A LIVELY ELECTRONIC COMPENDIUM OF RESEARCH, NEWS, RESOURCES, AND OPINION

Astronomy Education Review

Volume 6, Apr 2007 - Nov 2007
Issue 1

Analysis of the Astronomy Diagnostic Test

by **Erik Brogt**
University of Arizona
Darrell Sabers
University of Arizona
Edward E. Prather
University of Arizona
Grace L. Deming
University of Maryland, College Park
Beth Hufnagel
Anne Arundel Community College
Timothy F. Slater
University of Arizona
Received: 01/31/07, Revised: 04/13/07, Posted: 06/14/07

The Astronomy Education Review, Issue 1, Volume 6:25-42, 2007

Abstract

Seventy undergraduate class sections were examined from the database of Astronomy Diagnostic Test (ADT) results of Deming and Hufnagel to determine if course format correlated with ADT normalized gain scores. Normalized gains were calculated for four different classroom scenarios: lecture, lecture with discussion, lecture with lab, and lecture with both lab and discussion. Statistical analysis shows that there are no significant differences in normalized gain among the self- reported classroom formats. Prerequisites related to mathematics courses did show differences in normalized gain. Of all reported course activities, only the lecture and the readings for the course correlate significantly with the normalized gain. This analysis suggests that the ADT may not have enough sensitivity to measure differences in the effectiveness of different course formats because of the wide range of topics that the ADT addresses with few questions. Different measures of gain and their biases are discussed. We argue that the use of the normalized gain is not always warranted because of its strong bias toward high pretest scores.

1. INTRODUCTION

Conceptual diagnostic tests can be used to measure course effectiveness by assessing student understanding about a particular concept both before and after instruction. This is often called pretest and posttest design. Student achievement is measured prior to and after instruction, and a gain in students' scores as a result of instruction is calculated. In the context of physics education research, one of the most commonly used diagnostic tests is the Force Concept Inventory (FCI; Hestenes, Wells, & Swackhamer 1992). In a large meta-study, Hake (1998) obtained results from more than 60 courses, encompassing more than 6,000 students who were surveyed with the FCI, the Mechanics Baseline Test (MBT; Hestenes & Wells 1992), or the Mechanics Diagnostic test (MD; Halloun & Hestenes 1985). He used these results to measure the effectiveness of interactive engagement and traditional lecture-based course formats. Hake showed that interactive engagement methods in physics led to higher gains than traditional lecture-based methods.

In astronomy, tools like the FCI, MBT, and MD are less common. Furthermore, the student population is significantly different. Students taking introductory physics courses in which these physics conceptual diagnostics are administered are typically science, engineering, or pre-med majors. Most of these students are required to take the introductory physics series as a prerequisite for their degree program. The vast majority of students taking an introductory astronomy course are non–science majors fulfilling a general education science requirement; the course often will serve as their terminal course in science.

The most commonly used diagnostic to date in introductory astronomy courses is the Astronomy Diagnostic Test (ADT; Zeilik 2003). The ADT has 21 multiple-choice content questions covering a wide range of astronomy topics and is aimed at the introductory-level courses typically taught to non–science majors at colleges and universities. In addition, the ADT includes 12 demographic questions. Deming and Hufnagel (2001) constructed a database that contains more than 5,000 students' pretest scores and over 3,500 students' posttest scores on the ADT version 2.0. In addition, a vast array of instructor-reported information about the courses is available in the database.

In this study, we are interested in pursuing two questions: Are there differences in gain between the different course formats? Can we identify and quantify additional variables that may help predict student gains? It should be noted that this study is not similar to the Hake (1998) study in the sense that we do not make a distinction between interactive engagement and traditional lecture-based formats. There is not enough information in the database or other materials (Deming 2002) to suggest that the different course formats have true interactive engagement elements in them. Although we do not discount such an option, it is not an a priori assumption in this study.

The article is set up as follows: In Section 2, we briefly discuss the structure of the ADT database, in Section 3 we describe the methods we used, and in Section 4, we present the results. In Section 5, we discuss a further analysis on measures of gain. In Section 6, we discuss the results guiding our conclusions and offer recommendations for further study.

2. THE ADT DATABASE

The ADT database contains information about more than 5,000 pretest and 3,500 posttest results obtained from approximately 100 classrooms across the United States, reflecting a wide variety of institutions. In addition to the student responses to individual questions on the ADT and the total number of correct items, the database contains instructor-reported information about the following items: geography, institution type and size, class size and format, type of course (Solar System, universe in one semester, and so on), math prerequisite for the course, and information on how well the course topics align with ADT questions. To maintain protections afforded by human subjects policies, the database we worked with was completely absent of variables that might identify individual students. In this study, we were interested in the pretest and posttest scores as a function of course format. Of all the formats listed, four were useful for this analysis: lecture alone, lecture with mandatory laboratory, lecture with mandatory discussion or recitation sessions, and lecture with both laboratory and discussion sessions. We reduced the data set to contain only those entries that had these class formats, with 70 out of 100 classes meeting the requirements for this study.

3. METHODS

The participants for the collection and submission of ADT results for the database were instructors who volunteered to administer the ADT to their students at the beginning and/or end of their undergraduate introductory astronomy survey courses. As such, the sample represents one of convenience rather than a true random sample, and many instructors were obtained by personal contacts of the ADT design team. Students' pretest and posttest data are not matched; this restriction was imposed by removing identifying characteristics and partly by attrition in student numbers in the classes over the semester, as indicated by the difference in the number of pretests and posttests administered. We calculated class mean prescores and postscores and the normalized gain (Hake 1998) for each class, which is defined as

Normalized gain = (% post - % pre) / (100 - % pre)

We then averaged the normalized gains per instructional format. We have 16 classes characterized by lecture alone, 5 for lecture with discussion, 40 for lecture with lab, and 9 for lecture with both lab and discussion. If a course has multiple components, it is likely that a wider variety of student learning styles are being served. Based on the variety of opportunities to learn, we predicted that the lowest gains would be in the lecture-only format, and the highest gains would be in the lecture with both lab and discussion format. The lecture with only lab, and the lecture with only discussion formats were predicted to have gains between those two extremes. This assumption allowed us to do one-tailed tests, increasing the statistical power.

We chose a family-wise alpha level of $\alpha_{FW} = .05$. This means that the overall chance of finding significance when in fact the result is attributed to random chance is 5%. For the analysis, we used the Holm-Bonferroni planned contrast method. We chose this method because it is appropriate for the unequal sample sizes in the data obtained for this analysis. The high statistical power comes with a price: One is required to plan all contrasts prior to analysis to keep the alpha slippage (increasing the chance of claiming a significant result when it is not warranted) for the entire set of contrasts under control. Independent sample *t* tests are done for each contrast. One can argue that there is considerable overlap between the students doing the pretest and the posttest in each class and that an independent sample *t* test would lead to a decrease in statistical power. However, because there is no information available on which students did

the pretest and posttest, using an independent sample *t* test is the most conservative estimate that one can make. The resulting *p* values (the probability that the result, in this case the difference in gains, is the result of random chance rather than an actual effect) are rank ordered, with lowest value first, and compared with the threshold value. Because the first contrast is evaluated at an α level of $\alpha=(\alpha_{FW}$ / total number of contrasts), it is important to keep the number of contrasts low to increase the statistical power of the test. Each subsequent evaluation in this method is run at a slightly higher alpha level (denominator goes down by one for each evaluation), but it requires a statistically significant previous evaluation. When one of the evaluations yields a nonsignificant result, the subsequent evaluation will not be significant either. For this reason, we decided not to evaluate the contrast dealing with lecture with lab, and lecture with discussion because we could not a priori make a reliable prediction of which of those formats would yield a higher gain. Table 1 summarizes the planned contrasts, which are set up in the following form:

Gain (format 1) – Gain (format 2) < 0

Table 1. Planned contrasts evaluations

Format 1	**Format 2**
Lecture	Lecture + discussion
Lecture	Lecture + lab
Lecture	Lecture + lab + discussion
Lecture + lab	Lecture + lab + discussion
Lecture + discussion	Lecture + lab + discussion

Because we have five contrasts, the threshold for significance for the first contrast is $\alpha=.05/5 = .01$.

3.1 Other Variables in the Analysis

The database contains several variables that could potentially influence the normalized gain as well. Three variables had an a priori high face validity for further investigation. Those were class size, math prerequisite for the course, and the extent to which the course content mapped onto the topics covered in the ADT. All these data were self-reported by the course instructors or listed on the course syllabi. In the project that created the database used in this project, not all instructors reported all these variables. Therefore, because not all course formats in the database had data associated with these variables, it was not possible to use them as covariates; too many classes would have been eliminated from the analysis. Instead, we used a simple correlation to measure the effects of the variable on the entire set of classes.

3.1.1 Class size

Class sizes in the database varied from only a few students to over 300. In larger classes, it is generally accepted that students will be more anonymous than in smaller classes. This could lead to a lessened sense of relatedness to the class, one of the three fundamental ingredients for intrinsic motivation (Deci & Ryan 1985). Although one could argue that smaller lab or discussion sections would partly negate this effect, we expected to see a slight negative correlation between normalized gain and class size.

3.1.2 Math Prerequisite

In traditional instruction of introductory astronomy for non–science majors, some emphasis is placed on mathematical operations, usually in the form of solving algebraic equations and interpretation of graphs, as evidenced by introductory astronomy textbooks. This is poised to present a problem for those students with math anxiety and/or limited math skills. The courses in the original ADT database are coded for mathematics prerequisites. We expected a difference in gain scores between classes that have a formal university-level math prerequisite (algebra and trigonometry) and those that do not. We expected the former to have a higher normalized gain than the latter.

3.1.3 Course Content

In the original ADT study (Deming 2002; Hufnagel 2002), instructors were asked to rate on a scale from 1 to 11 the extent to which they thought that the different parts of their course (reading, lecture, homework, activities, and lab) aligned with the items on the ADT. The alignment does not indicate what fraction of the course was actually spent on topics covered on the ADT. However, we use the reported alignment as a first-order approximation because it seems reasonable to assume that a course with a higher reported alignment will produce a higher normalized gain than a course with a lower reported alignment. Because not all course designs had a sufficient number of classes to make a stratification, we aggregated all classes and calculated the Pearson correlation coefficients (a measure for a linear relation) between the different elements of the course and the normalized gain.

4. RESULTS

Summary statistics for the class formats can be found in Table 2. In the subsections below, we discuss the various results in more detail.

Table 2. Summary statistics for the different course formats

	Lecture	**Lecture + discussion**	**Lecture + lab**	**Lecture + lab + discussion**
# classes	16	5	40	9
# students pretest	1045	549	1730	723
Mean prescore (21 max)	6.65	6.09	6.70	7.31
Standard deviation	.71	.28	1.09	.90
# students posttest	758	369	1371	582
Mean postscore (21 max)	9.66	8.60	9.77	11.44
Standard deviation	1.61	1.13	1.88	2.01
Mean normalized gain	.2098	.1681	.2146	.3016
Standard deviation on normalized gain	.0870	.0628	.0973	.1221
Standard error on normalized gain	.0217	.0281	.0154	.0407
Lower limit 95% confidence interval	.1672	.1131	.1845	.2218
Upper limit 95% confidence interval	.2525	.2231	.2448	.3814

4.1 Homogeneity of the Class Formats

The original Hake (1998) study examined distinct populations: high school, college, and university students in both interactive engagement (IE) or traditional course format. In all populations, the normalized gains for the IE classes are higher than the gains achieved by traditional classes. This is shown in Hake's plot, pretest percentage score (the percentage of questions on the FCI or ADT answered correctly before instruction) versus normalized gain, in Figure 1. Moreover, the traditional classes in Hake's original study also occupied distinct areas in the plot, indicative of different populations (high school, college, and university students). In Figure 2, we plotted our data in a similar fashion to Figure 1. The different class formats show overlap, which we interpret to be indicative of a more homogeneous population. This result is not particularly surprising because the database contains only information about introductory astronomy students at the college/university level, making the population much more homogeneous than the populations in the original Hake study (Deming 2002; Hufnagel 2002).

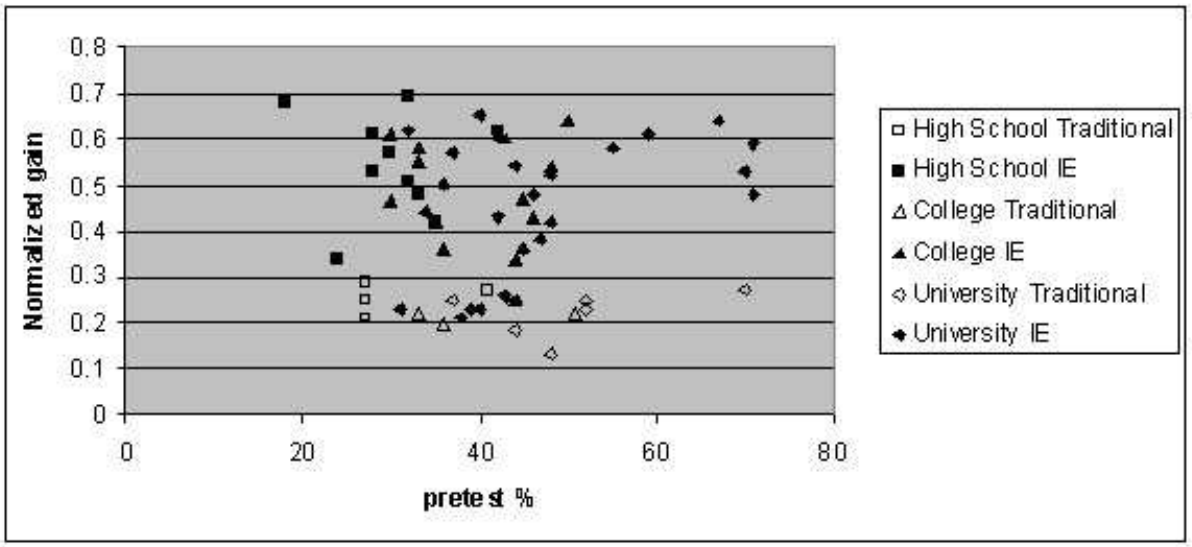

Figure 1. The Hake distribution of classes. Note that the different populations barely overlap. (Adapted from E. F. C. Dokter & S. R. Buxner, pers. comm.).

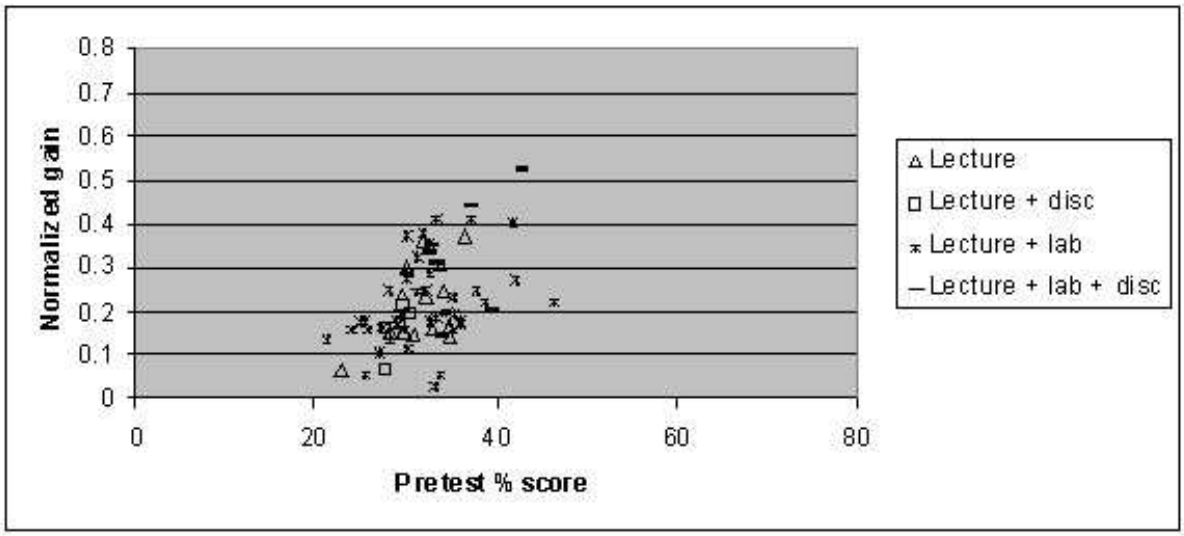

Figure 2. The ADT distribution of classes. Note that the different course formats do overlap.

4.1.2 Shape of the distributions

For all classes, we calculated skew (γ_1) and kurtosis (γ_2) of the distribution of scores (see Appendix A). The skew value measures in what direction the distribution is tailed, with $\gamma_1 < 0$ meaning that the distribution has a tail to the left, and $\gamma_1 > 0$ meaning that the distribution is tailed to the right. The kurtosis

is a measure of the flatness of the distribution, with $\gamma_2 > 0$ meaning the distribution has a high peak, and $g_2 < 0$ meaning that the distribution is less peaked. Overall, the classes showed a shift from pretest to posttest toward lower values for both γ_1 and γ_2. This is consistent with learning taking place (shift to the right in scores), but not everyone is learning at the same rate (flattening of the distribution and a larger standard deviation posttest as compared with pretest). Because in our study, an entire class is the unit of analysis, we did not consider the skew and kurtosis and their effects on the normality assumption in statistical tests of the distribution of an individual class. However, in an analysis on the classroom level, skew and kurtosis should be considered because they can undermine the assumptions of normality that underlie most statistical analyses.

4.2 Differences in Normalized Gain as a Function of Teaching Format

Results for the statistical tests using five planned contrasts are given in Table 3. The results show that even the first contrast is not significant, meaning that the other contrasts are not significant either. For one class that included lecture plus discussion, we noticed that the data showed extremely low gain that severely impacted the mean gain of the group (the sample size of this group is only 5). We recalculated the contrasts leaving out this anomalous value in Table 4, and again, no contrasts were significant.

Table 3. Ranked planned contrasts with obtained and critical (Holm-Bonferroni) *p* values

Format 1	Format 2	Obtained *t* value	Obtained *p* value	Rank of contrast	Critical *p* value
Lecture	Lecture + discussion	1.105 *	.16 (.84)	5	N/A
Lecture	Lecture + lab	-.202	.42	4	N/A
Lecture	Lecture + lab + discussion	-2.24	.017	2	N/A
Lecture + discussion	Lecture + lab + discussion	-2.31	.19	3	N/A
Lecture + lab	Lecture + lab + discussion	-2.35	.012	1	.01

Critical value for the first contrast is $p = .01$.
* The obtained *t* value indicates that it is located in the opposite tail of the distribution in the one-tail analysis, hence the ranking of 5.

Table 4. Ranked planned contrasts after the anomalous value in the group Lecture+Discussion was removed

Format 1	Format 2	Obtained *t* value	Obtained *p* value	Rank of contrast	Critical *p* value
Lecture	Lecture + discussion	.662 *	.26 (.73)	5	N/A
Lecture	Lecture + lab	-.202	.42	4	N/A
Lecture	Lecture + lab + discussion	-2.24	.017	3	N/A
Lecture + discussion	Lecture + lab + discussion	-2.60	.014	2	N/A
Lecture + lab	Lecture + lab + discussion	-2.35	.012	1	.01

Critical value for the first contrast is $p = .01$.
* The obtained *t* value indicates that it is located in the opposite tail of the distribution in the one-tail analysis, hence the ranking of 5.

4.3 Additional Variables

4.3.1 Class size

We plotted the normalized gain as a function of class size in Figure 3. A bivariate correlation yielded a nonsignificant Pearson *r* correlation coefficient (a measure for a linear relationship) for this distribution of $r = .05$. Class size does not appear to be a significant factor in predicting normalized gain scores in classes up to 50 students. The larger classes are not sufficiently sampled to draw a firm conclusion.

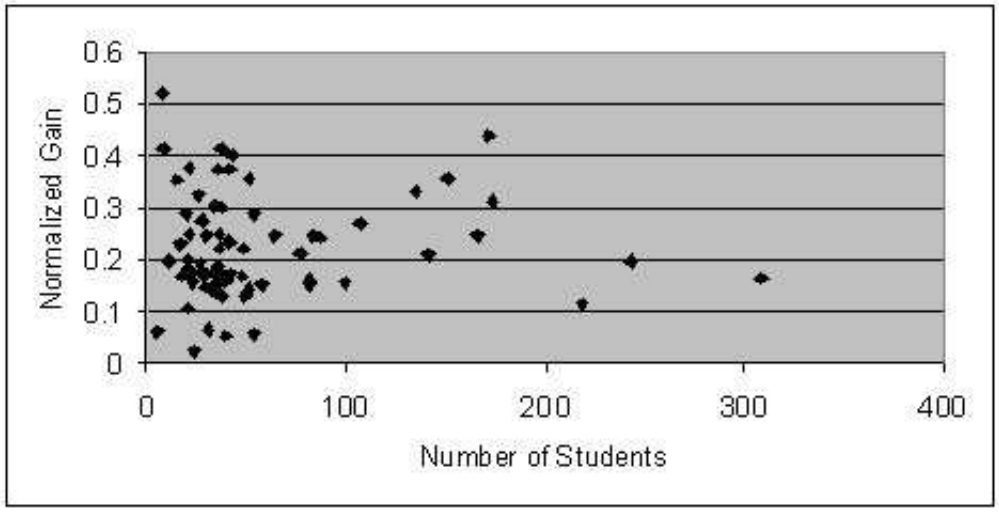

Figure 3. Normalized gain distribution as a function of class size.

4.3.2 Math prerequisite

The courses in the database are coded for a prerequisite in mathematics. Table 5 shows the results of a one-tailed independent sample *t* test of the average normalized gain in courses that did have a math prerequisite, and those that did not have a prerequisite yielded a significant ($p < .01$) difference. The Cohen's *d* effect size (in essence, the difference between the means in units of standard deviation) was calculated using the formula of Rosenthal and Rosnow (1991):

$$d = t\,(n1 + n2) / (df * n1 * n2)^{½}$$

in which *t* is the obtained *t* value, *n1* and *n2* the sample sizes and *df* the degrees of freedom. The effect size in Table 5 shows that we are dealing with a medium to large effect, keeping in mind that Cohen's classification of $d = .8$ as being a large effect should not be used as an absolute benchmark, following Thompson (2007).

Table 5. Results for the independent sample *t* test between math prerequisite and normalized gain

Prerequisite	***N***	**Mean normalized gain**	**Obtained *t* value**	***p* (one-tailed)**	**Cohen's *d* effect size**
Algebra + trigonometry	32	.26	2.796	.004	.70
No math prerequisite	33	.19			

We checked if this result was due to a higher pretest level of student content knowledge in the course that had a mathematics prerequisite. A one-tailed independent sample *t* test yielded a significant ($p < .05$) result. The Cohen's *d* effect size indicates that this is a medium effect. The results are summarized in Table 6. At least part of the difference in normalized gain between classes that had a math prerequisite and those that did not can be explained by the difference in pretest scores.

Table 6. Results for the independent sample *t* test between math prerequisite and pretest score

Prerequisite	*N*	Raw mean prescore	Obtained *t* value	p (one-tailed)	Cohen's *d* effect size
Algebra + trigonometry	32	6.92	1.741	.043	.44
No math prerequisite	33	6.50			

4.3.3 Course content mapping

Using a scale from 1 to 11, instructors self-reported the alignment of a course element with items covered on the ADT. We correlated the reported alignment on the various course elements with the normalized gain. However, in the database, some of the fields for a class were left blank, whereas others had the value zero. It was not clear whether a zero actually meant "not related at all to any item on the ADT" or if it simply was another way of denoting missing data (normally, fields missing data are left blank). Therefore, we calculated the Pearson *r* coefficients twice in Table 7: once with the original database, in which only blank values were ignored in the analysis, and once in which all the zero values were also ignored.

Table 7. Pearson *r* coefficients for the correlation of various parts of a class with the normalized gain

Form of content delivery	*r* (zeros included)	*r* (zeros ignored)
Reading	.22 (.40 **)	.19 (.40 **)
Lecture	.38 **	.38 **
Homework	.31 *	.02
Activity	.10	.05
Lab	.12	.23

* Significant at the .05 level. ** Significant at the .01 level.

One class in particular stood out. Class number 69 (see Appendix A) reported a rating of 1 (out of 11) for the reading on the ADT, yet has a normalized gain of 0.52. We judged this to be an anomaly. Leaving out this anomalous value leads to the Pearson *r* value reported in brackets.

Normally, one would expect combinations of factors—for example, a high score on the content mapping for reading and lecture—to have an effect as well. However, because all the data are self-reported, almost certainly leading to inconsistent values attached to similar mappings, we did not investigate such interactions.

5. The Use of Different Estimators

The normalized gain is biased toward high pretest scores. It is thus possible to find statistical significance between two normalized gains, which is an artifact of the different pretest scores. To investigate the effect of bias, we modeled three different measures of gain. These different measures of gain are biased toward different regions of pretest scores. If one finds statistical significance in one measure but not in others, the results can be suspect. However, if one finds significance on a multitude of measures, or if one fails to find significance on a multitude of measures, a much more compelling case can be made regarding to the validity of the results.

We evaluated the following measures of gain:

- Hake's normalized gain, defined as: gain = (post - pre) / (100 - pre)
- Gain 2, defined as: gain = (post - pre) / (post + pre)
- Gain 3, defined as: gain = (post - pre) / pre

For a detailed analysis of the biases of these measures of gain, see Appendix B.

A correlation between the pretest and posttest scores (Table 8) shows the biases involved in a different way. It is clear that there is a strong linear relation between pretest score and normalized gain.

Table 8. Correlations between pretest percent score and the various measures of gain

	Posttest %	**Normalized**	**Gain 2**	**Gain 3**
Pearson *r*	.754	.469	.018	.025

Note the strong correlation between the normalized gain and the pretest value.

To investigate the effect of these biases on our data, we used the different measures of gain to recalculate the planned contrasts in order to see if one of them would yield significance. The results of the planned contrasts analysis are given in Tables 9 and 10. No significance for any contrast was found with the estimators Gain 2 and Gain 3.

Table 9. Obtained *p* values using Gain 2

Format 1	Format 2	Obtained *t* value	Obtained *p* value	Rank of contrast	Critical *p* value
Lecture	Lecture + discussion	.466 *	.33 (.67)	5	N/A
Lecture	Lecture + lab	-.178	.43	3	N/A
Lecture	Lecture + lab + discussion	-1.61	.06	1	.01
Lecture + discussion	Lecture + lab + discussion	-1.63	.065	2	N/A
Lecture + lab	Lecture + lab + discussion	1.48 *	.07	4	N/A

Critical value for the first contrast is $p = .01$.
* The obtained *t* value indicates that it is located in the opposite tail of the distribution in the one-tail analysis, hence the ranking at the bottom of the list.

Table 10. Obtained *p* values using Gain 3

Format 1	Format 2	Obtained *t* value	Obtained *p* value	Rank of contrast	Critical *p* value
Lecture	Lecture + discussion	494 *	.31 (.69)	5	N/A
Lecture	Lecture + lab	-.245	.40	3	N/A
Lecture	Lecture + lab + discussion	-1.63	.06	2	N/A
Lecture + discussion	Lecture + lab + discussion	-1.71	.06	1	.01
Lecture + lab	Lecture + lab + discussion	1.48 *	.07	4	N/A

Critical value for the first contrast is $p = .01$.
* The obtained *t* value indicates that it is located in the opposite tail of the distribution in the one-tail analysis, hence the ranking at the bottom of the list.

6. CONCLUSIONS

Based on the results listed in the previous section, we reached the following conclusions. First, there are no significant differences in normalized gain between the four course formats. This can be interpreted in two ways. First, one can argue that the ADT only contains 21 questions that cover a wide range of astronomy topics. The ADT is thus not as tightly focused on a sample of related concepts as the FCI is. As such, the ADT cannot be considered a true diagnostic tool in the same sense of the FCI (Hestenes et al. 1992). There are probably not enough questions per concept covered in a typical introductory astronomy class to adequately probe student understanding of any one particular concept, if the ADT covers the concept at all. The low resolution of the ADT may thus influence the relatively low normalized gains that were observed. Gains lower than 0.3 are considered to be in the low region according to Hake (1998); the medium region is between 0.4 and 0.7, and gains larger than 0.7 are considered large. Only five classes (numbers 22, 28, 53, 67, and 69 in Appendix A) score a medium gain, and the rest of the classes are in the low region. Because of these low gains and low final scores (around 50%), there is ample room for growth, both positive and negative. This means that ceiling and floor effects in all the measures of gain are negligible.

Another way to interpret the results is via the argument that the four different formats that we investigated here are instructionally equivalent; all are essentially instructor-centered formats, without explicit interactive engagement elements in the courses in the sense of the Hake (1998) study. Therefore, it should not come as a surprise that all observed gains are statistically equivalent because one can argue that the

only pedagogically relevant variable among the courses potentially is time on task.

Second, the use of the normalized gain as a measure for course effectiveness may be suspect. The normalized gain is biased toward high pretest scores, as indicated in Table 7. The bias inflates differences, which makes it easier to find statistical significance. This can lead to claims about course effectiveness that may not be warranted. Other estimators used in this study were not so strongly biased toward pretest scores.

Third, the size of the class does not correlate with the normalized gain for class sizes smaller than 50, as illustrated by Figure 3 and by the low and nonsignificant Pearson *r* coefficient found for this distribution. Larger classes were not sufficiently sampled to draw any conclusion. Although this may indicate that class size does not influence student scores, we do not want to draw that conclusion because of the issues with the ADT as an instrument mentioned above. In addition, the sizes of lab and discussion sections were unknown. Part of the anonymity of a large lecture can be overcome in smaller, more personalized lab and discussion sections.

Fourth, prerequisites for mathematics show a positive correlation with the calculated normalized gain. This may be partly due to students entering such a class having higher pretest scores than students in classes with no mathematics prerequisite. We also suspect individual student demographics to be a factor because some mathematics prerequisites encourage students to take an astronomy course later in their academic careers. This may mean that students have developed more success skills for college courses. As such, they may have learned to get more out of a class, resulting in higher gains.

Fifth, the alignment of lecture and ADT items is positively correlated. The lecture is the most consistent factor in Table 7. This also is not surprising because it seems likely that lecture encompasses most of the time on task for the course, although the database does not provide direct evidence for this.

Last, it appears that the learning in all formats can be described best by a growth model of the form postscore = constant * prescore. This model accounts for over 50% of the variance in the cases in which a significant value (significantly deviating from zero) was found. The combination of low gains, the avoidance of the ceiling and floor effect regions, and a first-order relationship between pretest score and posttest score would argue for using a different measure for gain than the normalized gain. A case can be made for using Gain 3 because it is not biased toward pretest scores in this region, with this functional relationship between pretest and posttest scores. In general, choosing a measure for gain should depend on the relationship between pretest and posttest scores because different functional relations will bias different measures for gain in different ways.

6.1 Recommendations

To truly measure student understanding as a function of class format, more sensitive instruments will be needed. However, this is a double-edged sword. Concept inventories that focus on a single conceptual domain, like the ones on lunar phases (Lindell 2001), stars (Bailey 2006), greenhouse effect (Keller 2006), and light and spectra (Bardar 2006; Bardar et al. 2007), probe conceptual understanding of one particular topic and may be more sensitive to different instructional designs. For concept inventories to be successful, it is important that they are developed by people who are also experts in the discipline, as Hake (2007) argued. However, a word of caution is applicable in the use of concept inventories as measures for overall course effectiveness. Just as with the FCI in physics, a measurement of student understanding for a

single concept might not be representative of student overall understanding or course effectiveness. In a semester, many concepts are covered, and time spent on one of the topics that can be measured by one of the concept inventories listed above may be small. The alternative viewpoint is that if an instructor designs effective instruction for a particular conceptual domain, it is likely that students are receiving similarly effective instruction on other topics.

On a more logistical front, several recommendations can be made. If a large data-gathering project like the one by Hufnagel and Deming (1999) is undertaken again, some elements from that project could be improved to make the database product more useful to researchers. The first recommendation is to find avenues to develop and secure pretest and posttest data that are matched to individual student gains, yielding more powerful normalized gain scores. This would allow us to use repeated-measure statistics rather than independent-sample statistics, which would drastically reduce error terms. Moreover, Bao (2006) argued that using class averages rather than individual student scores can lead to different gain scores. This may require additional adjustments to determine how the attrition rate biases the data. An additional advantage would be that researchers can investigate individual classes rather than an aggregate of classes only. It would allow us to do a rigorous item analysis on the questions on the ADT. A second recommendation is to endeavor to obtain a more homogeneous determination of the mapping of the content on the ADT (rather than to rely on self-reports by the instructors) and to give an estimate of time spent on each of the mapping factors (the course elements). Although this would be difficult to do, it would allow researchers to make a more rigorous determination of which course elements influence gain scores most effectively.

Acknowledgments

EB would like to thank Joel Levin and Sanlyn Buxner for their advice, recommendations, and comments. The authors thank the anonymous referee for the thoughtful comments and suggestions.

References

Bailey, J. M. 2006, "Development of a Concept Inventory to Assess Students' Understanding and Reasoning Difficulties about the Properties and Formation of Stars," PhD dissertation, University of Arizona, Tucson.

Bao, L. 2006, "Theoretical Comparison of Average Normalized Gain Calculations," *American Journal of Physics*, 74(10), 917.

Bardar, E. M. 2006, "Development and Analysis of Spectroscopic Learning Tools and the Light and Spectroscopy Concept Inventory for Introductory College Astronomy," PhD dissertation, Boston University, Boston, MA.

Bardar, E. M., Prather, E. E., Brecher, K., & Slater, T. F. 2007, "Development and Validation of the Light and Spectroscopy Concept Inventory," *Astronomy Education Review*, 5(2), 103. http://aer.noao.edu/cgi-bin/article.pl?id=225.

Deci, E. L., & Ryan, R. M. 1985, *Intrinsic Motivation and Self-Determination in Human Behavior*, New York: Plenum Press.

Deming, G. 2002, "Results from the Astronomy Diagnostic Test National Project," Astronomy Education Review, 1(1), 52. http://aer.noao.edu/cgi-bin/article.pl?id=5.

Deming, G., & Hufnagel, B. 2001, "Who's Taking ASTRO 101? " *The Physics Teacher*, 39(6), 368.

Hake, R. R. 1998, "Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses," *American Journal of Physics*, 66(1), 64.

Hake, R. R. 2007, "Should We Measure Change? Yes!" To appear as a chapter in *Evaluation of Teaching and Student Learning in Higher Education*, American Evaluation Association monograph.

Halloun, I., & Hestenes, D. 1985, "Common Sense Concepts about Motion," *American Journal of Physics*, 53, 1056.

Hestenes, D., & Wells, M. 1992, "A Mechanics Baseline Test," *The Physics Teacher*, 30, 159.

Hestenes, D., Wells, M., & Swackhamer, G. 1992, "Force Concept Inventory," *The Physics Teacher*, 30, 141.

Hufnagel, B. 2002, "Development of the Astronomy Diagnostic Test," *Astronomy Education Review*, 1(1), 47. http://aer.noao.edu/cgi-bin/article.pl?id=4.

Hufnagel, B., & Deming, G. 1999, "The Astronomy Diagnostic Test: Comparing Your Class to Others," *BAAS*, 31(3), 937.

Keller, J. M. 2006, "Eliciting and Addressing Undergraduate Student Beliefs and Reasoning Difficulties Regarding the Atmospheric Greenhouse Effect," PhD dissertation, University of Arizona, Tucson.

Lindell, R. 2001, "Enhancing College Students' Understanding of Lunar Phases," PhD dissertation, University of Nebraska, Lincoln.

Rosenthal, R., & Rosnow, R. L. 1991, *Essentials of Behavioral Research: Methods and Data Analysis*, (2nd ed.), New York: McGraw Hill.

Thompson, B. 2007, "Effect Sizes, Confidence Intervals, and Confidence Intervals for Effect Sizes," *Psychology in the Schools*, 44(5), 423.

Zeilik, M. 2003, "Birth of the Astronomy Diagnostic Test: Prototest Evolution," *Astronomy Education Review*, 1(2), 46. http://aer.noao.edu/cgi-bin/article.pl?id=28.

APPENDIX A: Summary Statistics for the Classes Model

Click here for Appendix A in PDF.

APPENDIX B: Gain Behavior as a Function of Learning Model

Click here for Appendix B in PDF.

Appendix A: http://aer.noao.edu/auth/brogt-appendixa.pdf

Appendix B: http://aer.noao.edu/auth/brogt-appendixb.pdf

ÆR
25 - 42