Running head: Elusive Nature of Reliability

The Elusive Nature of Reliability: Problems and Pitfalls in Scoring Clinical Practice Action

Research Projects

David W. Moffett

Barbara K. Reid

Brewton-Parker College

Abstract

The Investigators studied scoring reliability of Candidates' ten day unit plans of instruction through prescribed action research projects, across three academic years. Scoring of the projects in year one provided opportunities for further refinement of the action research evaluation methods in year two. Across three terms in years one and two scoring reliability was realized. However, although reliability was realized in term two of year two, new scoring issues arose. Subsequently in year three of the study reliability was not realized across rater and interraters due to a lack of updating points earned on projects in mastery learning, lack of abiding by the rubric for scoring, and informal discussions between the rater and some interrators affecting scoring of projects. Discussion includes changes that likely led to reaching early satisfactory reliability levels, as well as reasons for not realizing reliability in the most recent year. Recommendations follow that may be helpful to education Units seeking to learn more about reliability issues when scoring Candidate items.

The Elusive Nature of Reliability: Problems and Pitfalls in Scoring Clinical Practice Action Research Projects

The education unit has eight assessments driving its assessment system. Assessment five is known as "Student Teacher Effect on Student Learning." In 2007 the Education Division faculty invested the summer in reviewing and revising unit assessments. The former assessment system was not resulting in data that could be analyzed to inform the Unit and lead to change.

Over the course of six months a new assessment five emerged. After considerable deliberation and research on best practices in measuring student teacher effect on student learning, the first generation of the current model was applied to the unit's clinical practices in the spring term of 2008. Reliability was not realized across the rater and interraters at the end of the first term of incorporating the new assessment five.

Additional changes were made prior to the second term of use (fall term, 2008-2009) and reliability was realized across rater and interraters. The third term of use (Spring Term 2008-2009) saw new faculty as interraters and this most likely affected reliability to a degree. Along with new faculty effects on reliability, the definitions for levels of scoring came into question as possibly having a negative impact on reliability.

Literature Review

How can we best measure candidate's effects on student learning and what effects do student teachers have on student learning? First, there must be an adequate data collection system. Also, what is the assurance that student teachers will alter practices based on the data they glean? Will data derived alter instruction provided by the candidates? How will we know? Finally, how can we cause this data to occur naturally so it is not merely an add-on to the already overwhelmed candidate and overworked cooperating and education faculty? The sheer number

of candidates going through teacher education programs can add to the maelstrom when trying to determine effect on learners. Determining the effect of student teachers on student learning can be very problematic (Williams & Balach, 2007).

Our best answer regarding how one could determine student teacher effect on student learning was to implement prescribed Action Research in the classroom by the student teacher. Action Research provides candidates with tools of systematic inquiry and beginning investigation skills, while providing the needed measuring stick to determine effect on student learning (Emery, Jumper, & Bruce, 2007). Such research has not been apparent at the undergraduate level in most teacher education programs. It has been traditionally housed in graduate programs, often as the capstone, but it does have a place in candidate clinical practice. Candidates formulate hypotheses in regards to their effect on student learning as student teachers and then test them by implementing unit guided Action Research in their clinical practice classroom(s).

In establishing an Action Research project for all student teachers we must be able to articulate to faculty and students what it is and what it is not (Ross-Fisher, 2008). Without proper clarity the undertaking of such an endeavor can easily go off the proverbial tracks. Action Research is not experimental and it is often messy and uncertain (Goodnough, 2008). It is the responsibility of the education Unit to ensure that the research projects don't become too messy or uncertain.

Again, what candidates incidentally learn from performing Action Research can cause them to feel as though they have a greater understanding of the big picture of what it means to be a professional educator. It offers candidates a professional identity they often don't possess

without engaging in such research, and it cultivates professional relationships and development (Warren, Doorn, & Green, 2008).

Candidates engaged in student teaching have a view of day to day classroom operations like few researchers can possess. Life in the teaching trenches offers candidates the opportunity to develop unique strategies for meeting individual student needs. However, teachers have been historically reluctant to engage in research (Nonis, 2008). As education faculty it is our responsibility and duty to instill in our candidates the expertise and initiative to be action researchers. Candidates should come to feel a sense of duty to research their classroom often and guide them with the data received.

Student teachers should engage in Action Research since they can experience success in it, which will lead to subsequent research attempts beyond their clinical experience. Student teachers who struggle with the daily realities of the clinical experience are the ones who probably need to internalize the results of Action Research most. However, they will be the ones who will most likely have the greatest difficulty implementing the research and deciphering the results that can ultimately improve their teaching (Monroe, Gali, Swope, & Perreira, 2007). This is only one of several conundrums we face in trying to adequately prepare candidates to be proficient educators.

Once student teachers collect and analyze the data garnered from Action Research where do they go from there? After all, how long do teacher effects persist anyway? In one study teacher effects on student learning are defined as, "Teacher specific residuals adjusted for student and treatment effects" (Konstantopoulos, 2007). Considering all of the variables that affect student learning, what is the effect size of teacher effect on student learning? Per the study, teacher effect is cumulative and the effects are evident beyond the current candidate/student

experience. It would appear as though longitudinal studies guided by state departments of education would be appropriate to best capture teacher effect on student learning, if this is the case. In other words candidates may well not see the total effect on student learning through their Action Research snapshot view of student learning. Still, our candidates should conduct Action Research and not speculate beyond the collected data, other than to recommend that their effect on students be studied across subsequent student years in school to capture total effect size.

Teachers as researchers can at least partially address the need for interventions intended to improve student performance in the classroom. Intervention can result from analyzing data collected by classroom teachers in their research. This value added self-assessment has the potential for resulting in teachers selecting professional development in areas they determine to be in need of improvement. Such intervention has promise for translating into continued, positive effects on student learning for several years, especially in students' early grades (McCaffrey, Lockwood, Mariano, & Sedodji, 2005).

Historically, there have been calls for studies on teacher effect on student learning. Effect on student learning includes the amount of student time on-task and this is correlated with effect on student learning and candidate characteristics (Fox, 1978). Candidate characteristics do indeed affect student learning. Empowering candidates to analyze their effect on student learning through Action Research can encourage reflective practice and incite personal, positive changes in pedagogy and practice.

<center>Term One of the Study</center>

The study began in the spring term of the 2007-2008 academic year. Twenty-four student teachers were assigned the task of completing prescribed Action Research projects in their

assigned classroom(s). The primary study investigator met with the student teachers at the beginning of the term to review the multiple-step research process that was to be implemented.

The steps included drafting a 10-day unit plan of instruction, developing and administering a pretest reflective of the content of the unit plan early in the term, analyzing the results of the pretest and revising the draft unit plan as needed, teaching of the 10-day unit plan, and administering a unit grand assessment with the pretest embedded therein.

One of the secondary investigators, who served as director of student teaching, oversaw candidate Action Research progress across the term. Some non-Action Research assignments were due across the semester as well and were listed in the calendar. However, everything pertaining to the Action Research project was due at the end of the term. Instructions for the Action Research project were embedded within the unit's 10 outcomes and were somewhat obscure. Basically the new "Unit Assessment Five" was plugged into existing assignments and this did not provide enough needed clarity for the candidates. Also, the clinical practice director was new to the concept of Action Research so they could not provide needed clarity for the candidates.

At the end of the term the director of student teaching printed the twenty-four Action Research projects from the electronic learning management system used in student teaching. The director and other education faculty members, who served as the interraters, had devised a 10-part evaluation instrument. Each of the 10-parts of the evaluation tool had a possible value of three points. The total possible points that could be awarded to an Action Research project were 30.

The director of student teaching scored all twenty-four candidate projects. Then, interraters were randomly assigned to score the projects, blind to the director of student teaching

scores. In cases where the scores across the director and interraters possessed a difference of 10 percent or more, a second interrater blind to both the director's score and the first interrater's score, evaluated and scored the project. Second interraters were randomly assigned to score projects just as initial interraters were randomly assigned.

Twenty-four student teachers participated in the assessment five Action Research project in its inaugural term, spring 2007-2008. Twenty-four action research projects were rated by the director of student teaching. Then, the projects were randomly assigned to other education division faculty interraters. Ten projects were interrated for a second time randomly by five of the faculty Interraters who had not previously evaluated the particular projects. These ten projects were interrated for a second time because the difference between the director of student teacher's score and first Interrater score was greater than ten percent of possible project points. All five second round Interraters participated in the first round of interrating, but the projects scored in round two were not scored by them in the first round. The total and mean of all the participants of the study by the rater were 317 and 13.31 respectively. The total and mean of all participants after the first round interrating were 315 and 13.13. The total and mean after the second round interrating for all participants were 276 and 11.50. As indicated by the data from the second round of interrating, eight of 10 action research projects continued to possess a score difference greater than ten percent. This difference can be explained in several ways but it was most likely the result of more training being needed, for both the rater and the interraters, in scoring the projects. Between the second scores and the first, four of the ten score differences were larger than in the first round. The total difference between the director of student teaching and the first interraters was -2, while the total difference between the director and the second interraters was -41. The total difference between round one Interrater scores and the second

round of Interrater scores was -39. An alpha cronbach reliability test revealed the alpha

coefficient for the student teaching director and the first Interrater was .40 and .53 for the

director and the second Interrater. The same test was applied to the first Interrater and second

Interrater and the alpha coefficient was found to be .74. There was reliability across Interraters

but not between the Rater and the Interraters.

Term One Results

Table 1

Descriptive Statistics of the Scores by the Rater, First Interrater, and the Second Interrater

Scores	Rater/Interrater n	n	M	SD
by Rater	1	24	13.21	4.37
by First Interrater	6	24	13.13	4.05
by Second Interrater	5	8	7.40	3.14

Note. N = 24.

Table 2

The Analysis of Variance (ANOVA) by Rater, First Interrater, and Second Interrater

Varible	df	SS	MS	F	p	α
Rater * First Interrater	23	376.23	24.29	1.52	.23	.40
Rater* Second Interrater	7	67.88	0.11	0.01	.93	-.06
First Interrater*Second Interrater	23	492.00	176.05	12.23	.00	.74

Note. N = 24.

In the first term of use of the new Assessment Five Action Research project results were

confounding on at least two levels. Overall quality of the Action Research projects was poor

with the mean below fifty percent of points possible. Perhaps this could be attributed to it being

the initial attempt of implementing such a project in student teaching. Or, it could be that the

assignments and expectations were not as clear as they needed to be. Or, the fact that the entire

project did not have to be submitted until the end of the term may have had something to do with

the low performance since candidates did not have the opportunity to refine particular pieces of

the project along the way and there was no consequence for submitting low quality product.

Whatever the case, or reason, the projects for the most part did not meet Unit expectations (see Tables 1 and 2).

A second area of concern was the lack of reliability across scores given by the director of student teaching and the interraters. The alpha cronbach reliability needed was not evident across director scores and those of the interraters. It could be speculated that the director was privy to subjective information regarding particular candidates and factored that knowledge into the scoring of the projects, while the interraters had no knowledge of information beyond the projects themselves. In other words, the director's scores may have included data that the interraters could not see or know when they scored the projects. To further defend this hypothesis, the reliability scores across interraters was at an acceptable rate to ensure reliability, while the difference between rater and interrater scores was not acceptable for reliability.

Conversely, the mean scores across director and first interraters were amazingly similar. On a thirty point scale the mean difference was less than .10. Still, the standard deviations were four or higher. This dissonance between the mean scores and standard deviations caused the investigators to greatly appreciate having more than one way of seeing the data. Had the investigators only evaluated differences between mean scores they would not have discovered the considerable differences within test scores between the rater and interrater scores.

In summary, the Action Research projects generated by the candidates in the first term of the study were not satisfactory. Secondly, there was no reliability across rater and interrater scores, thereby negating the assurance needed that the projects were scored in satisfactory and meaningful ways. Therefore, the data derived were not sufficient to inform the unit other than to cause it to revise the student teacher Action Research projects in substantial ways.

Much needed to be done to improve the clarity of expectations for the projects and the reliable scoring of them.

Term Two of the Study

In Fall Term 2008-2009 much was done to attempt to secure better projects and more reliable scoring. A report outline became evident in the electronic learning management system that houses the instructions for the Action Research projects. Also, a grading rubric was developed and shared with the candidates. It was very specific as to what must be included and it articulated the desired organizational method of the project. The primary investigator checked for candidate understanding in more meaningful ways when presenting the Action Research project to them at the beginning of the term in the initial student teaching seminar. The director who oversaw the projects possessed a clearer understanding of the concept of Action Research and was able to provide greater clarity regarding project expectations to the candidates in the second term attempt.

Additionally, each step of the Action Research project was turned in along the way and those were evaluated by the student teaching director. If the director found any portion of the project submitted to be unsatisfactory the candidate had to revise and resubmit that particular portion before proceeding.

Additional scoring training for the rater and the interraters was also implemented, to attempt to reach acceptable levels of reliability. All scorers came to know that they only score what is visible to all of those who are doing the scoring. This reduced some of the discrepancies evident in the first term of the project scoring.

The investigators were hopeful that the additional clarity provided to the candidates, rater, and interraters would result in acceptable reliability levels. Prior to the second term of use the student teaching director reflected on term one results. They concluded that it was obvious that there needed to be changes. In their reflections they commented that there was a wide fluctuation between rater and interrater scores and that the method of research design was too broad among candidates. Their test administration procedures, data collection, data analysis, and data interpretation were too reflective of a shotgun approach to completing the Action Research projects.

It became apparent the steps involved in the prescribed Action Research project were a better format for the candidates' electronic portfolio design. So, instead of plugging in the elements of the research into what already existed the director elected to have the steps in the research drive the clinical practice experience. Specific due dates for parts of the research became evident. As discussed previously in the first term everything was due at the end of the term so there was no time available for carefully evaluating the projects. With the new due dates along the way added, the director had time to carefully evaluate each prescribed step in the research across the term.

A more defined set of directions and procedures became apparent for candidates in the second term of the study. Greater clarity became evident in the areas of directions, procedures, and a rubric for grading. The rubric, timelines, and due dates across the term were shared with candidates at the beginning of the term. A single organizational template for the final report was also introduced.

<div align="center">Term Two Results</div>

Table 3

Descriptive Statistics of the Scores by the Rater and Interrater

Scorer	Rater/Interrater n	*M*	*SD*	Total
Rater	1	26.00	4.12	286
Interrater	5	22.86	4.82	251.1

Note. N = 11.

Table 4

The Analysis of Variance (ANOVA) Statistics by the Rater and Interrater

Variable	*Df*	*SS*	*MS*	*F*	*p*	*r*	*α*
Rater*Interrater	10	170	76.23	7.32*	0.02	0.67	0.80

Note. N = 11.
* *F* significant at .05 level.

Table 5

The t-Test Results of the Scores by Rater and Interrater

Variable	*df*	*M*	*SD*	*t*	*p*
Rater	10	26.00	4.12		
Interrater	10	22.86	4.82		
Rater*Interrater				20.91*	.00

Note. N = 11.
* *t* is significant at .05 level.

Term two results saw reliability realized across the rater and the interraters. The standard deviation continued to be approximately 4.00, as was the first term deviation. Interestingly, there was a greater range of mean scores across rater and interraters when compared with term one results (see Tables 3, 4, and 5).

<center>Term Three of the Study</center>

Although reliability was realized, more reflection was needed after the second term in the spirit of continuous improvement. Areas that were addressed prior to the third term of the study included more rater and interrater training on scoring the projects and more detail regarding the notebook arrangement containing the elements of the prescribed action research project. During this time there were faculty changes within the education division. Two new faculty members

joined the division replacing two professors who had been interraters in at least one term of the

study. Training was provided by the director of student teaching and spring term 2008-2009

projects were distributed to the interraters in the August, 2009 division meeting. The same

procedures were used to randomly assign candidate projects to the interraters. Interestingly, there

was still reliability realized across rater and interraters but the results revealed a drop of .07 from

the second term to the third.

<div align="center">Term Three Results</div>

Table 6

Descriptive Statistics of the Scores by the Rater, First Interrater, and the Second Interrater

Scores	Rater/Interrater n	n	M	SD	Total
by Rater	1	18	22.89	6.35	412
by First Interrater	6	18	22.11	5.89	398
by Second Interrater	6	9	20.33	6.36	183

Note. N = 18.

Table 7

The Analysis of Variance (ANOVA) by Rater, First Interrater, and Second Interrater

Variable	Df	SS	MS	F	p	α	r
Rater* First Interrater	17	589.78	195.57	7.94*	.01	.73	.58
Rater* Second Interrater	8	324.00	177.59	8.50*	.02	.85	.74
First Interrater* Second Interrater	17	589.78	323.84	19.48*	.00	.85	.74

Note. N = 18.

* *F* is significant at .05 level.

It appeared as though the drop in reliability from term two to three was due to change in

faculty and also because of misunderstandings regarding the definitions of the grading scale used

to assess the projects. At least one faculty member beyond the new faculty used the proficient

level as the target in assessing the projects, since the overall goal of the Unit assessment system

is to graduate proficient beginning teachers. It is interesting to note the difference between rater

and interrater scores and rater and second interrater scores. The rater/interrater reliability was

.73, while the rater/second interrater was .85. Interraters and second interraters were from the same pool of faculty so this infers that there are still scoring issues among some interraters. The Unit's goal is to cause candidates to reach proficient levels across assessments used in the system. However, the exemplary level as described in the scoring evaluation of the Action Research projects is actually reflective of proficient attributes. This may well be why the rater's scores were consistently higher than the interraters' scores (see Tables 6 and 7).

One could argue that as long as reliability is realized there is little need to continue to reflect on how to improve it. In fact such discussions have taken place among division faculty regarding this topic. On the other hand the drop in reliability from term two to three is such that the Unit is only .04 away from not realizing reliability across rater and interraters. In order for there to be merit in the results of the candidates' projects there must be reliability so the Unit must continue to continuously improve the Action Research evaluation process to ensure reliability.

Concerns about the drop in reliability in year two came to fruition in the third year of the study. In fall term 2009-2010 the rater applied mastery learning to candidates' projects. Students were stopped if particular portions of their projects were not satisfactory along the way and told to revise them as needed. The revisions took place but the rater's scores for same did not increase in all areas when the portions had been mastered. This led to an exaggerated discrepancy between rater and interrater scores on several of the projects which caused there to not be an acceptable level of reliability. The Chair of the division determined it would be futile to attempt to ensure reliability and asked the rater to simply go back and update the scores at either the acceptable/1 or proficient/2 level, based on where the initial score resided; with those scores

being close to the proficient level becoming proficient/2s and those near the acceptable level becoming acceptable/1's. Fall term 2009-2010 results:

Old Rater Score	New Rater Score	Interrater Score
4	10	26
17	20	15
19	20	15
19	20	30
18	20	21
13	20	16
20	20	15
17	20	23
19	20	16

Spring term of year three of the study saw the rater and interrater scores of Candidate Action Research not reach acceptable levels of reliability once again. The problems related to the rater not updating points earned on projects in the previous term were addressed, and solved, but new reliability roadblocks emerged.

The rubric used to score the projects consisted of a four point scale with 0 being Unacceptable, 1 being Acceptable, 2 Proficient, and 3 Exemplary. The target level for graduating Candidate's in the Unit is Proficient across all assessments. In some cases Action Research projects did not include all elements of one or more of the 10 categories used to assess and score them. In division meetings the Chair and an additional forty percent of the division faculty

thought everyone had agreed that if any element was missing that portion of the assessment would be scored as a zero. However, the rater determined on their own that if a Candidate submitted at least part of the required assignment they should earn more than a zero. Interestingly, this idea of awarding a 1/Acceptable for incomplete work is counter to the concept and philosophy of mastery learning which was applied to the Action Research projects in the previous term. If mastery learning is being used in the construction of the Action Research plans then this should mean there are no missing elements of the project.

Beyond the rater synthesizing the scoring rubric to better fit the situation they had casual conversations about their scoring technique with two of the interraters. These discussions regarding how the rater was scoring the assessments led to the two interraters having perfect correlation with the rater. However, those faculty who followed the Division's rubric scoring guide, had eighty to one hundred percent lack of correlation with the rater scores. This discrepancy led to the Unit not realizing satisfactory levels of reliability in the spring term of the third year of the study. Spring term 2009-2010 scoring results:

Number	Rater Score	Interrater #1	Score Difference	Interrater #2	Score Difference
1	16	17	1	17	
26	18	13	-5	17	1
13	20	16	-4	17	3
22	18	21	3	21	
7	23	10	-13	22	1
12	18	4	-14	19	-1
25	12	18	6	10	2

24	24	21	-3	21	
16	13	15	2	15	
19	18	19	1	19	
17	19	14	-5	10	9
11	18	16	-2	16	
15	17	15	-2	15	
21	13	14	1	14	
9	14	13	-1	13	
20	24	18	-6	19	5
8	15	17	2	17	
23	20	18	-2	18	
14	11	17	6	5	6
18	19	17	-2	17	
3	18	13	-5	12	6
5	12	11	-1	11	
6	15	9	-6	13	2
2	16	15	-1	15	
4	13	13	0	13	
				10 total	4 total
	424	374	-50		

Summary

For three years the Investigators studied issues with reliability across raters and interraters, in the scoring of clinical practice Unit Assessment 5 Action Research projects. In the first term of the study the new project was incorporated into what already existed in the clinical practice program. The director of the program had little knowledge of Action Research. Candidates submitted their entire project at the end of the term without formative and diagnostic feedback along the way. There was a range in quality of student work and reliability was not realized across the rater and interraters.

Prior to the second term of Action Research being used in clinical practice the format for the prescribed project became the actual format for the experience and the director of the experience became better versed in the concept of it. A rubric was developed and the rater and faculty interraters were further trained in scoring the projects. Reliability was realized in term two of the study.

Prior to the third term of the study and use of Action Research in clinical practice, the director further refined the delivery of the information to candidates and new faculty were hired in the Unit. Although reliability was realized in the third term, the results were lower than in the second term and this was attributed in large part to the change in faculty and the need for further training. The combination of new faculty evaluating the projects and misunderstandings regarding levels of scoring were likely the culprits resulting in the lower reliability scores, even though reliability was realized.

In the fall term of year three of the study the rater did not update points earned for portions of the Action Research projects, as students exhibited levels of mastery, which resulted in a rater undervaluing of points. The rater undervaluing of the projects led to their not being

reliability across rater and interrater scores. In spring term of the third year the rater formulated

a revised scoring system for the projects. The synthesized scoring system was shared with forty

percent of the faculty and perfect scoring correlation between the rater and those faculty who

were engaged in the informal discussion were perfect. However, those faculty who were using

the supposedly agreed upon rubric for the assessment had eighty to one hundred percent non-

correlation with the rater scores. These discrepancies led to there not being reliability in scoring

across rater and interrator scores in the second term of year three.

The Investigators agreed, earlier in the study, with the notion that preservice teachers can

engage in Action Research and be successful (Monroe, Gali, Swope, & Perreira, 2007). They

also believed exemplary projects should become available for candidate review, along with

continued work in the area of providing greater clarity regarding expectations (Murray, Grande

DiCamillo, Henry, & Henry, 2008). Such availability began in the third year of the study.

As reported in year two of the study, further research was planned to shift to the actual

purpose of the prescribed Action Research projects. The purpose of the projects is to determine

student and class wide gains or losses as a result of the candidates' 10-day unit plan of

instruction while student teaching or in an internship. However, the issues regarding lack of

reliability in scoring the projects must be addressed. Time must be devoted to revisiting the

scoring guide and rubric to cause all faculty members to apply the agreed upon instruments to the

projects. Or, the scoring instruments need to be modified to the point where faculty can agree to

use them, while the instruments themselves maintain integrity. Additionally, the concept of

mastery learning needs to be comprehended by all parties to ensure everyone agrees that the

elements needed in each portion of the scoring guide must be present for that portion to at least

be acceptable. Results of these efforts and others shall be shared in future papers.

References

Emery, M., Jumper, J., & Bruce, T. (2007). *Transformational teachers: Undergraduate education teacher candidates assessing their impact on student learning through action research*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, New York, NY. Retrieved October 21, 2008, from http://www.allacademic.com/meta/p142785_index.html

Fox, R. (1978). *Tracing teacher effects through student behavior to learning outcomes.* Paper presented at the Annual Meeting of the American Psychological Association, Toronto, Canada. Retrieved October 21, 2008, from http://eric.ed.gov/ERICWebPortal/custom/portlets/recordDetails/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED169039&ERICExtSearch_SearchType_0=no&accno=ED169039

Goodnough, K. (2008). Dealing with messiness and uncertainty in practitioner research: The nature of participatory action research. *Canadian Journal of Education*, *31*(2), V-VI.

Konstantopoulos, S. (2007). How long do teacher effects persist (No. 2893). Bonn, Germany: IZA.

McCaffrey, D. F., Lockwood, J. R., Mariano, L. T., & Sedodji, C. (2005). Challenges for value-added assessment of teacher effects. In R. Lissitz (Ed.)*, Value-added models in education: Theory and applications* (pp. 111-144). Maple Grove, MN: JAM Press.

Monroe, E. E., Gali, K., Swope, K., Perreira, I. (2007). Preservice teachers' use of action research to implement alternatives to round robin reading. *Journal of Reading Education, 32*(2), 13-17. Retrieved September 20, 2009, from Wilson OmniFile Database.

Murray, R., Grande, M., Lorrei DiCamillo, L., Henry,J., & Henry, D. (2008). The annotated unit: A description of a systematic approach to documenting candidate effectiveness in student teaching. *Action in Teacher Education*, *30* (3), 74-87. Retrieved July 20, 2009, from Wilson OmniFile Database.

Nonis, K. P. (2008). Breaking barriers: Building research partnerships between special education teachers and universities in action research in Singapore. *The Journal of the International Association of Special Education*, *9* (1), 28-37.

Ross-Fisher, R. (2008). Action research to improve teaching and learning. *Kappa Delta Pi, 44* (4), 160-164.

Warren, S., Doorn, D., & Green, J. (2008). Changes in vision: Teachers engaging in action research. *The Educational Forum*, *72*(3), 260-270.

Williams, A., & Balach, C. (2007). *Building capacity through teacher candidate action research: Documenting teacher candidate ability to positively impact student learning.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, New York, NY. Retrieve Oct 21, 2008 from http://www.allacademic.com/meta/p142771_index.html