Running head: COOK SCHOOL DISTRICT SIMULATION

Critique of “Exploring the Efficacy of the Cook School District Simulation”

Rod Myers

Indiana University

R690 – Application of Research Methods to IST Issues

Professor Frick

November 23, 2007

Description of "Exploring the Efficacy of the Cook School District Simulation"

In the journal article "Exploring the Efficacy of the Cook School District Simulation," Girod and Girod (2006) describe their study of the use of a Web-based classroom simulation in a teacher education program. The purpose of the study was to test the effectiveness of the simulation in enabling teacher candidates to practice connecting teaching and learning and to prepare for evidence-based assessment of their teaching. The authors contend that the study is significant because there is a trend toward using student learning as a measure of teacher/school effectiveness, resulting in increased pressure to demonstrate learning gains for all students (e.g. Public Law 107-110, commonly known as the No Child Left Behind Act of 2001 or NCLB). Therefore teacher education programs must ensure that their graduates are able to account for the efficacy of their instructional efforts.

While NCLB attempts to measure student learning through large-scale standardized testing, many teacher education programs are focusing on alternative methods for demonstrating the impact of teachers on learning achievement. One approach, which is used at the location of this study, is teacher work sample methodology (TWSM). This approach provides teachers with critical skills that enable them to gather evidence of the systematic connections between their teaching actions and student learning. However, the authors cite several researchers who have noted the lack of adequate apprenticeship opportunities in which future teachers can practice the skills and activities of TWSM. This led the authors to create the Cook School District simulation that is the focus of this study.

The Girod and Girod 2006 study was designed to answer four research questions related to the use of the simulation for teacher preparation:

When compared against teacher candidates who did not use the Cook simulation, do teacher candidates who used the Cook simulation

1. have different perceptions of their own skillfulness in connecting teaching and learning?
2. have different levels of value for the skills necessary to connect teaching and learning?
3. demonstrate a different ability to connect teaching and learning as evidenced within their teacher work sample?
4. teach differently in a real setting? (p. 487)

This study is relevant to the field of instructional systems technology (IST) because there is a long history of using simulations and games for instruction. In the past, these have been primarily analog artifacts such as card games, board games, and role-playing activities. However, advances in computing power and software have made it possible to simulate complex, interactive environments with many variables and personalized feedback. Researchers in IST are increasingly investigating the design and use of digital simulations and games for learning. In particular, there is a need for theory-based methodologies to guide the design of, to facilitate the use of, and to assess the effectiveness of instructional games and simulations.

The authors employed a quasi-experimental research design in which students self-selected for the treatment group. The sample was taken from Master of Arts in Teaching teacher candidates who were pursuing initial teacher licensure at a regional state university in the Western United States. The students had been admitted into three cohorts of 24-28 students. Up to half of the students from each cohort self-selected for treatment (treatment: n = 33, control: n

= 38). The treatment group and the control group had comparable numbers of students by gender, by age, and by content emphasis.

Both groups participated in the regular teacher preparation program, which consisted of "instruction in work sampling, assessment, classroom management, and instructional design" (p. 488) and student teaching experiences. Additionally, the treatment group used the simulation to practice designing and evaluating their teacher work samples and received feedback on their performance. This practice consisted of three 2-hour work sessions over three weeks between the preparation of two independent work samples.

During the first session, the teacher candidates learned how to use the simulation and became familiar with the virtual setting and the characteristics of their randomly-assigned simulated students. The second session consisted of working in the simulation to formulate objectives, to implement instructional strategies, and to create assessment items for their twenty simulated students. At the end of the session, the teacher candidates completed a worksheet that facilitated reflection on their actions and the outcomes in the simulation. The third session began with everyone focusing on five simulated students followed by a whole group discussion about the effectiveness of various strategies. The teacher candidates then returned to working with their twenty simulated students and practiced defending their instructional decisions using data they collected.

Both groups responded to pretests and posttests immediately before and after the three weeks in which the treatment group participated in the sessions. The "Skillfulness and Value Inventory" consisted of 22 items on a standard Likert scale. Eleven items measured the participant's self-perception of skillfulness in connecting teaching and learning, and eleven items measured the participant's perceived value of those skills. The teacher work samples were evaluated by seven part-time adjuncts whose only responsibility was supervision of student teachers in the field. Evaluations were based on a rubric that had been used for several years. A third source of data was an evaluation of lessons taught in the field placement settings by the teacher mentors. The evaluation form was “aligned with the teacher competencies articulated by our state teacher licensing agency” (p. 489) and had also been used for many years.

For their first two research questions (related to self-perception of skillfulness in connecting teaching and learning and perceived value of those skills), the authors found that both the treatment group and the control group were nearly identical on the pretest. Both groups made gains on the posttest, but the treatment group’s scores were significantly higher. Gender and GPA were not found to be predictive of the results. For the third research question (pertaining to the quality of the teacher work samples), both groups made gains on the posttest, with modest gains for all components of the evaluation but significant overall gains. For the fourth research question (teaching differently in a real setting), the treatment group scored higher on all pretest factors. Nevertheless, the authors noted that working with the simulation had a positive effect on the classroom performance of the treatment participants.

Critique of the Study

Criteria Used in Critique

The following criteria were used to critique the study by Girod and Girod (2006) of the effectiveness of a classroom simulation in enabling teacher candidates to practice connecting teaching and learning and to prepare for evidence-based assessment of their teaching.

Research methodology. According to Fraenkel and Wallen (2006), “[i]n intervention studies, a particular method or treatment is expected to influence one or more outcomes” (p. 15). The most appropriate methodology for this type of study is the experiment, which involves manipulating one or more independent variables and examining the effect on at least one dependent variable. Generally two (or more) groups are involved so that the effect of the treatment can be compared to the effect of no treatment. An important characteristic of experimental studies is the random assignment of subjects to treatment and control groups. The purpose of random assignment is to minimize the effects of confounding variables by dispersing them randomly throughout the sample, thereby strengthening any claims about causal relationships that the researcher may make.

Research questions. Fraenkel and Wallen (2006) state that research questions should be feasible, clear, significant, and ethical. A feasible research question is one that may be answered given a reasonable amount of time, money, and other resources. Clarity is an important characteristic of a research question because ambiguity signals lack of focus and can lead to confusion and miscommunication among researchers, subjects, and readers of the study. A research question is significant if it advances knowledge in a field, improves practice in some way, or improves the human condition. A worthwhile research question is neither capricious nor based solely on self-interest. Finally, a research question should not lead to physical or mental harm to people (and some would argue to animals as well) or the environment. Furthermore, the subjects must not be deceived and their rights and privacy must be respected. Creswell (2005) adds that good research makes clear the relationship between the research questions and the results.

Population and sampling. Because studying an entire population is generally not feasible, a researcher must first identify the population of interest and then select a sample of that population for study. It is important that the sample be representative if the results of the study are to be generalized to the population. Generalizability is a worthy goal because it means that the results of the research may inform practice in settings and with populations that are not exactly identical to those used in the research. Fraenkel and Wallen (2006) note that in practice researchers tend to sample not from the target population but from an accessible population. Regardless, it is ideal if the sampling is random for the reasons described above. Sometimes this is not possible and the researcher must resort to an alternative method such as purposive sampling or convenience sampling; however, these methods limit the degree to which the results may be generalized, and the researcher must hope that others will replicate the study to add support to the findings and increase their generalizability. While there are no strict rules for sample size, it is generally considered best to have around 30 subjects per group in experimental studies.

Instrumentation. The instruments used to gather data from subjects should be valid and reliable; otherwise the data may be erroneous and lead to incorrect conclusions. Validity is the ability to draw accurate and meaningful inferences from the data gathered with an instrument. The instrument must therefore be appropriate for the research questions it is intended to address. Evidence of an instrument’s validity may be content-related, criterion-related, and/or construct-related (Creswell, 2005; Fraenkel & Wallen, 2006). Reliability is the consistency of the data obtained with the instrument. Consistent measures should be obtained by related items or sets of items (internal consistency) as well as by separate administrations of the instrument to the same individuals. There are several methods for estimating the reliability of an instrument, including test-retest, equivalent forms, inter-rater reliability, and tests of internal consistency such as the Kuder-Richardson approach and the coefficient (Cronbach) alpha (Creswell, 2005; Fraenkel & Wallen, 2006). It is possible for an instrument to be reliable but not valid, i.e. it may deliver stable and consistent data which is not meaningful or relevant to a given study.

Data analysis and interpretation. The purpose of analyzing data is to answer the research questions posed at the beginning of the study. Analysis and interpretation allow the researcher to make meaning from the data and express it so that others may use the results to conduct further research or to improve their practices. There are two types of data, quantitative (or measurement) and categorical (or frequency), each with applicable methods of analysis. Various methods of statistical analysis are based on assumptions regarding the population, the sample, and the data. Using an inappropriate method of analysis or violating the assumptions of the method may lead to erroneous results. Therefore it is critical to select the appropriate methods of analysis based on the research questions, the research design, and the collected data.

Propositional knowledge claims. Frick (2005) provides a matrix for classifying types of propositional knowledge claims and evaluating them based on specific criteria. The article by Girod and Girod (2006) makes praxiological claims that are both situated and theoretical, one Type 3 claim regarding the effectiveness of a particular intervention and two Type 4 claims regarding general means for teacher preparation and student learning. A Type 3 claim may be justified by empirical evidence of the intervention's effectiveness, efficiency, and/or appeal. A Type 4 claim requires evidence of the instrumental value of the stated rules or principles.

Application of Criteria

In this section, the criteria specified above are applied to the study conducted by Girod and Girod (2006). In general, the article is well written and conforms to the expected format for the presentation of research. The topic—teacher preparation in connecting teaching and learning and providing evidence of effectiveness—is significant and worthy of study. There is increasing pressure for teachers and schools to provide evidence of learning for all students, so research into instructional methods that improve a teacher's ability to account for her pedagogical choices and actions is worthwhile. The theoretical framework for the intervention—cognitive apprenticeship and legitimate peripheral participation—is sufficiently addressed, although the authors might have provided stronger support for certain propositional knowledge claims, as described below.

Research methodology. While the research design for this study was quasi-experimental, due to convenience sampling and self-selection for treatment rather than randomized selection, in all other ways the study followed the experimental pretest-posttest control group design. The pretest was used to determine whether there were any significant differences between the two groups before the treatment was administered, as well as to provide a baseline for measuring the effect of the treatment compared to the traditional instruction. The duration of the treatment was limited to three weeks in an attempt to lessen the impact of the subjects' natural maturation. Other than the treatment, the groups received the same traditional instruction. The study utilized several methods to evaluate the effects of the treatment, including questionnaires, work samples, and judgments of performance in real-world settings, all designed to answer the research questions using a variety of measures.

Self-selection for the treatment group may have resulted in confounding variables such as a preponderance of high achievers or a proclivity for the use of technology in that group. The authors acknowledged that the quasi-experimental design and lack of random assignment limited the conclusions they were able to draw from the findings.

A significant flaw in the application of the treatment is that it resulted in additional learning time for the treatment group. The authors state that the time was relatively brief given the significance of the gains (six hours over three weeks), but they concede that further exploration is warranted to determine whether the treatment would be as effective when used as a replacement for traditional instruction rather than an addition to it.

Research questions. The research questions that defined the study meet the criteria described by Fraenkel and Wallen (2006), i.e. they are all feasible, clear, significant, and ethical. In particular, the research questions ensure that the results of the study can be used to advance knowledge and improve practice in the field of teacher education. The research questions cover Kirkpatrick’s second and third levels of evaluation. The first two questions concern the perceptions and attitudes of the subjects toward the content. The third question requires a demonstration that the subjects have learned and know how to apply the content. The fourth question asks whether the subjects apply what they have learned in a real-world situation. It would have been useful to address Kirkpatrick’s first level of evaluation by determining whether the treatment group found the instruction more appealing, engaging, and motivating than did the control group. Answering this type of question would contribute to our design knowledge (Reigeluth and Frick, 1999) and add further support for the use of simulations in education.

Population and sampling. The authors do not directly state the targeted population, but it may be inferred from the description of the simulation that they intend for their treatment to be used in teacher education programs that employ TWSM. The authors note that TWSM “has been adopted in other states and institutions around the country” (p. 482). However, they opted for a convenience sample from a single program and thereby diminished the external validity of their study. Convenience sampling is common in educational research, but it requires the caveat (which the authors provide) that replication of the study is necessary to determine the applicability of the treatment to other contexts. Frequencies for certain demographic variables are shown, but these are not compared to national figures for teacher education programs. The frequencies by gender, for example, indicate that the sample contained 38 males (54%) and 33 females (46%). By comparison, the American Association of Colleges for Teacher Education (AACTE, 2004) reported 21% males and 79% females in undergraduate education programs.

Instrumentation. The authors provide copies of the instruments in the appendices and describe their efforts to ensure that the instruments were valid and reliable. For the first two research questions (related to perceived skillfulness in connecting teaching and learning and perceived value of those skills), the subjects completed pre- and postmeasures with eleven items on a Likert scale for each research question. The two constructs were examined as separate factors, with the “Total Skillfulness” items achieving a Cronbach’s alpha of .82 and the “Value” items achieving a Cronbach’s alpha of .83, both acceptable levels of consistency.

For the third research question (pertaining to the quality of the teacher work samples), the subjects created teacher work samples before and after the treatment. These samples were evaluated by seven part-time adjunct instructors whose only responsibility was supervision of student teachers in the field. Evaluations were based on a rubric that had been used for several years. Cronbach’s alpha for the rubric was .88, again indicating good internal consistency. However, even though rater reliability was found to be high in the past, it had not been calculated in the past two years, leaving open the possibility of inconsistency among the raters.

For the fourth research question (teaching differently in a real setting), the subjects were evaluated on lessons taught in the field placement settings by their teacher mentors. The evaluation forms were designed to address five broad teacher competencies defined by state standards. The competency with the fewest items (*n* = 2) was the ability to evaluate pupil achievement, with a Cronbach’s alpha of .72. Because no significant differences between the groups were found for this competency, the Cronbach’s alpha is cause for concern. It may be that there were differences that were not detected due to the unreliability of the scale. The competency with the most items (*n* = 11) was the ability to establish a classroom climate conducive to learning, with a Cronbach's alpha of .89. Because a different teacher evaluated each student, there is a potential lack of reliability among the scores.

Data analysis and interpretation. The results of all measures are presented in tables of descriptive statistics that compare means and standard deviations for both groups from pretest to posttest as well as between groups from pretest to posttest. These scores were compared using analysis of covariance (ANCOVA) to examine factors that may have affected pretest scores. Neither gender nor GPA was found to be predictive of any outcome measures, but the authors acknowledge that there may have been other factors that influenced pretest scores and the effects of the treatment.

The analysis of work sample scores indicated a slightly greater improvement by the treatment group. The authors noted their concern about a significant decrease in standard deviations for the treatment group's work sample scores, which they hypothesized was due to a ceiling effect as the scores approached the maximum possible. The result would naturally affect the normality of the distribution and reduce the homogeneity of variance, so they used ANCOVA to control for pretest scores while examining the effect of the treatment.

Propositional knowledge claims. The primary claim in the article is that "the Cook simulation can play a role in the professional development, analytical sophistication, and success of teacher candidates as they face enormous pressures to facilitate student learning" (p. 493). This is a Type 3 claim regarding the effectiveness of a particular intervention, and it requires empirical evidence of the intervention's effectiveness, efficiency, and/or appeal. As discussed above, the authors provide several measures which indicate that the treatment had a significantly greater effect on the subjects' perceptions and abilities than did the traditional instruction. While the generalizability of the results may have been compromised by some aspects of the research design and methodology, the authors are careful to note, for example, that the simulation should not replace real-world practice but should supplement it.

There are also two related Type 4 claims regarding general means for teacher preparation and student learning. A Type 4 claim requires evidence of the instrumental value of the stated rules or principles. The first claim is that simulation is an effective pedagogy for teacher preparation. While this may be true, the authors would have done well to cite research in support of this claim. The second claim is that “one clear path to increased P-12 student learning is through improved teacher preparation” (p. 481). This claim is reasonable, although what is meant by “improved teacher preparation” may be argued. The claim is supported by two citations, an article published in a reputable journal by an author whose writings on educational leadership and accountability are widely cited and an unpublished paper prepared for the National Commission on Teaching and America’s Future meeting. It seems as though more and better evidence could be cited to support this claim.

Recommendation

There were several flaws in the study by Girod and Girod (2006) on the effectiveness of a simulation in enabling teacher candidates to practice connecting teaching and learning and to prepare for evidence-based assessment of their teaching. In particular, the sample selection and self-assignment for treatment diminish the generalizability of the findings and introduce possible confounding variables. However, no study is perfect, and the authors have noted the flaws and limitations of the study and have attempted to control for these as much as possible. Furthermore they have been careful not to overstate their findings. The article is well written and conforms to the expected format for the presentation of research. The topic is significant and worthy of study, and the theoretical framework for the intervention is sufficiently addressed. Overall the study makes a useful contribution to the fields of teacher preparation and instructional design.

References

AACTE. (2004). AILACTE characteristics: A report based on the 2002 Professional Education Data System of the American Association of Colleges for Teacher Education and the National Council for Accreditation of Teacher Education. Washington, DC: AACTE. (ERIC Document Reproduction Service No. ED484656). Retrieved November 27, 2007 from ERIC (Educational Resources Information Center database).

Creswell, J. W. (2005). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (2nd ed.). Upper Saddle River, NJ: Pearson.

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). Boston, MA: McGraw-Hill.

Frick, T. (2005). *Summary Table: Types of Propositional Knowledge Claims in Education: Elaboration of Typical Types of Researchers, and Primary Criteria for Evaluation.* Retrieved August 24, 2007, from http://www.indiana.edu/~istr690/frick07fall/resources/epistemologyclarified.doc.

Girod, M. & Girod, G. (2006). Exploring the efficacy of the Cook School District simulation. *Journal of Teacher Education, (57)*5, 481-497.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels.* San Francisco, CA: Berrett-Koehler.

Reigeluth, C. M., & Frick, T. W. (1999). Formative research: A methodology for creating and improving design theories. In C. M. Reigeluth (Ed.), *Instructional design theories and models: A new paradigm of instructional theory* (Vol. II, 633-651). Hillsdale, NJ: Lawrence Erlbaum Associates.