

# Understanding the Variable Effect of Instructional Innovations on Student Learning

Heidi L. Iverson

*CSU STEM Center, Colorado State University, Fort Collins, CO 80523-0001 USA*

**Abstract.** As a result of dissatisfaction with the traditional lecture-based model of education a large number of reform-oriented instructional innovations have been developed, enacted, and studied in undergraduate physics courses. While previous work has shown that the impact of instructional innovations on student learning has been overwhelmingly positive, it has also been highly variable. The purpose of this analysis is to investigate this variability. For this analysis, 79 published studies on undergraduate physics instructional innovations were analyzed with respect to the types of innovations used and the methodological characteristics of the studies themselves. The findings of this analysis have indicated that nearly half of the variability in effect size can be accounted for by study design characteristics rather than by the characteristics of the innovations used. However, a subsequent analysis illustrated that one specific innovation, Workshop/Studio Physics, appears to be particularly effective within the observed sample of studies.

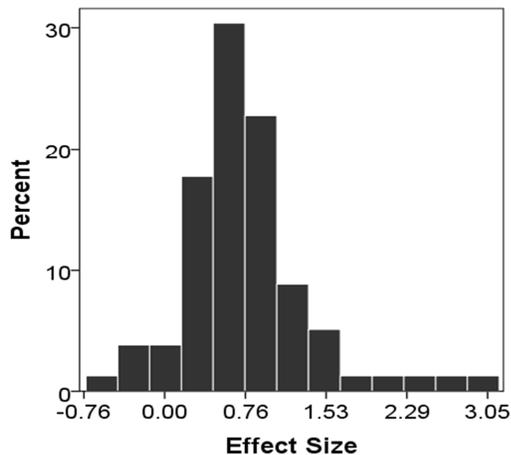
**Keywords:** Instructional Innovation, Synthesis, Meta-Analysis, Effect

**PACS:** 01.40.-d 01.40.Di 01.40.Fk 01.40.G- 01.40.gb

## INTRODUCTION

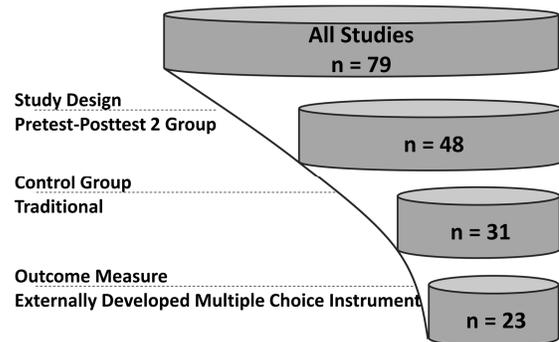
An increasing number of colleges and universities are moving away from the lecture-based model of physics instruction and moving towards the use of innovative instructional approaches intended to engage students in the learning process and help them take a more active role in their own learning.

Recent empirical studies have demonstrated that, in general, these novel instructional approaches have a positive impact on student achievement. For example, a recent meta-analysis has shown that the average effect size of undergraduate physics instructional innovations is a very large 0.76 [1]. However the effects found in these studies vary considerably, as illustrated in Figure 1.



**FIGURE 1.** Histogram of effect sizes from [1] centered on the mean (0.76); each bin is a half standard deviation (0.60).

The purpose of this study is to describe characteristics of the innovations that seem to have the largest impact on student performance. However, in order to do so it is crucial to first be sure that one is considering studies with similar methodological designs. In other words, while some of the observed variability in effect size shown in Figure 1 is due to the different types of innovations used, some amount of the variability may be due to three other factors which are controlled for in the present study: 1) study design, 2) control group, and 3) outcome measure. The first step in this analysis was to identify the largest possible pool of methodologically similar studies. This involved carefully whittling down the original pool of 79 studies – similar to the shape of a funnel. As the analysis proceeded, the pool of studies got smaller and smaller since only studies with similar methodological characteristics were kept. This approach is shown pictorially in Figure 2 and described in the sections below.



**FIGURE 2.** Visual depiction of funnel analysis

## FUNNEL ANALYSIS

### Study Design

There are three ways in which the 79 studies with effect sizes were carried out: (1) pretest/posttest with no control group ( $n = 8$ ), (2) posttest only with a control group ( $n = 23$ ), and (3) pretest/posttest with a control group ( $n = 48$ ). In the first type of study students' pretest scores are compared to their post-treatment scores. The major problem with this design is that it does not involve the use of a control group. Effect sizes calculated in this way [2] exaggerate the effect of the treatment. Furthermore, the underlying purpose of innovations is to improve upon the traditional model of physics instruction; without a control group this comparison is impossible.

In the other two study designs students' scores are compared across a treatment and control group which either involve a pretest and a posttest, or just a posttest. The pool of studies which include pretest and posttest scores for both a treatment and control group represent the largest group with 48 studies. This study design allows for more robust causal inferences to be made from the findings compared to studies that either do not include a control group or do not include pretest data, especially because the pretest conditions are taken into account in the effect size calculation [2]. Therefore studies which include pretest and posttest data across two groups represent the first cut-point in the analysis. Only these 48 studies were included in the subsequent analysis [Figure 2].

### Control Group Characteristics

In order to disentangle the possible differences in studies with different types of control groups, two types of control groups were coded: control groups with a traditional classroom experience ( $n = 31$ ), and those that experience something else (typically another innovation or some aspect of an innovation) ( $n = 17$ ). The largest group consisted of the 31 studies which use a traditional control group. This group allows for the best comparison across different innovations because the control group experience is held somewhat constant. While there are some variations amongst "traditional" classes, there are far more among the "other" treatment group category. In addition, one of the primary motivational factors for innovation has been to improve upon the traditional model of undergraduate physics education. Therefore studies which use a traditional control group allow for this comparison to be made directly. For this reason, only these 31 studies were included in the subsequent analysis [Figure 2].

## Outcome Measure Characteristics

Researchers use outcome measures which are either developed internally for the specific needs of a study, or those which have been developed externally, or both. There are several examples of papers in which the effect of a treatment is examined using two different outcome measures or just two variations of the same outcome measure. In these situations, the effect sizes reported for the same study can often be very different depending on the outcome measure used. For example, in one paper the Force Concept Inventory was used to evaluate the effect of an innovation but the authors reported their results in two different ways [5]: (1) using the full FCI, and (2) using a sub-section of the FCI which consisted only of a few items which were pertinent to the particular content covered in the study. When the full FCI was used to calculate the effect of the innovation the effect size was 0.65. However when the sub-section of the FCI was used, the effect size was 1.21. Clearly, as this study indicates, the outcome measure used in a study has an impact on effect size.

Within the sample of studies, those which involved internally developed outcome measures had higher mean effect sizes ( $n = 5$ , mean effect size = 1.09) than those using externally developed outcome measure ( $n = 26$ , mean effect size = 0.69). The difference across studies using different outcome measures may be because internally developed outcome measures are often very closely linked to the intended purpose of the innovation and tend to be more sensitive to changes in students' performance [4], as illustrated by the example in the previous paragraph.

An additional characteristic of outcome measures has to do with the type of items on the instrument. Within the 31 studies in this part of the analysis there were two different item types: multiple-choice ( $n = 28$ ) and open-ended ( $n = 3$ ). The three studies which used an outcome measure with open-ended items had a much lower mean effect size (0.26) than the 28 studies which used multiple-choice items (0.80). The three studies with open-ended items all came from the same paper in which the effect of three slightly different innovations was investigated. The outcome measure consisted of a single item for which student responses were scored prior to and after instruction. Single item instruments pose an issue for the validity and reliability of the measure, as well as for the study. Therefore these three studies were eliminated in the third cut.

In summary, the third cut of studies was made based on the characteristics of the outcome measure used. At this cut-point studies were eliminated if the outcome measure was not a multiple-choice test which

had been externally developed. There were 23 studies that were similar with respect to the methodological characteristics described above [Figure 2].

### Summary of Funnel Analysis

The preceding analysis helped to identify three methodological factors which impact the magnitude of the effect size of a study: study design, control group type, and outcome measure type. The results of a linear regression analysis shows that these factors account for 47% of the variance in effect size [6]. By reducing the overall pool of papers down to a methodologically similar sub-pool, the effects of these factors can be controlled for and the relationship between innovation type and effect size can be examined.

### ANALYSIS OF INNOVATION TYPE

Two different approaches to categorizing the types of innovations within the pool of papers were used. The first approach was to categorize the innovations into four non-mutually exclusive types developed and applied in previous research [1]: conceptually oriented tasks (COT), collaborative learning (CL), technology (Tech), and inquiry-based tasks (IBP). The second approach was to categorize innovations with respect to any known instructional innovations being used. These are the well-known innovations in physics education research, such as Peer Instruction and Tutorials.

When the 23 methodologically similar studies are analyzed with respect to the two aforementioned categorization approaches, two results stand out. First – studies that involve the use of collaborative learning combined with conceptually oriented tasks and technology (CL+COT+Tech,  $n = 10$ ) have a very large average effect size of 0.96 compared to the overall mean effect size of 0.77. Second – studies that involve the known instructional innovation Workshop/Studio Physics ( $n = 8$ ) also have a large average effect size of 1.02.

A closer look at the two groups of innovations with large effect sizes revealed that they actually involved the same studies. In other words, the 8 studies which involved the use of Workshop/Studio Physics were also in the pool of studies coded as involving the use of collaborative learning, conceptually oriented tasks, and technology. Furthermore, all of the collaborative learning, conceptually oriented tasks, and technology innovation studies that are above the mean of 0.77 are also Workshop/Studio Physics. This evidence indicates that the Workshop/Physics model has a high positive effect on student learning on average. Table 1 illustrates the overlap among the two different

categorization schemes used to analyze the data. The table lists studies in order of increasing effect size in order to illustrate visually how studies involving Workshop/Studio Physics have particularly high effect sizes.

**Table 1.** Pool of 23 methodologically similar studies, their effect sizes, and innovation type [7].

Study number	Effect Size	CL, COT, & Tech	Workshop Physics
1. Cheng et al. (2004)	-0.24		
2. Larkin (2005)	0.01		
3. Zhou et al. (2005)	0.49		
4. Wick et al. (2004)	0.49		
5. Hoellwarth et al. (2005)	0.52	X	X
6. Wick et al. (2004)	0.52		
7. Chang (2005)	0.52		
8. Chang (2005)	0.55		
9. Redish et al. (1997)	0.70	X	
10. Redish et al. (1999)	0.72		
11. Redish et al. (1997)	0.73	X	
12. Lenaerts et al. (2003)	0.75		
13. Chang (2005)	0.75		
14. Lasry et al. (2007)	0.79		
15. Sorensen et al. (2006)	0.89	X	X
16. Redish et al. (1999)	0.92	X	X
17. Beichner et al. (1999)	0.96	X	X
18. Cataloglu (2007)	0.99		
19. Beichner et al. (2007)	1.02	X	X
20. Beichner et al. (1999)	1.02	X	X
21. Johnson et al. (2001)	1.16		
22. Hoellwarth et al. (2005)	1.17	X	X
23. Hoellwarth et al. (2005)	1.62	X	X
Total Average ( $n = 23$ )	0.77		

### CONCLUSIONS & FUTURE WORK

The goal of this analysis was to understand the variable impact of undergraduate physics instructional innovations on student performance. This study took a closer look at some of the factors which might account for the large variability of innovation effectiveness found in a previous meta-analysis. The ultimate goal was to understand the relationship between the aspects of what happens in the classroom and the effect that it has on student performance. However, in order to investigate this relationship it was necessary to filter the overall pool of studies down to a methodologically similar pool of studies since nearly 50% of the variability in effect sizes could be explained based only on how the studies were carried out and not actually as a function of the innovations themselves.

Within the pool of methodologically similar papers analyzed in this research, there was a particular type of

innovation that stood out – Workshop/Studio Physics [8 & 9]. In order to gain a more comprehensive understanding of the critical features of the Workshop/Studio Physics instructional model a case study analysis has been conducted, the result of which will be described in future publications.

## ACKNOWLEDGMENTS

Many thanks to Ayita Ruiz-Primo, Derek Briggs, Robert Talbot, and Lorrie Shepard. Also thanks to Erin Furtak, Valerie Otero, and Noah Finkelstein. A portion of this research was funded by the Integrating STEM initiative at the University of Colorado at Boulder Graduate Student Fellowship program.

## REFERENCES

1. Ruiz-Primo, M.A., Briggs, D., Iverson, H.L., Talbot, R., & Shepard, L.A., *Science*, 331(6022), 1269-1270, 2011; Iverson, H.L., unpublished dissertation, March 28, 2011; Iverson, H.L., Briggs, D.C., Ruiz-Primo, M.A., Talbot, R.M., & Shepard, L.A., *Proceedings of the PERC 2009*, Ann Arbor, MI.

2. Effect sizes were calculated in three ways corresponding to the three different study designs. For pretest/posttest without a control group studies the following equation was used:  $ES_{SingleGroup} = \frac{\bar{X}_{Post} - \bar{X}_{Pre}}{SD_{Pre}}$ .

For studies involving a posttest with a control group design the equation used was:  $ES_{Glass's \Delta} = \frac{\bar{X}_T - \bar{X}_C}{SD_C}$ .

For studies involving a pretest and a posttest with a control group the following equation was used:

$$ES_{Pre-Post-Test Two Groups} = \frac{(\bar{X}_{T\_Post} - \bar{X}_{T\_Pre}) - (\bar{X}_{C\_Post} - \bar{X}_{C\_Pre})}{SD_{Pooled\_Pre}}$$

where  $\bar{X}$  represents a test score mean for treatment and control conditions (subscripts “T” and “C”) administered at the beginning and end of a study period (subscripts “Pre” and “Post”), and  $SD$  is computed as the weighted average of the standard deviations across treatment and control groups with the equation

$$SD_{Pooled\_Pre} = \sqrt{\frac{(n_{T\_Pre} - 1)SD_{T\_Pre}^2 + (n_{C\_Pre} - 1)SD_{C\_Pre}^2}{(n_{T\_Pre} - 1) + (n_{C\_Pre} - 1)}}$$

where  $n$  represents the sample size.

3. Hoellwarth, C., Moelter, M.J., & Knight, R.D., *American Journal of Physics*, 73, 459, 2005.
4. Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L., & Klein, S. *Journal of Research in Science Teaching*, 39(5), 369-393, 2002.
5. Redish, E., Saul, J., & Steinberg, R. *American Journal of Physics*, 65(1), p. 45-54, 1997.
6. A linear regression analysis was used to evaluate the degree to which these factors account for the variability in effect size.  $R^2 = 0.47$ ,  $RMSE = 0.45$ . Parameter estimates and significance:

	Parameters	Estimates	Sig.
	Intercept	0.75	0.00
Cut 1	Pretest-Posttest no control	1.11	0.00
	Posttest Only with control	-0.38	0.02
Cut 2	“Other” Control Group	-0.01	0.95
Cut 3	Internally Developed	0.10	0.50
	Open-ended Items	-0.21	0.19

7. Cheng, K.K., Thacker, B.A., Cardenas, R.L., & Crouch, C. *American Journal of Physics*, 72, 1447, 2004; Larkin, T. L. *Journal of STEM Education*, 6, 1-2, 2005; Zhou, G.G., Brouwer, W., Nocente, N. & Martin, B. *Journal of Interactive Learning Research* 16.1, 31, 2005; Hoellwarth, C., Moelter, M.J., & Knight, R.D., *American Journal of Physics*, 73, 459, 2005; Wick, D.P., & Ramsdell, M.W., *American Journal of Physics*, 72, 863, 2004.; Chang, W. *International Journal of Science Education*, 27(4), 0950-0693, 2005; Redish, E., Saul, J., & Steinberg, R., *American Journal of Physics*, 65(1), 45-54, 1997; Redish, E.F. & Steinberg, R.N., *Physics Today*, 52(1), 24, 1999; Lenaerts, J., Wieme, W., & Van Zele, E. *European Journal of Physics*, 24(1), 7-14, 2003; Lasry, N., & Aulls, M.W. *American Journal of Physics*, 75, 1030, 2007; Sorensen, C.M., Churukian, A.D., Maleki, S., & Zollman, D.A. *American Journal of Physics*, 74, 1077, 2006; Beichner, R., Bernold, L., Burniston, E., Dail, P., Felder, R., Gastineau, J., Gjertsen, M., & Risley, J. *American Journal of Physics*, 67, S16, 1999; Cataloglu, E. *European Journal of Physics*, 28(4), 767-776, 2007; Beichner, R. J., Saul, J. M. , Abbott, D. S. , Morse, J. J., Deardorff, D. L. , Allain, R. J. , Bonham, S. W. , Dancy, M. H., & Risley, J.S. *Reviews in PER* Vol. 1, 2007; Johnson, M. *American Journal of Physics*, 69, S2, 2001.
8. Although the pool of papers included in the present analysis were gathered in an exhaustive manner and steps were taken to ensure that all eligible papers were retrieved through these methods, a few papers may have been missed (please see Ref. 1 for more information). However, there is no reason to believe that the sample of papers gathered is somehow systematically different from the overall population of papers that exist which fit our criteria. Therefore it seems likely that they are representative of the field. However, care must be taken in assuming that these studies are representative of all classroom-based innovations and that the results of this analysis are generalizable to all college classrooms. In addition, it is likely that there is some publication bias among the studies in terms of their effectiveness because studies with null or negative effects are infrequently published.
9. Some readers may be curious as to why an additional cut wasn't made for the type of outcome measure. However, the effect size statistic is robust for comparisons across different outcome measures because scores are standardized (Lipsey & Wilson, 2001, *Practical Meta-Analysis*). When the analysis was repeated with the 15 studies that used the same outcome measure (the FCI), the results of the previous analysis held - i.e. studies involving Workshop/Studio Physics had a larger mean effect size ( $n = 6$ , mean effect size = 0.89) than the other studies ( $n = 9$ , mean effect size = 0.55).