5-1-2007

Interpreting force concept inventory scores: Normalized gain and SAT scores

Vincent P. Coletta
Loyola Marymount University, vcoletta@lmu.edu

Jeffrey A. Phillips
Loyola Marymount University, jphillips@lmu.edu

Jeffrey J. Steinert
Edward Little High School

Interpreting force concept inventory scores: Normalized gain and SAT scores

Vincent P. Coletta and Jeffrey A. Phillips
Loyola Marymount University, Los Angeles, California 90045, USA

Jeffrey J. Steinert
Edward Little High School, Auburn, Maine 04210, USA

(Received 22 May 2006; published 23 May 2007)

Preinstruction SAT scores and normalized gains (G) on the force concept inventory (FCI) were examined for individual students in interactive engagement (IE) courses in introductory mechanics at one high school ($N=335$) and one university ($N=292$), and strong, positive correlations were found for both populations ($r=0.57$ and $r=0.46$, respectively). These correlations are likely due to the importance of cognitive skills and abstract reasoning in learning physics. The larger correlation coefficient for the high school population may be a result of the much shorter time interval between taking the SAT and studying mechanics, because the SAT may provide a more current measure of abilities when high school students begin the study of mechanics than it does for college students, who begin mechanics years after the test is taken. In prior research a strong correlation between FCI G and scores on Lawson's Classroom Test of Scientific Reasoning for students from the same two schools was observed. Our results suggest that, when interpreting class average normalized FCI gains and comparing different classes, it is important to take into account the variation of students' cognitive skills, as measured either by the SAT or by Lawson's test. While Lawson's test is not commonly given to students in most introductory mechanics courses, SAT scores provide a readily available alternative means of taking account of students' reasoning abilities. Knowing the students' cognitive level before instruction also allows one to alter instruction or to use an intervention designed to improve students' cognitive level.

DOI: 10.1103/PhysRevSTPER.3.010106 PACS number(s): 01.40.Fk

I. INTRODUCTION

The force concept inventory (FCI) is a 30-question multiple-choice test,[1,2] used as a measure of student understanding of Newtonian concepts in introductory mechanics and usually given both at the beginning and at the end of an introductory mechanics course. The wrong answers on the test are based on extensive student interviews and correspond to common student misconceptions. Students usually score higher on the test when it is taken the second time, following instruction. Interpretation of FCI results is facilitated by use of the normalized gain[3,4] (G), defined as the change in score divided by the maximum possible increase:

$$G = \frac{(\text{postscore } \%) - (\text{prescore } \%)}{100 - (\text{prescore } \%)}.$$

For example, using this measure, we equate the conceptual gains of students with pre$\rightarrow$post scores of $20\% \rightarrow 60\%$, $40\% \rightarrow 70\%$, and $80\% \rightarrow 90\%$; all correspond to $G=0.5$. It should be emphasized that G is the single student normalized gain and is not the same as Hake's normalized gain $\langle g \rangle$, obtained from the class averages of pretest and posttest scores. Hake[4] discusses the mathematical relationship of $\langle g \rangle$ to the class average of individual students' G's and states that the two are usually within 5%.

One way to describe G is that it is a measure of the fraction of the concepts learned that were not already known at the beginning of the course. Thus we are able to use G as a measure of learning Newtonian concepts, independent of a student's initial state of understanding. The validity of this interpretation is justified by the fact that, *when other important factors such as reasoning ability are either accounted for or averaged over*, students' normalized gains are not correlated with preinstruction scores. For example, in a study of 12000 high school students' FCI scores, Hestenes[5] found that there was no significant correlation between G and FCI prescores (correlation coefficient $r=0.00$). However, in college introductory mechanics courses, G is often positively correlated with prescores.[6] We believe that this is not because higher prescores tend to *cause* higher G's, but rather because in college classes both high prescores and high G's tend to be achieved by those students with the strongest reasoning skills. Higher prescores are often a reflection of the greater conceptual learning achieved by stronger reasoners in their high school physics courses, and higher G's are achieved by stronger reasoners in their college courses. Thus the correlation between G and prescore in many college classes appears to be simply a by-product of a correlation between conceptual learning and reasoning skills, as discussed below. A more detailed explanation of the relationship between prescores, normalized gain, and reasoning ability may be found in our recent article.[6]

A considerable body of pedagogical research over the past decade has demonstrated that traditional physics instruction does not meet the needs of the great majority of students who take introductory physics courses. This research[3,7] also shows that many of the active learning, or interactive engagement (IE), strategies that have been developed in recent years are considerably more effective than traditional approaches. Traditional courses consistently result in class average G's of only about 0.2, whereas IE classes produce consistently higher class average G's, typically in the range 0.3–0.6.

We wondered whether this broad range of G's might be at least partly due to population effects. Our research has been

concerned with the effect of different student populations on values of G observed in IE classes. In two quite different populations we have seen very similar, strong positive correlations[8] between G and preinstruction scores on Lawson's Classroom Test of Scientific Reasoning.[9] In both groups, the upper quartile by Lawson score (averaging approximately 90%) achieved average G's over 0.6 and the lowest quartile by Lawson score (averaging approximately 45%) achieved average G's of less than 0.3. These results have now been replicated at the University of Colorado[10] and at the University of Central Florida.[11] We think it is quite likely that much of the variation in class average G's in different IE classes across the country may well be due to variations in the composition of classes with regard to reasoning level, and it is important that this be taken into account when interpreting gains. For example, it may be incorrect to conclude that teaching methods used in a class with a normalized gain of 0.6 are necessarily more effective than those which produce a gain of 0.3 in a different class, because the backgrounds of the students in the two classes may be a more important factor than the specific IE methods used in the classes. We are gratified that many physics instructors are beginning to use the Lawson test to help interpret their FCI results. But we are also aware that many other instructors find the addition of another diagnostic test too great a burden. The purpose of this paper is to offer an alternative to using valuable class time to administer the Lawson test, making use of SAT data that are already available in most student files.

Piaget's model of cognitive development states that an individual progresses through discrete stages, eventually developing the skills to perform scientific reasoning.[12] The penultimate stage is known as concrete operational. During this stage a person has the ability to make sense of concrete experiences but not yet form hypotheses or understand abstract concepts.[13] In the final stage, known as the formal operational stage, an individual has the ability to form an hypothesis and test it with carefully designed experiments, using hypothetico-deductive reasoning.[14] Although Piaget believed that the formal stage is typically reached between ages 11 and 15, many high school and college students never reach this stage.[15,16] For example, Arons and Karplus state that only 1/3 of college students have reached the formal stage.[17] The majority of students either remain confined to concrete thinking or are only capable of partial formal reasoning, often described as transitional. In other studies focusing on physics students, similar results have been seen.[18–20] It seems clear that while formal reasoning skills are not sufficient for a physics student, they are necessary. Students who lack the ability to understand abstract concepts will struggle even with Newton's second law.[21]

The SAT Reasoning Test, formally known as the Scholastic Aptitude Test, is a standardized test widely used in college admissions. The test is comprised of mathematical reasoning and verbal thinking sections, and although a writing section was recently added to the test, none of the data presented here are from this "new SAT." By focusing on general skills that will be used in college, rather than competence in specific subjects, the SAT strives to be a predictor of college success. According to the creators of the SAT, the test measures a student's "college readiness."[22] At least one study has interpreted this readiness as general intelligence g and observed significant correlations between measures of g and SAT scores.[23] Many studies have looked at the correlations between SAT scores and freshman grade point averages (FGPAs), the most often used measure of college success. While the reported correlation coefficients have varied, two large studies[24,25] have reported values around 0.35. Among engineers, there is a stronger correlation between the math section of the SAT and FGPA,[26] $r=0.43$. We have studied the correlation between cumulative math and verbal SAT scores and scientific reasoning ability, as measured by Lawson's test, for our own students, and found $r=0.746$ and $r=0.680$, respectively, for the university and high school students in our study. Since SAT scores correlate with Lawson scores and Lawson scores correlate with G, we decided to test for a correlation between SAT scores and G.

II. DATA

We analyzed preinstruction math, verbal, and cumulative SAT scores and FCI normalized gains for 292 students in various IE introductory mechanics classes at Loyola Marymount University (LMU) and for 335 students in IE modeling physics classes at Edward Little High School (ELHS). Of the 292 LMU students, 117 were taught by one of us (Coletta), using a method in which each chapter is covered first in a "concepts" class, in a Socratic style very similar to Peer Instruction, and then again in a "problems" class, featuring estimation problems and group problem solving. Another author (Phillips) taught 89 students in a learning cycle format, with lectures and small group activities, such as using conceptual worksheets, performing short experiments, and working context-rich problems. The other 86 LMU students were taught by professors Bulman and Sanny, who both lecture with a strong conceptual component and with frequent class dialogue. Half of the classes were calculus based, primarily composed of engineering majors; the other half were algebra based, with mostly biology and natural science majors.

All of the 335 ELHS students were taught by one of us (Steinert) in algebra-based regular or honors physics classes using modeling instruction. Modeling[27] engages students in constructing and using scientific models to understand the physical world by providing them with conceptual tools to represent physical objects and processes in multiple ways. Instruction is organized into modeling cycles,[28] which move students through the phases of model development, evaluation, and application in concrete situations, promoting an integrated understanding of a small set of models as the content core of physics. Students at ELHS collaborate in planning and conducting experiments and solving problems, and are required to justify their thinking in oral and written presentations of their laboratory conclusions and homework solutions. Socratic questioning techniques are used to probe for misconceptions and guide student inquiry.

The average SAT score of LMU students in the calculus-based course was 1192 ± 8 (s.e.), and the average of those in the algebra-based course was 1114 ± 12. Combined data from

TABLE I. The correlations of math (M), verbal (V), and combined math and verbal (M+V) SAT scores with FCI *G*.

	SAT M & FCI *G*	SAT V & FCI *G*	SAT M+V & FCI *G*
LMU	0.46	0.35	0.46
ELHS	0.57	0.45	0.56

both LMU courses provided a wide range of cumulative SAT scores: 720 to 1550, with an average of 1164±8. The cumulative SAT scores among the ELHS students ranged from 720 to 1540, with an average of 1109±9.

Typically students take the SAT exam during the spring of their junior year or fall of their senior year of high school. The ELHS students took introductory mechanics in the fall of their senior year of high school. (Note that although SAT scores for ELHS students achieved in their senior year were available, only scores earned prior to the start of their high school physics course were used in this study.) LMU students in calculus-based physics typically take introductory mechanics in the spring semester of their freshman year of college, and LMU students in algebra-based physics typically take introductory mechanics in the fall of their junior year of college. Thus all students in both schools took the SAT exam prior to the beginning of their mechanics course, but the time delay between testing and taking physics was much shorter for the high school students (typically less than 6 months) than for the university students (either almost 2 years or about 3 and 1/2 years).

We considered separately the correlations of math, verbal, and combined math and verbal scores with FCI *G* (Table I). We found highly significant correlations for all three at both schools, with significance levels $p<0.0001$; the probability that *G* and SAT scores are not correlated in these populations is less than 0.0001. The correlation coefficients for the math scores are considerably greater than the correlation coefficients for the verbal scores (0.46 vs 0.35 at LMU and 0.57 vs 0.45 at ELHS). The correlation coefficient for the combined math and verbal score is the same as for the math score alone at LMU, and the correlation for the combined score is nearly the same as for the math score at ELHS. Thus a student's SAT math score alone and her or his cumulative SAT score seem to be of equal value in predicting whether she or he will succeed in introductory physics.

For both schools, we graphed each student's normalized gain *G* versus the student's cumulative SAT math and verbal score [Figs. 1(a) and 2(a)]. There are, of course, other factors affecting an individual's value of *G*, and so there is a range of *G*'s for any particular SAT score. The effect of the SAT score on *G* can be seen more clearly by binning the data, averaging values of *G* over students with nearly the same SAT scores [Figs. 1(b) and 2(b)]. We formed bins with the same number of students in each bin, as nearly as possible, so that each data point on a graph of binned data has equal weight. Ideally, one wants the bins to contain as many students as possible, to produce a more meaningful average for each point on the graph. However, one also wants as many data points as possible to improve the statistics for the graph.

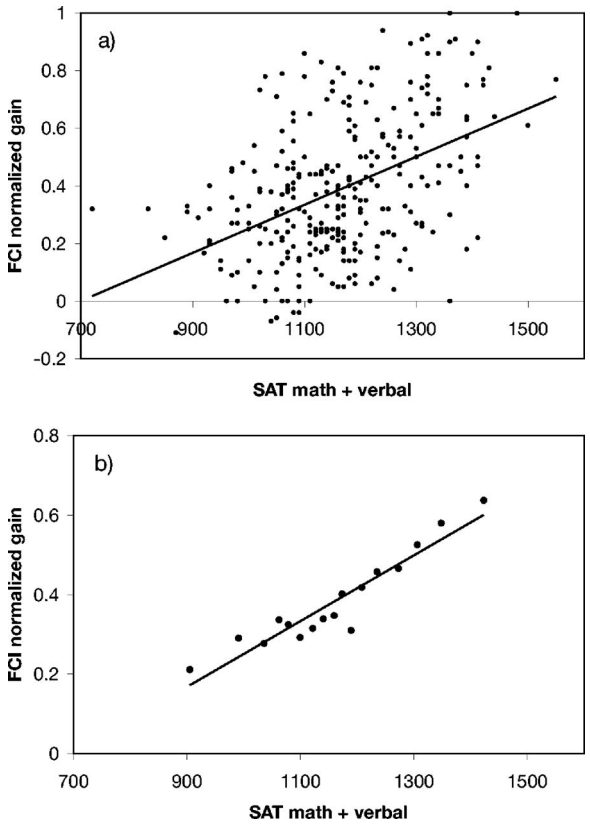

FIG. 1. (a) Plot of individual students' normalized FCI gains versus SAT scores for 292 LMU students: slope=0.00084, $r=0.46$. (b) Plot of normalized FCI gains versus SAT scores, with individual student data averaged within 17 bins.

Our approach is to make the number of bins roughly equal to the square root of the total number of students in the sample, so that the number of bins and the number of students in each bin are roughly equal. However, varying the bin size had very little effect on the slope of the best-fit line.

The slopes of the best-fit lines in Figs. 1(a) and 2(a) are 0.000 84 and 0.000 89, respectively, and the correlation coefficients *r* equal 0.46 and 0.57, respectively. Both the distribution of SAT scores and the regression lines were similar for the two data sets, and so we decided to combine data from the two schools. Figure 3 shows a graph of the combined, binned data from LMU & ELHS ($N=627$). A linear regression for this graph gives $r=0.94$. However, binning the data also reveals that the variation of *G* with SAT score is not linear. A quadratic function, shown in Fig. 3, provides a better fit to the data, with a correlation coefficient of 0.97. For purposes of comparison, we also combined the available Lawson and FCI data from both schools, again binned the data, and plotted a graph of FCI *G* versus Lawson score (Fig. 4, $N=297$). Again a quadratic equation provides a better fit to the data than a linear one: $r=0.89$, linear, and $r=0.95$, quadratic.

III. CONCLUSIONS

We conclude that, when one takes account of reasoning ability in interpreting FCI gains, use of SAT scores offers a

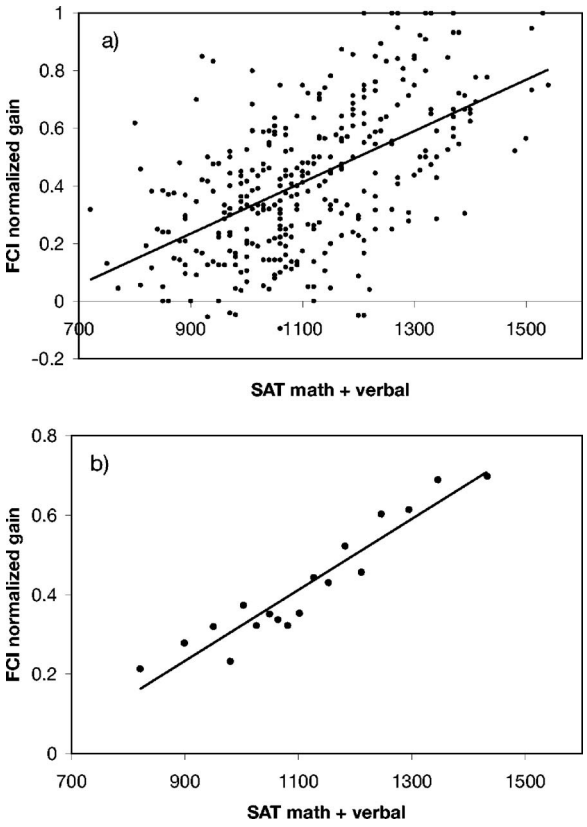

FIG. 2. (a) Plot of individual students' normalized FCI gains versus SAT scores for 335 ELHS students: slope=0.00089, r =0.57. (b) Plot of normalized FCI gains versus SAT scores, with individual student data averaged within 18 bins.

reasonable alternative to use of Lawson's test scores. We were able to obtain over twice as much SAT and FCI data as Lawson and FCI data. For the subset of LMU students for whom we have both Lawson and SAT scores (N=98), the correlation between SAT scores and FCI G's (r=0.46) is weaker than the correlation between Lawson scores and FCI G's (r=0.54). However, for the subset of ELHS students for whom we have both Lawson and SAT scores (N=199), the correlation between SAT scores and FCI G's (r=0.57) is

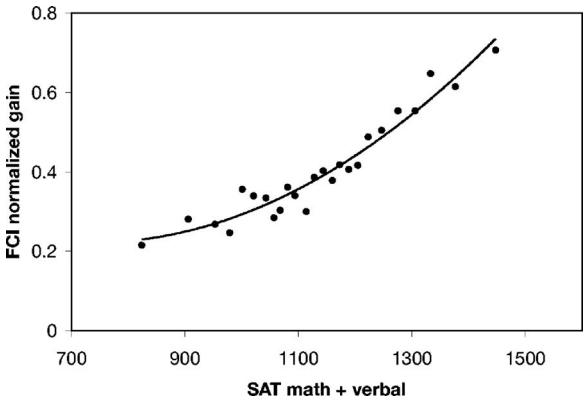

FIG. 3. Plot of normalized gains versus SAT scores for 627 LMU and ELHS students, with individual student data averaged within 25 bins.

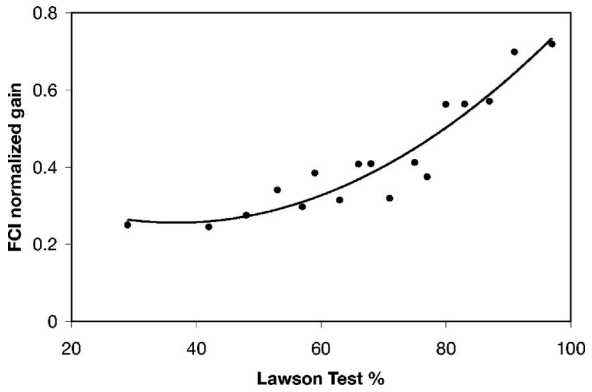

FIG. 4. Plot of normalized gains versus Lawson Test scores for 297 LMU and ELHS students, with individual student data averaged within 17 bins.

even stronger than the correlation between Lawson scores and FCI G's (r=0.53). Because the correlations between Lawson scores and FCI G's are so similar for the two schools, we conclude that the weaker correlation between SAT score and G observed at LMU is likely due to the greater time delay between taking the SAT exam and the beginning of introductory mechanics for college students. During that long delay one might expect that developmental experiences of students would vary, and so the SAT, taken up to 3 and 1/2 years earlier, would be a less accurate indicator of their initial state in a physics class.

SAT scores are used by colleges to predict college success, where success is measured by the FGPA, which is not a direct measure of student learning. With this study we see that the SAT is, in fact, correlated with student learning, as measured by the normalized gain on the FCI. The correlation we observed between SAT scores and FCI G's is larger than the correlation typically seen between SAT scores and FGPA.

Instructors may want to assess their class average FCI normalized gains by taking into account their students' reasoning ability either, by using class average Lawson test scores or class average SAT scores. Figures 3 and 4 provide a means to do so. For example, these figures show that for a class with an average SAT score of 1100 or an average Lawson score of 65%, a class average G of about 0.35 would be equal to the average G achieved by students in our study with the same average SAT or Lawson scores. For a class with either an average SAT score of 1400 or an average Lawson score of 95%, a class average G of 0.7 would be equal to the average G achieved by students in our study with the same average SAT or Lawson scores.

Several interventions have been developed to address cognitive development. (i) Feuerstein[29] developed an intervention for dramatically improving the reasoning of Israeli children with low IQ's. His methods have been applied by others and shown to be effective in improving the cognitive levels of normal children. (ii) In Great Britain, Adey[30,31] developed an intervention program for 12–14 year old children using science as a means of improving cognitive skills. He produced substantial long-term improvement in grades in science, mathematics, and English. (iii) In the U.S., Karplus[32,33] developed an intervention to improve the pro-

portional reasoning skills of middle school children. He demonstrated dramatic long-term improvement. Feuerstein, Adey, and Karplus were all strongly influenced by the work of Piaget. Our research demonstrates that many high school and college students, who have not attained the level of formal reasoning identified by Piaget as necessary for understanding science, could benefit from such interventions. We are currently working to adapt some of the interventions mentioned above for our students.

ACKNOWLEDGMENTS

We wish to thank John Bulman and Jeff Sanny for sharing their data.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[2] E. Mazur, *Peer Instruction: A User's Manual* (Prentice Hall, Upper Saddle River, NJ, 1997). We used Mazur's version of FCI.
[3] R. R. Hake, Interactive-Engagement vs Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses, Am. J. Phys. **66**, 64 (1998).
[4] R. R. Hake (unpublished), http://www.physics.indiana.edu/~hake/PERC2002h-Hake.pdf.
[5] D. Hestenes (private communication).
[6] V. P. Coletta and J. A. Phillips, Interpreting FCI Scores: Normalized Gain, Pre-instruction Scores, and Scientific Reasoning Ability, Am. J. Phys. **73**, 1172 (2005).
[7] E. Redish and R. N. Steinberg, Teaching physics: Figuring out what works, Phys. Today **52** (1), 24 (1999).
[8] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Why you should measure your students reasoning ability, Phys. Teach. **45**, 235 (2007).
[9] A. E. Lawson, The development and validation of a classroom test of formal reasoning, J. Res. Sci. Teach. **15**, 11 (1978). An updated multiple choice version of the test is in the appendix of Ref. 6.
[10] M. A. Dubson and S. J. Pollock, Can the Lawson Test Predict Student Grades?, AAPT Announcer **36**, 90 (2006).
[11] P. M. Pamela and J. M. Saul, Interpreting FCI Normalized Gain, Pre-instruction Scores, and Scientific Reasoning Ability, AAPT Announcer **36**, 89 (2006).
[12] J. W. Renner and A. E. Lawson, Piagetian theory and instruction in physics, Phys. Teach. **11**, 165 (1973).
[13] B. Inhelder and J. Piaget, *The Growth Of Logical Thinking From Childhood To Adolescence; An Essay On The Construction Of Formal Operational Structures* (Basic Books, New York, 1958).
[14] A. E. Lawson, The generality of hypothetico-deductive reasoning: Making scientific thinking explicit, Am. Biol. Teach. **62**, 482 (2000).
[15] D. Elkind, Quality conceptions in college students, J. Social Psych. **57**, 459 (1962).
[16] J. A. Towler and G. Wheatley, Conservation concepts in college students, J. Genet. Psychol. **118**, 265 (1971).
[17] A. B. Arons and R. Karplus, Implications of accumulating data on levels of intellectual development, Am. J. Phys. **44**, 396 (1976).
[18] H. D. Cohen, D. F. Hillman, and R. M. Agne, Cognitive level and college physics achievement, Am. J. Phys. **46**, 1026 (1978).
[19] J. W. McKinnon and J. W. Renner, Are colleges concerned with intellectual development?, Am. J. Phys. **39**, 1047 (1971).
[20] A. E. Lawson and J. W. Renner, A quantitative analysis of responses to Piagetian tasks and its implications for curriculum, Sci. Educ. **58**, 545 (1974).
[21] J. W. Renner and A. E. Lawson, Promoting intellectual development through science teaching, Phys. Teach. **11**, 273 (1973).
[22] The College Board, *SAT Program Handbook, 2005* http://www.collegeboard.com/prod_downloads/prof/counselors/tests/sat/2005-06-SAT-program-handbook.pdf.
[23] M. C. Frey and D. K. Detterman, The Relationship Between the Scholastic Assessment Test and General Cognitive Ability, Psychol. Sci. **15**, 373 (2004).
[24] B. Bridgeman, L. McCamley-Jenkins, and N. Ervin, *Predictions of Freshman Grade-Point Average From the Revised and Recentered SAT I: Reasoning Test* (College Entrance Examination Board, New York, 2000) http://www.collegeboard.com/research/pdf/rr0001_3917.pdf.
[25] S. Geiser with R. Studley, *UC and the SAT: Predictive Validity and Differential Impact of the SAT I and SAT II at the University of California*, 2001 http://www.ucop.edu/sas/research/researchandplanning/pdf/sat_study.pdf.
[26] J. S. Shoemaker (unpublished).
[27] M. Wells, D. Hestenes, and G. Swackhamer, A modeling method for high school physics instruction, Am. J. Phys. **63**, 606 (1995).
[28] R. Karplus, Science teaching and the development of reasoning, J. Res. Sci. Teach. **14**, 169 (1977).
[29] R. Feuerstein, Y. Rand, M. B. Hoffman, and R. Miller, *Instrumental enrichment: An intervention program for cognitive modifiability* (University Park Press, Baltimore, 1980).
[30] P. S. Adey and M. Shayer, *Really Raising Standards: Cognitive intervention and academic achievement* (Routledge, London, 1994).
[31] P. S. Adey, M. Shayer, and C. Yates, *Thinking Science: The curriculum materials of the CASE project*, 3rd ed. (Nelson Thornes, London, 2001).
[32] B. Kurtz, Ph.D. dissertation in science and mathematics education, University of California, 1976.
[33] B. Kurtz and R. Karplus, Intellectual development beyond elementary school vii: teaching for proportional reasoning, Sch. Sci. Math. **79**, 387 (1979).