Educational research

Predicting at-risk students in general chemistry: comparing formal thought to a general achievement measure

Scott E. Lewis and Jennifer E. Lewis*,

Department of Chemistry, University of South Florida, FL, 33620, USA
e-mail: jlewis@cas.usf.edu

Received 10 August 2006, accepted 1 January 2007

Abstract: This study is an investigation into the ability of pre-assessment measures of formal thought ability and general achievement to predict students at-risk of poor performance in college-level general chemistry. Over a three year period, data on formal thought ability (as measured by the Test of Logical Thinking, or TOLT) and/or general achievement (as measured by the Scholastic Aptitude Test, or SAT) was collected from over 3000 students as they entered a general chemistry course. The outcome measure was an American Chemical Society general chemistry exam at the end of the course. Findings indicate that both the formal thought and the general achievement measure can successfully identify at-risk students in this setting, with neither measure being superior in doing so. The presence of distinct groups of students correctly predicted to be at-risk by only one of the measures demonstrates that formal thought ability and general achievement each represent an independent hindrance to success in chemistry. Therefore, efforts to help at-risk students should include a focus on the development of formal thought as well as a content review. [*Chem. Educ. Res. Pract.*, 2007, **8** (1), 32-51]

Keywords: Assessment, formal thought, at-risk students, performance predictors, college chemistry

Introduction

All too often, substantial numbers of students in college fail to demonstrate sufficient understanding of chemistry to proceed beyond the introductory course, general chemistry. This circumstance hinders not only the individual student but also the field of chemistry. While the costs to the individual are immediate and obvious (not only the regrettable lack of knowledge of chemistry but also a closed door to any major field of study requiring that knowledge), the costs to chemistry are also significant. With each year this trend continues, chemistry loses numerous individuals who now will not contribute to the growth of the discipline. Indeed the ramifications stretch beyond chemistry, as other science curricula require general chemistry prior to course work within their program (Tai et al., 2005). Students who cannot muster an acceptable understanding of general chemistry are prevented from contributing to many science fields. On a more systemic level, the inability of students to continue in science-oriented courses because of low performance in general chemistry represents a major setback in efforts to create a scientifically-informed populace and a technically-proficient workforce. For these reasons, unsatisfactory student performance in college-level general chemistry remains a critical area of concern.

Since basic constructivism indicates that the prior knowledge and skills with which students enter a course play a role in success (or its absence), it is both possible and valuable to identify students who are at-risk of not succeeding in a course at the point when they first enter the course. To do so provides the opportunity for assisting these students early on,

while success is still possible. Further, knowledge about the factors contributing to low (at-risk) performance can inform the design of interventions aimed toward reducing the challenges faced by these students. The first task is identifying the at-risk population with reasonable accuracy, and the second is suggesting potential interventions. Ideally, the measure used for identification contains within itself implications for a potential remedy. This paper compares the accuracy, degree of overlap, and implications for potential interventions of two measures that can be used to identify students at-risk of not succeeding in general chemistry. It therefore joins a long history of 'predictor papers' but is unique in its combination of generalizability, a focus on at-risk students, and consideration for the implications of choosing a particular predictor.

The need for more work with predictors

Extensive work has been done on the ability to predict success in college chemistry. Past studies of college chemistry have examined the ability of SAT (Pederson, 1975; Pickering, 1975; Bender and Milakofsky, 1982; Craney and Armstrong, 1985; Nordstrom, 1990; Bunce and Hutchinson, 1993; Spencer, 1996), ACT (Carmichael et al., 1986; Nordstrom, 1990 House, 1995), high school GPA (Carmichael et al., 1986), high school chemistry grade (Ozsogomonyan and Loftus, 1979; Craney and Armstrong, 1985; Nordstrom, 1990), personality characteristics (House, 1995) and Piagetian tasks (Bender and Milakofsky, 1982; Bunce and Hutchinson, 1993) to predict final chemistry course grade. In all these studies, however, the use of chemistry grade as an outcome variable relies on the ability of chemistry grade to approximate chemistry understanding. The extent to which this approximation is valid depends on several decisions peculiar to the course, the instructor and the institution. Decisions such as grading on a curve or an absolute scale, grading based completely on exam performance versus consideration of student homework, the allowance of extra credit, and even the method by which each exam was created, can all alter the extent to which chemistry grades reflect true student understanding of chemistry. As a result, the generalizability of the above studies depends on whether all of these factors are handled the same way at other institutions. Of the studies presented above, the work by Bender is the only one to provide detailed evidence of the grading procedures employed so that a replication could be attempted at another institution.

A more replicable option for an outcome variable is the use of a single exam as a measure of students' chemistry understanding. The exam questions can readily be made available for scrutiny in order to provide a clear picture of what constitutes success in chemistry, with none of the ambiguity surrounding course grade. In addition, the scoring of a single exam lends itself readily to the statistical procedures commonly used with predictors. One example of such a procedure is present in Yager et al.'s examination of the effects of taking high school chemistry (Yager et al., 1988). In this study, students were measured on a standard exam, a course final exam, and by a final course grade to provide multiple measures of success in chemistry. In particular, the use of a standard exam allows for a ready assessment of generalizability.

Finally, considerable work has gone into the development of chemistry-based diagnostic exams for course placement and prediction of performance (Russell, 1994; McFate and Olmsted III, 1999). These instruments tend to incorporate both math and chemistry questions and can be said to measure chemistry ability rather than incoming chemistry-specific knowledge. Such instruments seem to have a reasonably high success rate in predicting chemistry grades (McFate and Olmsted III, 1999; Legg et al., 2001; Wagner et al., 2002), but leave open the question of what can be done to assist the students who score low on such measures. Some suggestions put forth include recommending increased study time or remedial coursework for such students. Similar suggestions are presented in Yager's study**.**

But these suggestions do not necessarily lead to specific remedies: for example, what should be the design and intent of remedial coursework? How should the increased study time be spent? Even in the case of chemistry-focused exams that can identify specific deficiencies indicating that students did not achieve a sufficient understanding of chemistry via their earlier chemistry courses, how should we construct a second attempt to teach these concepts so that it will be successful? These questions are of particular importance, since recent research has suggested that remedial coursework may offer only marginal improvements in chemistry success (Bentley and Gellene, 2005; Jones and Gellene, 2005).

Our dissatisfaction with the remedies that can be offered on the basis of chemistry diagnostic exams or high school chemistry GPAs led us to the most important facet of our study. In particular, we wanted to compare two potential methods of identifying at-risk students that would suggest slightly different remedies in order to consider whether either, neither, or both remedies are tenable. Our study therefore compares two predictors, one with a long history of success at predicting course grades (SAT score), and another with a sound theoretical underpinning (formal thought ability) but with less information available as to its efficacy as a predictor in a college chemistry setting. As will be discussed in the next section, formal thought ability has a theoretical link to specific chemistry topics, and a research base aimed at improving formal thought performance (Lawson and Nordland, 1976; Adey and Shayer, 1990; Shayer and Adey, 1992a, 1992b, 1993) means that interventions to improve formal thought could be readily applied. In a similar fashion, the role of Math SAT in predicting performance implies that math skills are responsible for success, which would lead to specific suggestions of additional math course work or tutorials.

Although our study joins a long history of predictor papers, no one has yet offered a replicable predictor study that contains within itself clear guidelines for the construction of remedies. Further, our focus is predicting students at-risk of performing poorly in general chemistry, since it is for these students that interventions are needed. This means we look specifically at how well the predictors in our study identify students who perform at the lower end of our outcome measure, something which few previous studies have done. (Notable exceptions are Legg (Legg et al. 2001) and Wagner (Wagner et al., 2002)). For our outcome measure, as a result of considering the limitations of previous studies based on course grades (discussed above), we chose a standard exam designed to measure student understanding of chemistry. Our results are therefore generalizable to the extent that the content of this exam matches the desired outcomes at other institutions. The exam is available to the public, so this determination can be made (Examinations Institute of the American Chemical Society, 1997). Further, our study allows us to see whether, in the specific case of college chemistry performance, a simple paper and pencil measure of formal thought ability, the Test of Logical Thinking (TOLT), can stand up against the SAT's successful history.

Formal thought and science achievement

With the intent of identifying at-risk students in a way that would inherently suggest a particular remedy, our predictor selections had to be focused on measures that have the potential to describe a large hindrance for students. Because of its basis in a well-described learning theory, the construct of formal thought offers the ability to suggest specific difficulties students face, leading to specific remedies. Formal thought has been described as one of a series of factors necessary for a successful performance (Lawson, 1979, 1983; Chandran et al., 1987), so the *absence* of formal thought would definitely be expected to lead to a poor performance – exactly what is necessary for a good predictor of at-risk status.

Formal operational thought is the last stage of cognitive development as described by Piaget, in which 'deduction no longer refers directly to perceived reality but to hypothetical statements' (Inhelder and Piaget, 1958). In formal thought, possibilities are regarded as

hypothetical at first, and then verified by empirical evidence: in short, deductive reasoning. Contextually, this leads to the meaningful manipulation of empirical results, as well as a familiarity with the abstract. Also taken from Piaget's work is a series of reasoning patterns that would describe formal thought operations. Adey and Shayer (1994) grouped the reasoning patterns into three main categories. The first category, the handling of variables, includes the control and exclusion of variables, the recognition of multiple classification schemes, and the description of combinatorial possibilities. The second category, relationships between variables, includes the use of ratios, and proportion (comparing of two ratios), as well as compensation (use of inverse relationships), correlation and probability. The final group, formal models, describes the creation of an abstract representation of complex behaviours. Also included in this last group is the use of logical reasoning. Within Piagetian theory, the onset of formal thought would be characterized by the development of all the cognitive operations at about the same time, a postulate that has been supported by empirical evidence (Lawson and Renner, 1975; Lawson and Nordland, 1976; Lawson et al., 1978).

Certain aspects of formal thought have been suggested as explaining the difficulty some students face in chemistry. The second category, relationships between variables, for example could explain an inability to relate mathematical formulas to underlying concepts, a task frequently required in chemistry. In keeping with this idea, some researchers have hypothesized which chemistry concepts require formal thought (Herron, 1975), and others have investigated links between formal reasoning ability and conceptual understanding of specific topics in chemistry (Abraham and Williamson, 1994; Demerouti et al., 2004). Neo-Piagetian theories of learning still incorporate formal thought ability as one of several critical cognitive factors important for problem-solving in chemistry (Niaz, 1987, 1996; Tsaparlis 2005). Tsaparlis et al. investigated the effects of several cognitive variables on student performance on several types of molecular equilibrium problems and found that developmental level in terms of formal thought ability was the most important predictor of success; however, additional work led to the identification of developmental level as a potentially confounding factor in studies using chemistry problems with complex logical structures to investigate the importance of working memory capacity (Tsaparlis et al., 1998; Tsaparlis and Angelopoulos, 2000).

Formal thought has been postulated as a necessary condition, either directly or indirectly, for conceptual change to occur (Oliva, 2003). Thus, in addition to describing students' incoming abilities, formal thought may play a role in whether and how students actively incorporate new information presented in the course. One early example is the work of Lawson and Renner (1975), who showed that students at the concrete operational stage are unable to develop an understanding of formal concepts, and that students at the formal operational stage demonstrate an understanding of both formal and concrete concepts. Lawson (1982, 1985) pointed out that such results could be interpreted largely as a spurious correlation, describing what might be a more general intelligence measure underlying the success seen on both measures. In the 1982 study, a partial correlation between formal thought and biology achievement while controlling for fluid intelligence revealed a significant relation, illustrating that it was the formal thought measure that better corresponded to this biology achievement measure.

We continue this line of investigation by examining whether formal thought features a unique relationship to overall achievement in college chemistry. While Lawson demonstrated that controlling for a general intelligence measure did not remove the relationship between formal thought and biology achievement, no one has investigated whether a general achievement measure, such as SAT, may be at the heart of that relationship. The potential overlap between general achievement and formal thought has important classroom

implications for assisting at-risk students. If formal thought and general achievement have a high degree of commonality in relating to course performance, then formal thought maps onto a broader range of general abilities and any potential remedies should consider this. For example, efforts to promote formal thought alone may have limited utility, as other factors such as math ability would still hamper success. However, if students with low formal thought are hindered in the course regardless of scores on the general achievement measure, then formal thought still represents a series of specific traits that are independent of more general measures. In this case, interventions targeted solely toward the development of formal thought would have significant potential to assist at-risk students, whereas those that focus solely on developing math skills (such as algebraic manipulation) would not.

Although the contrast between a general achievement measure and a formal thought measure provides theoretical interest, the primary goal of this study is to produce a generalizable model for identifying at-risk students that will be useful for recommending specific interventions leading to success in general chemistry. It is in this frame that we have discussed the two potential predictors and the implications arising from their comparison. Our goal therefore leads to a series of research questions:

- Which predictor, SAT or a formal thought measure, is better able to identify at-risk students?
- Are the at-risk students identified by each predictor distinct groups, which may lead to more specific interventions geared for each group of students?
- Can a combination of SAT and formal thought measures provide an advantage in identifying at-risk students?
- And, to what extent are all at-risk students identified by this set of predictors?

Methods

Instruments: predictor and outcome variables

SAT

The SAT is a college entrance exam common in the U.S., typically administered in a student's final year of high school (Educational Testing Service, 2006). When the SAT data were obtained, the mathematics portion of this multiple-choice exam covered basic topics in mathematics, including algebra, geometry, data analysis, and probability and statistics, while the verbal portion involved reading comprehension and vocabulary skills. SAT sub-scores were obtained from the university's registrar as they were reported from the Educational Testing Service. SAT sub-scores have been found to have reliability coefficients exceeding 0.9 and a large body of research has demonstrated predictive validity towards college grades, convergent validity with other predictors used in admissions, and construct validity by panel reviews and item analysis (Cohen and Cronbach, 1985).

Test of Logical Thinking (TOLT)

Several measures of formal thought have been developed, validated and utilized in the research literature. What these measures share is an attempt to approximate the original Piagetian interviews. Emulating a Piagetian interview is problematic, especially with large numbers of students, due to the time-intensive nature of the interview procedure. As a result, written exams, in particular, have been constructed to take the place of these interviews. Perhaps the closest approximation to the interview procedure is Shayer and Adey's Science Reasoning Tasks (Shayer and Adey, 1981), in which students are asked to make written predictions before they witness demonstrations and then are asked to explain what they saw in each case. Depending on the task, questions may be free-response or require students to select from a set of responses.

However, for the present study, with class sizes approaching 200 students, we were concerned about the timing for student responses and doubtful that all students would be able to witness a demonstration adequately. As a result, we elected to choose a completely written exam. Among the possibilities are the Inventory of Piagetian Developmental Tasks, IPDT (Bender and Milakofsky, 1982); the Group Assessment of Logical Thinking, GALT (Roadrangka et al, 1983); the Test of Logical Thinking, TOLT (Tobin and Capie, 1981); and the Piagetian Logical Operations Test, PLOT (Staver and Gabel, 1979). Of these choices, the TOLT was selected because of its ease of administration (normally taking 40 minutes of class time); two-tiered question design, which reduces the possibility of students' guessing the correct answer (Treagust, 1988); published validity (Tobin and Capie, 1981), and use in the research literature (Haidar and Abraham, 1991; Yarroch, 1991; Williamson and Abraham, 1995; BouJaoude et al., 2004). Additionally, a Spanish language TOLT has been developed and validated, making the instrument available to a larger audience (Acevedo and Oliva, 1995). Because of the ease of administration and bilingual availability, the TOLT may be seen as a preferential predictor to the SAT from an international perspective. For this case and others in which SAT scores for entering students are not widely available, the TOLT can be given in less than one class period.

The TOLT was developed and validated by Tobin and Capie to measure what they termed formal reasoning ability. In order to do so items previously used by Lawson (Lawson, 1978; Lawson et al., 1979) were selected so that the test comprised two items for each of five modes of formal reasoning: controlling variables, proportional reasoning, probabilistic reasoning, correlational reasoning and combinatorial reasoning. To receive a correct score for each item, students need to select the correct answer from up to 5 choices and select the correct reason for the answer from 5 possible reasons. The only exceptions are the combinatorial reasoning questions, where students are required to list all the correct combinatorial possibilities without any replication. The validation of TOLT was done by relating student scores on the TOLT with student performance via interviews, for students ranging from grade six to college (Tobin and Capie, 1981).

ACS 'Special' Exam

As noted, a large variety of predictor papers rely on student grades as an outcome variable. As a research base, the results of these studies are generalizable only to the extent one can assume that student grades at the research institution match the desired student outcomes of other locales. More importantly, without extensive detail, this assumption becomes impossible to assess. With the desire to produce a generalizable model for identifying at-risk students, we selected an exam produced by the American Chemical Society (ACS), whose Division of Chemical Education features an Examinations Institute, which provides exams to chemistry teachers and administrators in high schools, colleges and universities. This exam is copyrighted and kept secure, so it can be given to candidates year after year, making it easy to make valid comparisons of student scores from different years and institutions.

The Examinations Institute offers more than fifty exams covering general chemistry, organic chemistry, analytical chemistry, physical chemistry, inorganic chemistry, biochemistry, polymer chemistry, and high school chemistry (American Chemical Society, 2006). The first semester general chemistry exams include various lengths of a conventional exam and a special examination (SP97A) meant to combine conceptual knowledge questions with the conventional (algorithmic) type questions. Given the recent push towards conceptual understanding of chemistry in the research literature (Pickering, 1990; Sawrey, 1990; Nakhleh, 1993; Nakhleh et al., 1996) our view is that both conceptual and conventional assessment methods play an important role in the objectives of most general chemistry

courses. As a result we selected the ACS special examination as the outcome variable for this study. Though this exam played a large role in determining student grades (it served as the final exam for the course, 25% of student grades), there were other factors that also contributed to student grades. Thus, it would be possible, though unlikely, for a student to complete the course successfully despite a poor performance on the exam.

Participants

The TOLT was administered during the first week of classes in 22 classes of the first semester of general chemistry at a large southeastern public urban research university over the course of three academic years. Students were given 45 minutes to complete the TOLT. Taking the TOLT comprised a small portion of the students' grades, and students were not graded based on their performance on the TOLT. These administration procedures resulted in TOLT scores for 3798 students out of an estimated 4180 students enrolled in the 22 classes. Of the 3798 students, 56.0% of the students were in their first year in college, 62.1% were female and 74.1% reported having at least one full year of high school chemistry.[a] A majority of the students described their major or intended major as pre-med or allied health professions. Finally, when asked about the grade they expected to receive in the course, 97.2% responded with either an A or a B. Only one student reported anticipating failing the course.

At the end of the course, students took the ACS exam as a final exam to measure student academic achievement. Of the 3798 students, ACS exam scores were available for 2871 students (75.6%). Since completing the ACS exam was a course requirement, the likely reason for not obtaining ACS exam scores was students not completing the course. Finally, student SAT scores were obtained from institutional records. Among the 2871 students that took the ACS exam, SAT scores were available for 2284 students. The most likely causes for missing SAT scores were students taking the ACT in place of the SAT or students enrolling in the course after SAT records were pulled. The focus of the analysis was the 2284 students for whom complete data was available. The decision to omit missing data will be revisited in a later section, particularly since the missing data may disproportionately represent at-risk students.

Examining the data

Among the complete data, steps were taken to determine if there were any outliers in the data, so that no single data point would have an unusually large effect on the results of the analysis. Outliers were determined by evaluation of the standardized residuals for a multiple regression model that included both SAT sub-scores and TOLT. An examination for any standardized residuals greater than 3 (Stevens, 1999), revealed nine students found to be inconsistent with the general pattern, and these were omitted from future analysis, resulting in 2275 students.

Prior to examining the trends between variables, descriptive statistics were evaluated and are presented in Table 1.

Table 1. Descriptive statistics for measures used.

	TOLT (0-10)	Math SAT (200-800)	Verbal SAT (200-800)	ACS Exam (0-100)
Mean	6.80	559.14	540.58	52.02
St. Dev.	2.613	83.505	82.648	16.638
Skewness (Std. error = 0.048)	-0.664	-0.048	0.070	0.240
Kurtosis (Std. error = 0.097)	-0.452	-0.166	-0.115	-0.690

The normality tests indicate that the TOLT scores feature a significant negative skew, indicating the scores were more heavily distributed at the higher values. This may be a result of the setting of the study, since the TOLT was designed for grades 6 through college, while the sample consists entirely of college students. While most statistical tests rely on a normality assumption, the tests employed are also very robust to violations of normality (Cohen et al., 2003).

Analysis procedures

As described previously, several research questions guided the nature of this investigation. To investigate each question, inferential and descriptive statistics were used. Inferential statistics established the utility of the predictors by relating the predictors to performance and assisted in interpretation of the descriptive statistics. Where possible, effect sizes were reported as a standardized measure of the differences seen, and operationalized using Cohen's qualitative terms: small, medium and large effects. As Cohen describes them, small effects are where the effect is small relative to the effect of uncontrollable extraneous variables (noise), medium effects are thought to be large enough to be visible to the naked eye and large effects are described as clearly visible (Cohen et al., 2003). Descriptive and inferential statistics were used to relate the ability of the models to identify at-risk students.

The first step in identifying at-risk students is to classify what would constitute an at-risk student. Typically, a grade of 'C' is meant to denote an average performance, and students whose scores fall substantially below an average performance can be considered at-risk. For this study, substantially below average was considered to be scoring below the 30th percentile on the ACS exam. For this sample, students who scored below 43.3% correct on the ACS exam (i.e. more than 0.525 standard deviations below the mean), represent the at-risk group. 802 out of the 2275 students (35.3%) in the sample scored below this cut-off. Since it would have been conceivable to make the decision regarding the at-risk cut-off differently, the effects of choosing different cut-offs are discussed in a later portion of this paper.

Results

Which predictor, SAT or a formal thought measure, is better able to identify at-risk students?

First the extent to which TOLT and SAT sub-scores have a linear relationship with academic performance was determined via correlation analysis. The results from the relevant correlations are presented in Table 2.

Table 2. Comparison of correlation coefficients.

	TOLT	VSAT	MSAT	ACS Exam
TOLT	---			
VSAT	0.492	---		
MSAT	0.654	0.625	---	
ACS Exam	0.510	0.527	0.608	---

all coefficients $p < 0.001$

The presence of significant positive correlation coefficients is indicative of a relationship with academic performance among all the predictors. Correlation coefficients also provide an indication of the strength of relationship between the predictors and the outcome variables. Using Cohen's effect size operation, each of the predictors features a large effect size with the outcome variable, and a large effect size between each predictor. Thus each predictor is

believed to be a reasonable construct in explaining ACS exam score, and consideration will need to be given to the possibility of the predictors over-lapping. This is consistent with our goal to determine whether formal thought and general achievement can be thought of as distinctly different in terms of their bearing on college chemistry achievement.

In order to determine the ability of the predictors to identify at-risk students, two linear regression models were used. The first model relates TOLT to students' scores on the ACS exam, and the second relates the SAT sub-scores to the ACS exam. The combination of both SAT sub-scores in one model was chosen to represent the practical option for data available to instructors. The results from the two regression models are shown in Tables 3 and 4 below:

Table 3. TOLT model results.

Coefficient	Slope	Std. Error	t-value	p
Constant	29.936	0.837	35.748	<0.001
TOLT	3.245	0.115	28.249	<0.001

$R^2 = 0.260$ Model F = 798.006

Table 4. SAT model results.

Coefficient	Slope	Std. Error	t-value	p
Constant	-25.196	1.995	-12.630	<0.001
SATV	0.04864	0.004173	11.657	<0.001
SATM	0.09107	0.004130	22.050	<0.001

$R^2 = 0.405$ Model F = 774.514

Using each model, it is possible to depict which students would be identified as at-risk by each set of predictors. For the TOLT model, TOLT scores of 4 or less are predicted to be below the cut-off. By this criterion, the TOLT model identifies 471 students in the sample to be at-risk. Of those 471 students, 332 students had an actual ACS exam score below the cut-off, indicating 70.5% of those predicted were correctly classified (see Table 5). Of the 139 incorrectly classified, 75 scored below average on the ACS exam.

The SAT model predicts students to be below the cut-off via a variety of SAT score combinations, so it is not appropriate to name a single set of SAT cut-offs. However, as rule of thumb, scores below 500 on both the math and verbal portion would qualify as at-risk in this context, although a score below 500 on one portion could potentially be offset by a higher score on the other. Application of the SAT model led to a classification of 451 students as at-risk based on the combination of SAT sub-scores, slightly lower than the number of students the TOLT model predicted. Of the 451 students, 327 were correctly classified, a 72.5% success rate, a rate slightly higher than the TOLT model (see Table 5). Of the 124 incorrectly classified in this group, 69 scored below average.

Table 5. The model predictions: at-risk students.

	Predicted At-risk	Actually At-risk	% correct predictions
TOLT model	471	332	70.5%
SAT model	451	327	72.5%

The similar success rates in identifying at-risk students is curious, given the lower R^2 of the TOLT model compared to the SAT model. As a measure of goodness of fit, it is expected that the higher the R^2 value, the better the model would be at predicting scores. This expectation would hold true if predictions for the entire sample were considered. However, in looking only at the at-risk students, a subset of the sample is being examined. Why the TOLT

model is able to identify at-risk students better than expected based on its R^2 value may be understood by considering the following scatter-plots, in which the relationship between ACS exam score and TOLT can be compared with the relationship between ACS exam score and Math SAT. Because of the large number of data points, a 20% random sample of the data is used in these plots.

Figure 1. Low TOLT identifies at-risk.

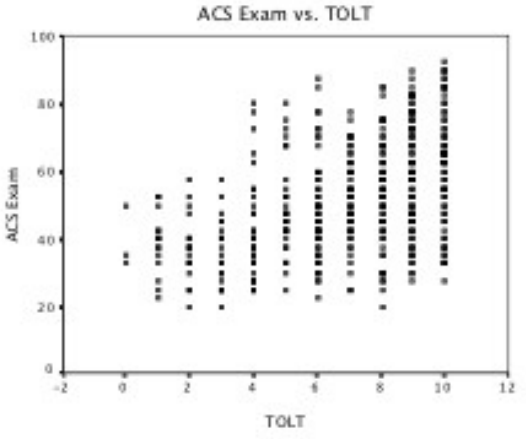
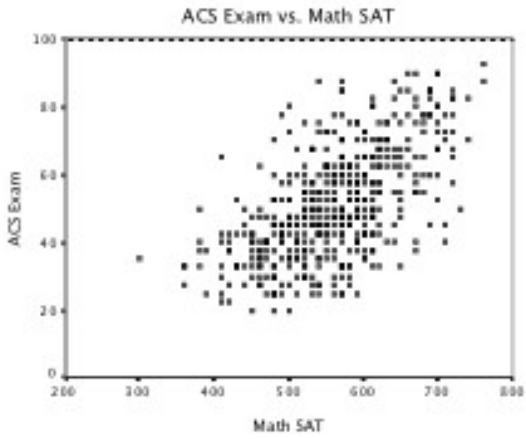


Figure 2. SAT predicts entire range.



The TOLT plot (Figure 1) demonstrates that the variability of ACS exam scores is low for the low TOLT scores, while the ACS exam scores span almost the entire range for the high TOLT scores. This broad distribution at the high end is the likely cause for the lower R^2 for the TOLT model as compared to the SAT model. In the Math SAT plot (Figure 2), a more linear trend is observed: high Math SAT scores correspond to higher ACS exam scores, while lower Math SAT scores correspond to lower ACS exam scores, which would lead to a higher R^2. (Verbal SAT has a similar distribution, as does a combination of the two SAT sub-scores using the weighting found in the regression model; only one of these three possible plots is shown for simplicity). For this reason, SAT would be better suited for identifying successful students than TOLT, while the at-risk students in this sample are comparably identified by each model.

Are the at-risk students identified by each predictor a distinct group, which may lead to more specific interventions geared at each group of students?

Since all three predictors (TOLT, Math SAT, and Verbal SAT) used in the two models feature strong correlations with each other (Table 2), it is tempting to hypothesize that poor performance on any one of the three is indicative of poor performance on all, so that a student predicted to be at-risk by one model would also be predicted to be at-risk by the other. However, this turns out to hold true for just over one-half of the cases predicted to be at-risk: of the 471 students predicted to be at-risk by the TOLT model and the 451 students predicted to be at-risk by the SAT model, only 266 of the students were classified as at-risk by both models. It is useful to consider three exclusive categories: students predicted to be at-risk by both models, at-risk by only the TOLT model, and at-risk by only the SAT model. Table 6 shows the number of students that fall into each category (top row), and the resulting performance on the ACS exam for each category (middle two rows). The bottom row of the table presents the rates of correct prediction in each category for comparison.

Table 6. The overlap between models.

Model Predictions of At-risk Status	Only TOLT At-risk (n=205)	Only SAT At-risk (n=185)	Both models At-risk (n=266)
Correct (scored Below Cut-off)	113	108	219
Incorrect (scored Above Cut-off)	92	77	47
% Correct	55.1%	58.4%	82.3%

From Table 6 it appears that each model, TOLT and SAT, describes a distinct trait that hinders success in chemistry. There is a distinct group of 113 students that performed poorly on the TOLT and on the ACS final exam while performing satisfactorily on the SAT measure. A similar situation occurs for 108 students who performed poorly on the SAT and on the ACS final exam while performing reasonably well on the TOLT. These two cases demonstrate that the two models identify different groups of students as being at-risk, even though 219 students were correctly predicted by both models to be at-risk. It should also be noted that neither of the models identifies all students who are at-risk: out of the 1619 students not predicted to be at-risk by either model, only 1257 (77.6%) in fact performed above the cut-off. This will always be the case: the models attempt to identify factors that are necessary for success in general chemistry, but necessary does not mean sufficient.

Statistical comparisons between percent correct predictions employed an arcsine transformation to stabilize variances (Cohen, 1988). The highest percent correct is for those who would be classified as at-risk by both the TOLT model and by the SAT model, demonstrating that a combination of low scores on both measures leads to a greater chance of students performing poorly on the ACS final exam. The differences in correct prediction rate between this 'both' category and each of the two 'only' categories were significant with a medium effect size. No evidence supporting a significant difference in percent correct between the only TOLT category and the only SAT category was found, indicating that neither model isolates a distinct group of at-risk students better than the other.

Can a combination of SAT and formal thought measures provide an advantage in identifying at-risk students?

The previous discussion has shown that, if both the TOLT model and the SAT model predict a student to be at-risk, that is very likely to be the case! However, this post-hoc combination of the predictions of two different models may be too conservative, identifying only a relatively small number of at-risk students. It may be possible to construct a single model using both sets of predictors that will retain a high success rate and identify a larger number of at-risk students. To investigate this possibility, a model (shown in Table 7) was constructed to use both SAT sub-scores and TOLT scores:

Table 7. Combined TOLT and SAT model.

Coefficient	Slope	Std. Error	t-value	p
Constant	-19.477	2.106	-9.250	<0.001
SATV	0.04410	0.004163	10.594	<0.001
SATM	0.07253	0.004738	15.310	<0.001
TOLT	1.0440	0.13574	7.691	<0.001

$R^2 = 0.420$ Model F = 549.274

Each predictor entered the model significantly, indicating that, even when controlling for the variability that comes from the other predictors, both SAT sub-scores and TOLT still

feature a significant relation with the ACS exam, which is consistent with the interpretation of the data in Table 6 that TOLT and SAT map onto performance in distinct ways. (Note that the incorporation of interaction terms TOLT*MSAT and TOLT*VSAT into the model adds only 0.009 to R^2; therefore, these terms were not retained in the model.) Similar to the SAT model, this new model that combines both SAT sub-scores and TOLT scores has many combinations of predictor scores that would result in an at-risk prediction. The combined model predicts 489 students to be at-risk, higher than the 266 predicted by the overlap of the two individual models. Of those 489 students, 354 scored below the ACS final exam cut-off and thus were correctly classified. This leads to a success rate of 72.4%, which is only slightly higher than the 70.5% seen with the TOLT model, and essentially equivalent to the 72.5% rate of the SAT model. Further, of the 354 students correctly classified by this combined model, 351 had been identified by one of the two previous models. Thus the combination of both predictors in a single regression model fails to provide an improved way to identify at-risk students, since only three additional students were correctly found by combining the two sets of predictors. Of the 135 misclassified, 71 scored below average on the exam.

To what extent are all at-risk students identified by this set of predictors?

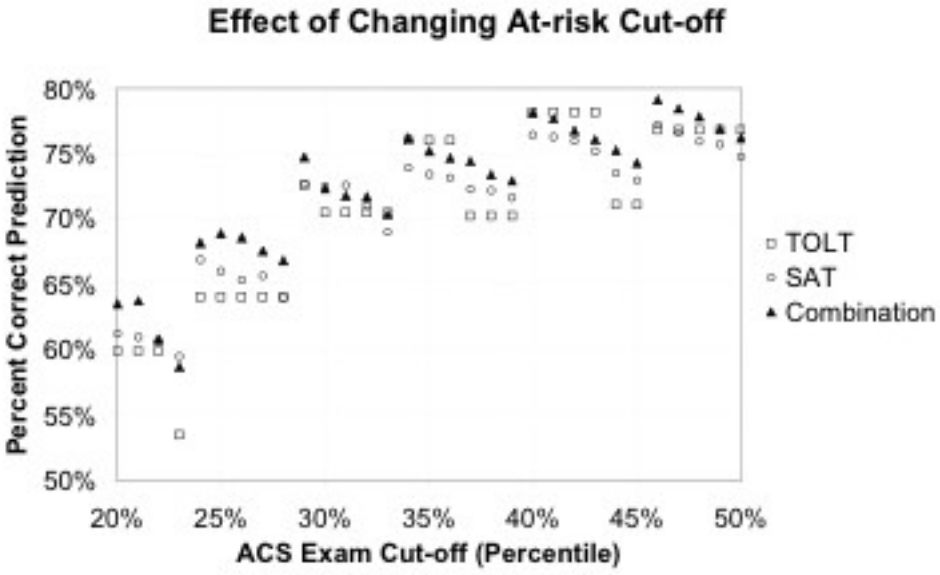
Of the 2275 students, 802 students finished the course below the ACS cut-off. Of these 802 students, 443 students (55.2%) were identifiable based on scores from either TOLT, SAT or a combination of the two. Thus a sizable portion of the students that performed poorly on the ACS exam was not identifiable by these models. We believe this finding may be representative of a need to include non-cognitive predictors, such as affective measures like motivation or confidence, if the goal is to predict all at-risk students. However, it is important to recognize that many non-cognitive factors may feature a strong correlation with general achievement (House, 1995). Rather than attempting to include non-cognitive factors in this study, it would be more appropriate to focus an additional study on the degree to which affective factors are distinct from general achievement, in the same manner as we have set out the comparison between formal thought and general achievement.

At-risk cut-off

It is recognized that the decision to employ the cut-off at the bottom 30% of the sample is somewhat arbitrary, as other values such as the bottom 25% or bottom 33% could reasonably suffice. To address these concerns and to understand the impact of this decision on the conclusions reached, a SAS program was developed to calculate the percent correct predictions for each model as the cut-off point is changed. The results have been plotted in Figure 3.

First, note from Figure 3 that the models switch places depending on the cut-off decision, but all of them remain relatively close together, so that no model offers a distinct advantage over the others in terms of accuracy in identifying at-risk students. Also note the general upward trend of percent correct predictions as the cut-off decision increases. This can be attributed to chance guessing. For example, if the cut-off is placed at 20%, randomly selecting students would get 20% correct prediction in identifying at-risk students. However, if the cut-off was 40%, there would be a 40% chance of identifying at-risk students by random selection. In general, each model stays approximately 35 - 45% above the random selection method of identifying students. Now that we can describe the role of the models in the identification of at-risk students for whom SAT scores were available, we will now turn our attention to another important aspect of this study, the consideration of students for whom SAT scores were not available.

Figure 3. Correct predictions for each model versus the cut-off point.



Those without SAT scores

As mentioned, cases for which SAT scores were unavailable were omitted so that the previous comparisons between the SAT model and the TOLT model could be undertaken for the same group of students. This omission presents some interesting implications for the study. A chief concern with missing data is the presence of a trend in those students who have missing data, because the presence of any such trend represents a limitation in the generalizability. By omitting students without available SAT scores, it is necessary to check if the group omitted differs from the group studied. If so, then the applicability of the analysis to those omitted may be questionable. Table 8 presents the results from this comparison

Table 8. Comparison of those with SAT scores to those without.

	Avg. score for those with SAT (n, st dev)	Avg. score for those without SAT (n, st dev)	t-test	p-value	d-value
TOLT	6.65 (2957, 2.656)	6.24 (841, 2.645)	3.934	0.000	0.155
ACS Exam	52.11 (2284, 16.724)	49.92 (587, 16.178)	2.839	0.005	0.133

The students without SAT scores scored significantly lower on both the TOLT measure and the ACS exam measure than students with SAT scores. The d-value is the effect size for comparing two means, with both values representing a small effect. Because of the differences between students with SAT scores and those without, the students without SAT scores likely represent a non-random population. For this reason we will examine these students separately in terms of the conclusions presented so far.

There were 841 students in the original sample without SAT scores, and 587 of those took the ACS exam. While no claim can be made regarding what the SAT model would have predicted for these students, the role of TOLT in identifying at-risk students can still be investigated. To do this, a new regression model (Table 9) was fitted for just these 587 students.

Table 9. TOLT model for students without SAT scores only.

Coefficient	Slope	Std. Error	t-value	p
Constant	31.910	1.599	19.962	<0.001
TOLT	2.799	0.230	12.149	<0.001

$R^2 = 0.201$ Model F = 147.587

As does the previous TOLT model, this model indicates a positive linear relationship between TOLT and ACS exam scores. An examination of the standard error associated with the TOLT coefficient (0.230) and the intercept (1.599) in this model indicates that it cannot be considered different from the original model. It appears the conclusions reached regarding the previous TOLT model also apply to the students without SAT scores. Of the 587 students, 150 scored at or below a 4 on the TOLT and would therefore be characterized as at-risk. Of these 150 students predicted to be at-risk, 101 scored below the ACS final exam cut-off, giving a 67.3% success rate in classification. Of the 49 incorrectly classified, 20 scored below the average score. In short, the inclusion of students for whom SAT scores were not available reveals nothing inconsistent with the previous findings regarding the utility of TOLT. The conclusion that TOLT as a formal thought measure identifies a barrier to the success of chemistry students holds true for those in our sample without SAT scores. The distinct advantage of the TOLT model over the SAT model here is that, even though SAT scores were unavailable, the ease of administration of the TOLT made it possible to identify correctly an additional 101 students as being at-risk.

Those who did not finish the course

As mentioned earlier, those who did not finish the course represent a significant portion of the at-risk student population. However, while leaving the course may be a function of academic performance during the course, there are also a variety of other reasons for such a departure, ranging from personal health to financial trouble. For this reason, we are hesitant to classify all students who did not finish the course as at-risk students. However, given the nature of the conclusions reached regarding the ability of TOLT to predict performance, it will only be necessary to examine those whose TOLT scores fell at or below 4, to determine if those students were in fact performing poorly when they left the course. This will be approximated by reviewing students' scores on four instructor-generated, multiple-choice in-course tests, in comparison to the class performance on the same test.

Of the 3798 students in this study, 927 students (24.4%) did not take the ACS exam. Of those 927 students, 263 students (28.4%) scored at or below a 4 on the TOLT, which was the criterion previously used for at-risk classification. Forty-two of the 263 students did not take any of the tests, so their decision to drop the course came relatively early, and unfortunately, little else can be said of them. However, of the remaining 221 students, 194 students scored in the bottom 30% within their class on every test they took. Of the remaining 27 students, 22 scored above this mark only once. While it is not possible to extrapolate an exact reason for leaving a course from the data available, and indeed the decision is likely attributable to a number of factors, the data do indicate that low academic performance probably played a role in the decision and that low performance on the TOLT would have served as a warning sign for this population.

Discussion

What we have identified with this study is threefold:

1) Formal thought (as measured by TOLT) and general achievement (as measured by SAT) represent separate and distinct factors, each of which can be used to identify at-risk students.
2) Neither the formal thought measure (TOLT) nor the general achievement measure (SAT) is clearly superior in terms of percent correct identification of at-risk students.
3) The ease of administration of the TOLT makes it possible to use this measure to identify additional at-risk students for whom SAT scores are not available

The fact that both general achievement and formal thought represent distinct factors (see discussion of Table 6) in this study is important. It seems there are at least three groups of at-risk students: those who do not have an appropriate knowledge of mathematics and language for success in chemistry, those who do not have the requisite reasoning skills for success in chemistry, and those who lack both. In other words, even if a student performs reasonably well on the SAT, low formal thought ability can still hinder his or her success in chemistry, and the reverse is also true. Since this result shows that mathematics achievement and reasoning ability represent different barriers to success, effective remediation aimed at these two different groups of students will incorporate a review of relevant mathematical and verbal skills as well as the opportunity to work on developing formal thought ability. A remedial course focused solely at reviewing fundamental mathematical rules in the abstract (e.g. how to isolate variables in an equation, how to manipulate logarithms) is definitely too narrow. Connecting mathematical manipulations with concrete observables in chemistry could provide some assistance to both groups; however, the chemistry content review in a standard remedial course has not typically led to success for a large proportion of students (Bentley and Gellene, 2005). We suggest that attention should be paid to the sequencing of concepts in the chemistry content review, presenting the most abstract only after a concrete foundation has been established (Tsaparlis, 1997; Sanger et al., 2001). Chemistry educators have also achieved some success with improving cognitive skills such as formal reasoning ability in the context of chemistry courses via the integration of carefully sequenced problem-solving activities, incorporating algorithms of increasing complexity, with conceptual instruction sufficient to explain the 'how' and 'why' of the calculations at each level (Tsaparlis, 2005).

From a pedagogical perspective, chemistry review lectures on concrete concepts but with few graphics, animations, or demonstrations require students to create their own mental models of these concepts, a skill associated with formal rather than concrete thinking. Taking advantage of the wide array of animations available for illustrating major concepts in chemistry is perhaps the simplest way to scaffold learners with low formal thought ability in the large lecture setting (Williamson and Abraham, 1995; Tasker et al., 1996; Sanger and Greenbowe, 2000; Wu et al., 2001; Stieff, 2005). Another alternative is the incorporation of computer-assisted learning activities to supplement lectures. Simulations with a focus on manipulating variables and repetition have shown promise with science learners of low formal reasoning ability (Huppert, 2002). In general, researchers have recommended the use of active learning practices to avoid over-dependence on lectures (Chandran et al., 1987; Shayer, 2003). Tien et al. (2002) and Lewis and Lewis (2005) provide examples of effective active learning reforms that de-emphasize lectures without moving completely from the lecture format.

Finally, in all cases, both quantitative and qualitative investigations of the effects of the reform on different groups of students are needed to provide insight into whether and how these reforms assist those who need to develop formal thinking skills as well as chemistry knowledge. How should these investigations be conducted? Indeed, considering formal

thought as a factor distinct from general achievement in college chemistry has implications for research as well as for teaching. The relationship between formal thought as measured by TOLT and chemistry performance (displayed graphically in Figure 1) is important: TOLT is a better predictor of at-risk students than of successful students. Therefore, studies that investigate a relationship between TOLT and academic performance through the use of linear regression or correlation (BouJaoude et al., 2004) may be underestimating the importance of formal thought. A large amount of variation in academic performance for students with high TOLT scores, while congruent with cognitive development theory, would lead to a reduced proportion of variance explained by TOLT as compared with other predictors of performance. In other words, researchers may be misled into thinking formal thought is not relevant for a given situation, when, in fact, the association of low formal thought ability with poor performance is masked by the large variability in performance for those at the higher end of the TOLT. A suggestion for researchers who are considering such models is to dichotomize TOLT scores, creating a low TOLT score and high TOLT score classification, an approach that is better aligned with the theory of developmental stages. Another option would be to consider students with low TOLT scores as a unique subset of students. This latter option should be of particular use when evaluating whether pedagogical reforms are able to help different groups of at-risk students.

Even though this study focuses on college-level general chemistry, it is also worthwhile to consider broader teaching implications. Longitudinal work from Novak has shown that complex science instruction among elementary age students can show improved understanding at the high school level on similar concepts, far removed from the intervention (Novak, 2005). Therefore initiatives to improve formal thought ability could also be instituted earlier in the educational stream, with the strong possibility for improving the trends witnessed here at the college level. One such initiative, Shayer and Adey's Cognitive Acceleration program (Adey and Shayer, 1994), has shown promising results in promoting cognitive development among middle school students.

Conclusions

In this study, formal thought has been found to have a unique relationship to chemistry achievement apart from SAT sub-scores, even though the two constructs share a medium-sized correlation. Low formal thought ability impedes success in chemistry as much as low SAT sub-scores, and formal thought has been shown to represent a necessary factor for success in college-level general chemistry for a distinct group of students. Recommendations for remediation and for future research were discussed in light of these findings. It is important to note that, while both measures used in the study had reasonable success at identifying students at-risk of performing poorly in college-level general chemistry, there was an additional group of students whose poor performance was not predicted by either measure. Therefore, factors that are unaddressed in this paper are also likely to play a role in success in chemistry. Research into affective aspects of chemistry learning with specific emphasis on at-risk students would complement the cognitive approach taken in this paper. In particular, identifying those affective components that prevent students from achieving success despite high cognitive abilities may help identify other distinct groups of at-risk students and lead to the development of targeted remedies for these groups.

Notes

a. It was found that those students completing one year of high school chemistry or more scored significantly higher, with a consistent small effect size (d=0.2), than those who did not on all four measures (Table 10):

Table 10. Comparison of high school chemistry takers with non-takers.

	Average Score (standard deviation, n)		t-value	p-value	d-value
	< 1 full year of high school	≥ 1 full year of high school chemistry			
TOLT	6.19 (2.73, 936)	6.73 (2.61, 2678)	5.437	0.000	0.202
Math SAT	533.5 (84.8, 691)	558.2 (83.1, 2184)	6.770	0.000	0.294
Verbal SAT	523.8 (83.1, 691)	539.6 (81.7, 2184)	4.416	0.001	0.197
ACS Exam	48.0 (15.1, 660)	53.0 (16.9, 2107)	6.737	0.000	0.312

The high school chemistry taking population is therefore likely not representative of the college chemistry taking population. This also supports Chandran et al.'s (1987) postulate that students taking high school chemistry represent a population of students that are likely to succeed in chemistry.

References

Abraham M.R. and Williamson V.M., (1994), A cross-age study of the understanding of five chemistry concepts, *Journal of Research in Science Teaching*, **31**, 147-165.

Acevedo J.A. and Oliva J.M., (1995), Validacion y aplicaciones de un test de razonamiento logico, *Revista de Psicologia General y Aplicada,* **48,** 339-352.

Adey P.S. and Shayer M., (1990), Accelerating the development of formal thinking in middle and high school students, *Journal of Research and Development in Education,* **27**, 267-285.

Adey P.S. and Shayer M., (1994), *Really raising standards: cognitive intervention and academic achievement* (First ed.), New York: Routledge.

Bender D.S. and Milakofsky L., (1982), College chemistry and Piaget: the relationship of aptitude and achievement measures, *Journal of Research in Science Teaching,* **19**, 205-216.

Bentley A.B. and Gellene G.I., (2005), A six-year study of the effects of a remedial course in the chemistry curriculum, *Journal of Chemical Education,* **82**, 125-130.

BouJaoude S., Salloum S. and Abd-El-Khalick F., (2004), Relationships between selective cognitive variables and students' ability to solve chemistry problems, *International Journal of Science Education,* **26**, 63-84.

Bunce D.M. and Hutchinson K.D., (1993), The use of the GALT (Group Assessment of Logical Thinking) as a predictor of academic success in college chemistry, *Journal of Chemical Education,* **70**, 183-187.

Carmichael J.W.J., Bauer J.S., Sevenair J.P., Hunter J.T. and Gambrell R.L., (1986), Predictors of first-year chemistry grades for Black Americans, *Journal of Chemical Education,* **63**, 333-336.

Chandran S., Treagust D.F. and Tobin K.G., (1987), The role of cognitive factors in chemistry achievement, *Journal of Research in Science Teaching,* **24**, 145-160.

Cohen J., (1988), *Statistical power analysis for the behavioural sciences* (2nd ed.), Hillsdale, Lawrence Erlbaum Associates.

Cohen J., Cohen P., West S.G. and Aiken L.S., (2003), *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd ed.), Mahwah, NJ, Lawrence Erlbaum Associates, Inc.

Cohen S.J. and Cronbach L.J., (1985), College Board Scholastic Aptitude Test and Test of Standard Written English, In Boros Institute of Mental Measurement (Ed.), *The Mental Measurements Yearbook* (9th ed.), New Brunswick, Rutgers University Press.

Craney C.L. and Armstrong R.W., (1985), Predictors of grades in general chemistry for allied health students, *Journal of Chemical Education,* **62**, 127-129.

Demerouti M., Kousathana M. and Tsaparlis G., (2004), Acid-base equilibria, Part II. Effect of developmental level and disembedding ability on students' conceptual understanding and problem-solving ability, *The Chemical Educator*, **9**, 132-137.

Educational Testing Service, (2006), About the SAT, Accessed December 2006; http://www.collegeboard.com/student/testing/sat/about.html

Examinations Institute of the American Chemical Society, Division of Chemical Education (1997), *First term general chemistry (special examination)*, Clemson, SC, Clemson University.

Haidar A.H. and Abraham M.R., (1991), A comparison of applied and theoretical knowledge of concepts based on the particulate nature of matter, *Journal of Research in Science Teaching*, **28**, 919-938.

Herron J.D., (1975), Piaget for chemists, *Journal of Chemical Education*, **52**, 146-150.

House J.D., (1995), Noncognitive predictors of achievement in introductory college chemistry, *Research in Higher Education*, **36**, 473-490.

Huppert J., (2002), Computer simulations in the high school: students' cognitive stages, science process skills, and academic achievement in microbiology, *International Journal of Science Education*, **24**, 803-821.

Inhelder B. and Piaget J., (1958), *The growth of logical thinking from childhood to adolescence*, New York, Basic Books Inc.

Jones K.B. and Gellene G.I., (2005), Understanding attrition in an introductory chemistry sequence following successful completion of a remedial course, *Journal of Chemical Education*, **82**, 1241-1245.

Lawson A.E., (1978), The development and validation of a classroom test of formal reasoning, *Journal of Research in Science Teaching*, **15**, 11-24.

Lawson A.E., (1979), The developmental learning paradigm, *Journal of Research in Science Teaching*, **16**, 501-515.

Lawson A.E., (1982), Formal reasoning, achievement, and intelligence: an issue of importance, *Science Education*, **66**, 77-83.

Lawson A.E., (1983), Predicting science achievement: the role of developmental level, disembedding ability, mental capacity, prior knowledge, and beliefs, *Journal of Research in Science Teaching*, **20**, 117-129.

Lawson A.E., (1985), A review of research on formal reasoning and science teaching, *Journal of Research in Science Teaching*, **22**, 569-617.

Lawson A.E. and Nordland F.H., (1976), The factor structure of some Piagetian tasks, *Journal of Research in Science Teaching*, **13**, 461-466.

Lawson A.E. and Renner J.W., (1975), Relationships of science subject matter and developmental levels of learners, *Journal of Research in Science Teaching*, **12**, 347-358.

Lawson A.E., Adi H. and Karplus R., (1979), Development of correlational reasoning in secondary schools: do biology courses make a difference?, *The American Biology Teacher*, **41**, 420-425.

Lawson A.E., Karplus R. and Adi H., (1978), The acquisition of propositional logic and formal operational schemata during the secondary school years, *Journal of Research in Science Teaching*, **15**, 465-478.

Legg M.J., Legg J.C. and Greenbowe T.J., (2001), Analysis of success in general chemistry based on diagnostic testing using logistic regression, *Journal of Chemical Education*, **78**, 1117-1121.

Lewis S.E. and Lewis J.E., (2005), Departing from lectures: an evaluation of a peer-led guided inquiry alternative, *Journal of Chemical Education*, **82**, 135-139.

McFate C. and Olmsted III J.; (1999), Assessing student preparation through placement tests, *Journal of Chemical Education*, **76**, 562-565.

Nakhleh M.B., (1993), Are our students conceptual thinkers or algorithmic problem solvers?, *Journal of Chemical Education*, **70**, 52-55.

Nakhleh M.B., Lowrey K.A. and Mitchell R.C., (1996), Narrowing the gap between concepts and algorithms in freshman chemistry, *Journal of Chemical Education*, **73**, 758-762.

Niaz M., (1987), Relation between M-space of students and M-demand of different items of general chemistry and its interpretation based upon the neo-Piagetian theory of Pascual-Leone, *Journal of Chemical Education*, **64**, 502-505.

Niaz M., (1996), Reasoning strategies of students in solving chemistry problems as a function of developmental level, functional M-capacity, and disembedding ability, *Journal of Chemical Education*, **64**, 502-505.

Nordstrom B.H., (1990), *Predicting performance in freshman chemistry*, Paper presented at the American Chemical Society National Meeting, Boston, Massachusetts.

Novak J.D., (2005), Results and implications of a 12-year longitudinal study of science concept learning, *Research in Science Education*, **35**, 23-40.

Oliva J.M., (2003), The structural coherence of students' conceptions in mechanics and conceptual change, *International Journal of Science Education, 25*(5), 539-561.

Ozsogomonyan A. and Loftus D., (1979), Predictors of general chemistry grades, *Journal of Chemical Education*, **56,** 173-175.

Pederson L.G., (1975), The correlation of partial and total scores of the Scholastic Aptitude Test of the College Entrance Examination Board with grades in freshman chemistry, *Educational and Psychological Measurement*, **35**, 509-511.

Pickering M., (1975), Helping the high risk freshman chemist, *Journal of Chemical Education*, **52**, 512-514.

Pickering M., (1990), Further studies on concept learning versus problem solving, *Journal of Chemical Education*, **67**, 254-255.

Roadrangka V., Yeany R.H. and Padilla M.J., (1983), *The construction and validation of Group Assessment of Logical Thinking (GALT)*, Paper presented at the Annual Meeting of the National Association of Research in Science Teaching, Dallas.

Russell A.A., (1994), A rationally designed general chemistry diagnostic test, *Journal of Chemical Education*, **71**, 314-317.

Sanger M.J., Brinks E.L., Phelps A.J., Pak M.S. and Lyovkin A.N., (2001), A comparison of secondary chemistry courses and chemistry teacher preparation programs in Iowa and Saint Petersburg, Russia, *Journal of Chemical Education*, **78**, 1275-1280.

Sanger M.J. and Greenbowe T.J., (2000), Addressing student misconceptions concerning electron flow in aqueous solutions with instruction including computer animations and conceptual change strategies, *International Journal of Science Education*, **22**, 521-537.

Sawrey B.A., (1990), Concept learning versus problem solving: revisited, *Journal of Chemical Education*, **67**, 253-254.

Shayer M., (2003), Not just Piaget; not just Vygotsky, and certainly not Vygotsky as alternative to Piaget, *Learning and Instruction*, **13**, 465-485.

Shayer M. and Adey P.S., (1981), *Towards a science of science teaching*, London: Heinemann Educational Books.

Shayer M. and Adey P.S., (1992a), Accelerating the development of formal thinking in middle and high school students II: post-project effects on science achievement, *Journal of Research in Science Teaching*, **29**, 81-92.

Shayer M. and Adey P.S., (1992b), Accelerating the development of formal thinking in middle and high school students III: testing the permanency of effects, *Journal of Research in Science Teaching*, **29**, 1101-1115.

Shayer M. and Adey P.S., (1993), Accelerating the development of formal thinking in middle and high school students IV: three years after a two-year intervention, *Journal of Research in Science Teaching*, **30**, 351-366.

Spencer H.E., (1996), Mathematical SAT test scores and college chemistry grades, *Journal of Chemical Education*, **73**, 1150-1153.

Staver J.R. and Gabel D.L., (1979), The development and construct validation of a group administered test of formal thought, *Journal of Research in Science Teaching*, **16**, 535-544.

Stevens J.P., (1999), *Intermediate statistics: a modern approach* (2nd ed.), Mahwah, New Jersey, Lawrence Erlbaum Associates.

Stieff M., (2005), Connected chemistry: a novel modeling environment for the chemistry classroom, *Journal of Chemical Education*, **82**, 489-493.

Tai R.H., Sadler P.M. and Loehr J.F., (2005), Factors influencing success in introductory college chemistry, *Journal of Research in Science Teaching*, **42**, 987-1012.

Tasker R.F., Chia W., Bucat R.B. and Sleet R., (1996), The VisChem Project: visualising chemistry with multimedia, *Chemistry in Australia*, **63**, 395-397.

Tien L.T., Roth V. and Kampmeier J.A., (2002), Implementation of a peer-led team learning instructional approach in an undergraduate organic chemistry course, *Journal of Research in Science Teaching*, **39**, 606-632.

Tobin K.G. and Capie W., (1981), The development and validation of a group test of logical thinking, *Educational and Psychological Measurement*, **41**, 413-423.

Treagust D.F., (1988), Development and use of diagnostic-tests to evaluate student misconceptions in science, *International Journal of Science Education*, **10**, 159-169.

Tsaparlis G., (1997), Atomic and molecular structure in chemical education: a critical analysis from various perspectives of science education, *Journal of Chemical Education*, **74**, 922-925.

Tsaparlis G., (2005), Non-algorithmic quantitative problem-solving in university physical chemistry: a correlation study of the role of selective cognitive factors, *Research in Science and Technological Education*, **23**, 125-148.

Tsaparlis G. and Angelopoulos V., (2000), A model of problem-solving: its operation, validity, and usefulness in the case of organic-synthesis problems, *Science Education*, **84**, 131-153.

Tsaparlis G., Kousathana M. and Niaz M., (1998), Molecular-equilibrium problems: manipulation of logical structure and of M-demand and their effect on student performance, *Science Education*, **82**, 437-454.

Wagner E.P., Sasser H. and DiBiase W.J., (2002), Predicting students at risk in general chemistry using pre-semester assessments and demographic information, *Journal of Chemical Education*, **79**, 749-755.

Williamson V.M. and Abraham M.R., (1995), The effects of computer animation on the particulate mental models of college chemistry students, *Journal of Research in Science Teaching*, **32**, 521-534.

Wu H.K., Krajcik J.S. and SolowayE., (2001), Promoting understanding of chemical representations: Students' use of a visualization tool in the classroom, *Journal of Research in Science Teaching*, **38**, 821-842.

Yager R.E., Snider B. and Krajcik J.S., (1988), Relative success in college chemistry for students who experienced a high-school course in chemistry and those who did not, *Journal of Research in Science Teaching*, **25**, 387-396.

Yarroch W.L., (1991), The implication of content versus item validity on science tests, *Journal of Research in Science Teaching*, **28**, 619-629.