansased.org/teachers/pdf/im_observations_0107.pdf

Hatcher, L., & Stepanski, E. J. (1994). A step-by-step approach to using the SAS System for

univariate and multivariate statistics. Cary, N. C: SAS Institute Inc.

Kerlinger, F. K., (1973). *Foundations of Behavioral Research, 2nd ed*. New York: Holt, Rinehart, and Winston, Inc.

National Council for Accreditation of Teacher Education. (2000). Planning instrument (Revised 2002 edition.) Washington, D. C.: NCATE.

National Council for Accreditation of Teacher Education. (2008). Abbreviated planning instrument for 2008 NCATE standards. Washington, D. C.: NCATE.

U. S. Census Bureau. *Pope County quickfacts from the U. S. Census Bureau*. Retrieved on November 5, 2009, from http://quickfacts.census.gov/qfd/states/05/05115.html .