Effect of written presentation on performance in introductory physics

John Stewart* and Shawn Ballard
Physics Department, University of Arkansas, Fayetteville, Arkansas 72701, USA
(Received 15 August 2009; published 27 October 2010)

This study examined the written work of students in the introductory calculus-based electricity and magnetism course at the University of Arkansas. The students' solutions to hourly exams were divided into a small set of countable features organized into three major categories, mathematics, language, and graphics. Each category was further divided into subfeatures. The total number of features alone explained more than 30% of the variance in exam scores and from 9% to 15% of the variance in conceptual posttest scores. If all features and subfeatures are used, between 44% and 49% of the variance in exam scores is explained and between 22% and 28% of the variance in conceptual posttest scores. The use of language is consistently positively correlated with both exam performance and conceptual understanding.

DOI: 10.1103/PhysRevSTPER.6.020120 PACS number(s): 01.40.Di, 01.40.Fk

I. INTRODUCTION

Writing is a fundamental mode of intellectual expression and producing students with superior scientific writing skills is a goal of many physics programs. The need for development of good writing skills in all students has resulted in the writing-across-the-curriculum initiative that recognizes the importance of developing writing skills in all classes and the usefulness of writing as a learning tool [1–3]. Improvement of the technical writing skills of physics students is often a key goal of physics departments [2,4,5] and other science departments [6,7]. Writing is also frequently proposed as an important method for improving student thinking skills or understanding of a topic. The way in which writing is used and the instructional claims made for the efficacy of writing are diverse including "folder activities" using different prompts to use writing as a mechanism for students to confront misconceptions [8], reflective essays as a mechanism to improve textbook comprehension [9], and argumentative essays are a promoter of conceptual change [10]. These interventions are part of the writing-to-learn movement that gained popularity in the 1970s and has generated hundreds of examples of the use of writing to accomplish instructional goals [11–13]. A meta-analysis of studies of writing-to-learn programs in actual classrooms where a control group was used found 75% positive outcomes over the control groups, but only a small average effect size of 0.26 [14]. This meta-analysis found that the details of the writing-to-learn program were important with interventions that required metacognitive thinking yielding an effect size of 0.44 ($p=0.02$). Writing has also been investigated as a key difference in the approach to physics of students of different gender [15]. This study introduces a quantitative characterization of student writing that could be applied across the diverse implementations of the writing-to-learn philosophy in science classes. In this study, the characterization technique is used to examine the amount of the variance in student performance explained by the countable features of student writing.

Many studies have examined the extent to which students' performance in a physics class can be predicted by students' experience previous to entering the class and by the students' native abilities as evidenced by past success or by examination. The studies which follow used regression analysis and the results of the studies will be summarized using the R^2 statistic; this statistic is the ratio of variance explained by the regression model to the original variance in the data.

The solution of physics problems requires the application of logical reasoning. Liberman and Hudson investigated the effect of formal operational reasoning ability on final exam scores [16]. Formal operational reasoning was measured with the Tomlinson-Keasy and Campbell test, which contains questions requiring a range of reasoning skills. The study found $R^2=0.24$ for a regression of a logical reasoning posttest on the course final exam. As part of a separate study, Hudson and Liberman [17] also investigated the effect of formal operational reasoning on final course grade using the same instrument as a pretest. They found $R^2=0.189$ for a regression of the logical reasoning pretest on the final course grade.

Performance in physics classes also involves the application of mathematics. Hudson and Liberman investigated the impact of mathematical reasoning ability on performance in a physics course [17]. Students were given a mathematics pretest involving algebraic reasoning. Regression of the pretest score on the final course grade yielded $R^2=0.119$. Wollman and Lawrenz [18] report $R^2=0.21$ for the correlation between mathematics pretest score and course test performance. Halloun and Hestenes [19] found $R^2=0.26$ for a regression of a mathematics pretest score on course performance. Meltzer found R^2 values 0.01, 0.09, 0.14, and 0.21 for the relation between mathematics pretest and normalized conceptual gain [20].

Physics classes also build on knowledge developed in previous classes, either pre-requisite college classes or high school classes. Halloun and Hestenes [19] found $R^2=0.3$ in a regression of physics pretest score on course performance.

Time-on-task has also been examined as an important factor in student performance in many academic disciplines. Schmidt [21] investigated a combination of detailed time-on-task measures, demographic data, and past performance measures such as ACT score. These variables accounted for $R^2=0.44$ of the variation in the multiple-choice scores on a final examination in an economics class. Admiraal *et al.* [22] used

*johns@uark.edu

detailed time-on-task measurements to calculate an effect size for two different class structures in graduate legal education. Stewart *et al.* [23] found the amount of time students spent on the various activities associated with a physics class (reading, working homework, studying, etc.) produced R^2 from 0.26 to 0.41 when regressed upon exam performance.

Many studies have also investigated the effect of combinations of factors on student performance. Hudson and Liberman found $R^2=0.268$ for a regression of both mathematical and logical reasoning pretest scores on final course grades [17]. Wollman and Lawrenz report $R^2=0.569$ using a mathematics pretest, ACT score, and total GPA in a regression on course test grades [18]. Champagne *et al.* investigated the combined roles of preexisting conceptual physics knowledge, logical reasoning ability, and mathematical reasoning ability and found that these three variables yielded $R^2=0.325$ when regressed against an achievement statistic that averaged the scores on three hourly exams and the scores on the mechanics part of the final exam [24]. Halloun and Hestenes measured $R^2=0.4$ for the math and physics pretests combined, $R^2=0.15$ for previous physics and math courses, and $R^2=0.49$ combining all these factors when regressed upon course performance [19]. Sadler and Tai [25] examined a combination of demographic information, high school GPA, course decisions, including the decision to take honors or AP courses in high school, physics teacher's pedagogical choices, and many other factors. They found $R^2=0.26$ of the variance of introductory physics grades for students who took physics in high school was explained by these factors. This rose to $R^2=0.36$ for students who did not take physics in high school.

In summary, logical reasoning accounts for 19–24 % of the variance in student performance in a physics class, mathematical reasoning 12–26 %, physics pretest scores up to 30%, and time on task 26–41 %. Combinations of factors can explain up to 57% of the variance in performance.

The above studies [1–7] show that integration of writing into science classes and the details of that writing is a longstanding and active avenue for improvement for science classes. This study seeks to place the amount of variance explained by quantitative features of that writing among other measured features of science students including logical reasoning [16,17], mathematical reasoning [17,18,24], mathematics and physics preparation [19], and time-on-task [23]. This study seeks to answer two questions: (1) to what extent is the variance in student performance explained by the quantitative features of student writing and (2) what features of student writing are most important for performance?

II. METHODOLOGY

This study examined the written work of students in the introductory calculus-based electricity and magnetism course at the University of Arkansas.

A. Sampling

This study counted the writing elements found in the hourly exams given in the second-semester calculus-based introductory electricity and magnetism class at the University of Arkansas during the fall 2006 and fall 2007 semesters. This class, reformed under a CCD grant, and further refined as part of the PhysTEC project, has longitudinal statics over the last eight years that indicate it produces a stable performance each semester.

Four hourly exams were given each semester. Each exam contained nine multiple-choice problems and three open-response problems. Only the writing that was part of the solution of the open-response questions was recorded. The test score used in the analysis was the score of all twelve questions on the exam; therefore, the multiple-choice questions on the tests were counted toward the exam score.

Students took the Conceptual Survey in Electricity and Magnetism (CSEM) [26] as a pretest and posttest to assess conceptual learning. The normalized gain was calculated for each student to measure conceptual learning gains,

$$\text{normalized gain} = \frac{\text{post-test} - \text{pretest}}{100 - \text{pretest}}. \quad (1)$$

Only students who completed the class were included in the study: $N=87$ for fall 2006 and $N=135$ for fall 2007. In fall 2006, 81 students, 93%, completed the posttest and 74 students, 85%, completed both the pretest and posttest allowing the calculation of a normalized gain. In fall 2007, 127 students, 94%, completed the posttest and 121 students, 90%, completed both the pretest and posttest allowing the calculation of a normalized gain. Therefore, the population of students used in correlation and regression analysis is slightly different for the test average, posttest score, and normalized gain. These populations are distinguished by the lab sessions missed, since the pretest and posttest are given in lab. The different populations may produce some bias in the analysis but since in all cases the vast majority of the class was included in any subset of students used for analysis, these biases are expected to be small.

The open-response questions typically have multiple parts requiring either mathematical, graphical, or linguistic responses. A complete response may (and preferably should) include other modes of expression than those explicitly required by the solution; a quantitative problem is enhanced by a drawing and linguistic description. An analysis of the composition of the open-response questions for each semester is presented in Table I.

Some instruction in scientific presentation is provided by the class. Students are encouraged to give solutions symbolically before performing numeric calculations. These numeric calculations are also expected to include proper units where appropriate. Additionally, certain exam questions ask specifically for graphical or linguistic description. The importance of good presentation for successfully mastering physics is also discussed. Presentation standards are only passively enforced with students only losing points for egregious presentation, therefore the measurements presented should be more representative of the students' native writing style than the effects of any course writing policy. Each physical subdiscipline has unique solution patterns. The combination of language, mathematics, and graphics found in this class should be fairly representative of the students' native writing habits,

TABLE I. Summary of hourly exams—the tests are broken down into the primary form of communication expected in the answer. The expected communication was determined from the statement of the question; questions asking for a drawing were classified as graphical, questions asking for an explanation without a calculation, linguistic, and questions asking for a calculation, mathematical. The "Points" columns report the percentage of the free-response points allocated to each type of question. The "Score" column reports the average percentage scored on each type of problem. The "Parts" column reports the total number of parts for all free-response problems for each semester.

		Linguistic		Mathematical		Graphical	
Semester	Parts	Points (%)	Average (%)	Points (%)	Average (%)	Points (%)	Average (%)
Fall 06	40	5	63	76	69	19	76
Fall 07	40	8	93	63	78	29	86

but the analysis may be biased in some manner by the topical material of the class.

B. Measurement

A student's solution to a physics problem is a complex combination of language, mathematics, and drawings. To allow reliable measurement, each solution was divided in a small set of countable features. By focusing on countable features, the subjective issues of writing analysis such as handwriting, neatness, and style are eliminated. The major classes of writing found in students' work, mathematics, language, and graphics, were further divided into subcategories. The mathematics group was divided into symbols, operators, relations, and numbers; the graphical group into graphed objects, graphed symbols, and graphed words; and the language group into words and sentences. The definition of the constituents of each group follows.

1. Mathematical elements

Symbols. Variables (x, y, t), constants (μ_0, ϵ_0, e), and letters in languages other than English (π, Δ, γ).

Operators. Mathematical operations including +, −, /, trigonometric functions, cross and dot products, integrals, derivatives, and absolute values. Implied multiplication is not counted.

Relations. Relations in mathematical expressions, i.e. inequalities and equal signs.

Numbers. Constants (2, 3), fractions ($\frac{3}{4}$), simple numbers with units (1.2 m, 5000 V), and numbers in scientific notation (4×10^{-12} F). If a number includes a unit, the unit is counted with the number so 4×10^{-12} F would be counted as one number.

2. Graphical elements

Graphed objects. Lines and curves, such as circles or exponential functions, as well as special composite objects, such as solenoids or batteries.

Graphed symbols. Presentation on a graph that would normally fall under mathematics, such as symbols, numbers, or other symbolic or numeric graph labels.

Graphed words. Words written on a graph.

3. Language elements

Words. English letters listed together in a recognizable pattern such that this string is distinguishable from a chain of variables.

Sentences. Complex word strings that include a verb and/or some form of punctuation.

The above categories were developed previous to measurement by observing homework. The top-level division into language, mathematics, and graphics was natural. The subdivision of language into words and sentences was traditional and also represented a strong delineation of the way language is used in the solution of physics problems. The division of the graphical objects simply formed subcategories of objects found within graphical presentation based on the main divisions language, mathematics, and graphics: language in graphs (Graphed Words), mathematics in graphs (Graphed Symbols), and purely graphical objects (Graphed Objects). Observation of student work suggested mathematics be divided into numbers and symbols. Symbols were further divided into symbols that stand or could stand for a number, symbols that accomplish an operation, and symbols that express a relation. The symbols that express a relation, mostly equal signs, were separated in analogy to the sentence separation in language category. Operators were divided out of symbols because it seemed the number of operations performed might affect features like the complexity of the mathematics differently than the number of simple placeholder symbols.

The tests were observed in the five day period between when the tests were given and when they were returned. One to two physics undergraduates, former members of the class being studied, performed the measurement using the description of the various elements presented earlier and with the prescription that all writing presented as part of the solution found on the tests must be counted in some category. Writing such as crossed-out work not meant as part of the solution was not counted. This forced the researchers to resolve any grey areas of the descriptions for themselves; however, given the excellent researcher-to-researcher reliability (reported below), any ambiguity in the description of elements did not generate significant errors. Counts for each element were recorded and converted to electronic form for analysis.

The use of a measurement of reliably countable features of student writing to investigate the correlation between writ-

ing and performance had several motivating factors. The method is highly reliable and does not require the observer to make judgments about the quality of the writing. While relatively value-judgment free, the categories can be used to characterize a surprisingly rich set of features that have been advanced at some point as characteristics of good physics problem presentation [27]. For example, the linguistic complexity of the writing can be measured by the words per sentence. The abstract nature of the presentation can be quantified by the ratio of numbers to symbols. The multirepresentational nature of the presentation can be quantified by the relative size of the language, mathematics, and graphics categories. Regression analysis, then, on the counts in the various categories can be used to find the most important combination of the variables without inserting our own model of important textual features. This was desirable because as was shown in the introduction, the prescriptions on the use of writing to improve instruction are exceptionally diverse, and the writing examined in this study is not a result of a specific writing intervention, but rather represents the students' normal written response to physics questions.

C. Reliability

To test the reliability of the above definitions, two researchers, junior physics undergraduates who had taken the class in previous semesters, observed 30 tests using the above definitions and under the restriction that all written elements meant to contribute to the solution had to be placed in one subcategory and found a 99% agreement in the total element counts. Agreement for individual features such as word count or symbol count ranged from 100% to 98.5% except for Graphed Objects which had an agreement of 96.3%. The graphed objects were, by their definition, the most difficult to count reliably. Still, this represents an exceptionally reliable observation, as would be expected from the very specific description of each category. The division into the nine categories performed well in this test with all written elements assigned to one of the categories by each researcher; as such a default "other" category was not introduced.

D. Composite variables and scaling

Four new variables were introduced as defined in Eq. (2). These variables represent each subsection total and the total of all presentation elements.

$$\text{Total Language Elements} = \text{Words} + \text{Sentences}$$

$$\text{Total Math Elements} = \text{Numbers} + \text{Symbols} + \text{Relations} + \text{Operators}$$

$$\text{Total Graphical Elements} = \text{Graphed Objects} + \text{Graphed Symbols} + \text{Graphed Words}$$

$$\text{Total Elements} = \text{Total Language Elements} + \text{Total Math Elements} + \text{Total Graphical Elements} \quad (2)$$

The measured features of student writing and the group totals introduced in Eq. (2) all depend on the verbosity of the student, how much he or she writes. We would also like to study how the relative features of a student's writing affect his or her performance. For example, we would like to determine if relatively more of the elements in a solution were in graphical presentation does this correlate to superior conceptual evaluation performance. To accomplish this, a new set of variables was introduced by dividing each of the measured features and the group totals by the total number of presentation elements. For example, the variable Number Ratio=Numbers/Total Elements and Total Mathematical Element Ratio=Total Mathematical Elements/Total Elements. These variables will be called "Scaled Variables."

E. Statistical analysis

The SAS statistics system was used to calculate correlations between all variables and the test average, the posttest score on the CSEM, and the normalized gain on the CSEM. SAS was also used to construct linear models of the test average, posttest, and the normalized gain first using unscaled variables and then using the scaled variables. All variables used in the linear regression analysis and the correlation analysis were standardized by subtracting the mean and dividing by the standard deviation. Care is required in multiple linear regression analysis when selecting the best model; the addition of more regressors usually improves the R^2 of the model. All models reported in Table III and in Table V are the combination of variables that maximize R^2_{adj}. The R^2-adjusted statistics corrects R^2 for the loss of degrees of freedom resulting from the addition of regressors. All regressions were performed using the SAS REG procedure.

III. RESULTS FOR TOTAL PRESENTATION

The correlation analysis for the two-semester measurement is shown in Table II. Table II presents the correlation of features measured from student work and the composite statistics defined in Eq. (2). The correlation with the student test average, r_{test}, with the posttest score, r_{post}, and with the normalized gain, r_{gain}, is presented.

Correlation of the total number of presentation elements, Total Elements, with test average was strong ($r=0.60$ and $r=0.57$) and consistently significant at the $p<0.0001$ level for both semesters. The correlation with posttest ($r=0.39$ and $r=0.29$) and normalized gain ($r=0.31$ and $r=0.32$) was weaker, but still significant at the $p<0.05$ level. All components of mathematical presentation showed a strong correlation with test average. Mathematical presentation was less correlated with posttest score ($r=0.36$ and $r=0.24$) and normalized gain ($r=0.32$ and $r=0.25$), but still significantly correlated at the $p<0.05$ level. Language use had a weaker correlation with test performance ($r=0.46$ and $r=0.37$) than mathematics use ($r=0.60$ and $r=0.60$), but still maintained a

TABLE II. Total test presentation—r_{test} is the correlation with test average, r_{gain} is the correlation with the normalized gain on the CSEM, and r_{post} is the correlation with the posttest score. For fall 2006, $N=87$ students were included in the study. A test average was calculated for all students, 81 students completed the posttest, and a normalized gain was calculated for 74 students. For fall 2007, $N=135$ students were included in the study. A test average was calculated for all students, 127 students completed the posttest, and a normalized gain was calculated for 121 students. All variables were normalized.

	Fall 2006			Fall 2007		
Variable	r_{test}	r_{gain}	r_{post}	r_{test}	r_{gain}	r_{post}
Words	0.46[a]	0.23	0.37[b]	0.37[a]	0.37[a]	0.31[b]
Sentences	0.39[b]	0.24[b]	0.34[b]	0.41[a]	0.39[a]	0.39[a]
Total language elements	0.46[a]	0.23[b]	0.36[b]	0.37[a]	0.37[a]	0.32[b]
Numbers	0.52[a]	0.27[b]	0.27[b]	0.55[a]	0.21[b]	0.20[b]
Symbols	0.62[a]	0.33[b]	0.38[b]	0.49[a]	0.21[b]	0.18[b]
Relations	0.60[a]	0.29[b]	0.34[b]	0.48[a]	0.19[b]	0.18[b]
Operators	0.51[a]	0.31[b]	0.34[b]	0.63[a]	0.30[b]	0.31[b]
Total mathematical elements	0.60[a]	0.32[b]	0.36[b]	0.60[a]	0.25[b]	0.24[b]
Graphed objects	0.07	0.02	0.04	0.30[b]	0.14	0.12
Graphed symbols	0.19	0.06	0.09	0.19[b]	−0.04	−0.01
Graphed words	0.22[b]	0.10	0.20	0.18[b]	0.01	0.01
Total graphical elements	0.17	0.05	0.09	0.26[b]	0.00	0.03
Total elements	0.60[a]	0.31[b]	0.39[b]	0.57[a]	0.32[b]	0.29[b]

[a]Correlation with $p<0.0001$.
[b]Correlation with $p<0.05$.

correlation at the $p<0.0001$ level for most variables.

The use of graphics was weakly correlated with all performance measures. This observation is somewhat curious, since one would expect graphical reasoning to play a role in successful conceptual reasoning. This measurement seems to indicate that more graphical reasoning or more thorough graphical reasoning as evidenced by increased use of words or symbols within the graphics was not important in either a student's qualitative or quantitative mastery of the material. This observation is somewhat explained by data presented in Table I which shows students performed better on graphical problems. If the graphical reasoning is used in problems that score higher on average, it may not correlate with the test totals for which graphical problems are only a small component.

To examine the total variance accounted for by all the presentation measures, linear regression analysis was used to build linear models for the test average, normalized gain, and posttest score using the un-scaled variables. The result of this analysis is presented in Table III. For each dependent variable, three regressions were performed: first, only the total count of presentation elements was used (Total Elements), second, only the totals of each major subgroup were used (Total Language, Total Math, and Total Graph), and finally all variables were used. All models reported in Table III and later in Table V use standardized variables and maximize R^2_{adj} as described in Sec. II E.

Regression on test average with only Total Elements yields $R^2=0.32$ and 0.36. The linear models for the test average were both significant at the $p<0.0001$ level. Regression on normalized gain and posttest yields R^2 from 0.08 to 0.15. These models were significant at the $p<0.05$ level.

The second level of detail regresses the group totals, Total Language Elements, Total Mathematical Elements, and Total Graphical Elements. In all cases, R^2 improved for the maximum R^2_{adj} models. In four of the six regressions, all three subgroup totals were included. For the two models that included only two subgroup totals, only Total Mathematical Elements was select for both models. The models for test average continued to be significant at the $p<0.0001$ level as did the fall 2007 normalized gain model; the other models were significant at the $p<0.05$ level.

The most detailed regressions used all variables. In all cases, the models at this level had higher R^2 than the less detailed models. The variables selected for the maximum R^2_{adj} models were different for each model. These regressions yielded $R^2=0.44$ and 0.49 for test average, a very large R^2 when compared with studies presented in the introduction. Regression on posttest yielded $R^2=0.22$ and 0.28 and regression on normalized gain yielded $R^2=0.15$ to 0.27. Normalized gain contains a measure of pre-preparation, the pretest score, which may be more difficult to model than the posttest score using written presentation. The analysis shows that the number and distribution of written elements in students' work explains a substantial amount of the variation in student performance in a physics class and student conceptual mastery of the subject. All models were significant at the $p<0.0001$ level except fall 2006 normalized gain and fall 2006 posttest.

IV. RESULTS FOR SCALED PRESENTATION

The strong positive correlation of the components of student's exam presentation found in Table II is hardly surpris-

TABLE III. Regression with Standardized Variables—Subset models may contain only the variables Total Language, Total Graph, or Total Math.

	N	R^2	R^2_{adj}	Variables	Model
Fall 06 Test	87	0.36[a]	0.36	Total	Total Elements
	87	0.45[a]	0.43	Subset	Total Language, Total Math, Total Graph
	87	0.49[a]	0.46	All	Relations, Graphed Words, Total Graph, Total Elements
Fall 07 Test	135	0.32[a]	0.32	Total	Total Elements
	135	0.37[a]	0.36	Subset	Total Language, TotalMath
	135	0.44[a]	0.42	All	Operators, Sentences, Symbols
Fall 06 Gain	74	0.10[b]	0.08	Total	Total Elements
	74	0.13[b]	0.10	Subset	Total Math, Total Graph
	74	0.15[b]	0.11	All	Sentences, Operators, Graphed Symbols
Fall 07 Gain	121	0.10[b]	0.09	Total	Total Elements
	121	0.17[a]	0.15	Subset	Total Math, Total Language, Total Graph
	121	0.27[a]	0.24	All	Sentences, Symbols, Operators, Graphed Words, Total Graph
Fall 06 Post-test	81	0.15[b]	0.14	Total	Total Elements
	81	0.20[b]	0.17	Subset	Total Language, Total Graph, Total Math
	81	0.22[b]	0.18	All	Numbers, Graphed Words, Total Graph, Total Elements
Fall 07 Post-test	127	0.09[b]	0.08	Total Elements	Total Elements
	127	0.13[b]	0.11	Subset	Total Language, Total Graph, Total Math
	127	0.28[a]	0.25	All	Sentences, Operators, Symbols, Graphed Words, Total Elements

[a]Model that is significant with $p < 0.0001$.
[b]Model that is significant with $p < 0.05$.

ing. In general, one would expect that a student who could work more of the exam would write more. While the strong correlation was expected, it was not guaranteed. Many instructors have had the experience of the student who had no idea how to address a problem trying to cover by writing an extensive solution [28]. This unfortunate behavior is, however, not so prevalent as to obscure the relation between more presentation and better exam performance. To examine the effect of the relative features of writing, the analysis of the previous section was repeated with the scaled variables. Since the scaled variables measure the relative amount of writing that goes into different modes of communication, they can be used to investigate how the relative features of the writing impact performance, eliminating the effect of more writing implying more performance. For example, the scaled variables can be used to investigate whether students who use relatively more words than numbers in their writing perform better on a conceptual posttest.

Correlation analysis was repeated for the scaled variables and is presented in Table IV. The correlations presented in Table IV are quite different than those presented in Table II, lending support to the proposition that the source of the strong correlations in Table II was the effect of more writing implying better performance. A few patterns emerge. Mathematics use is no longer strongly positively correlated with test performance ($r=-0.11$ and $r=-0.10$). The ratio of Graphed Objects to Total Elements is strongly and significantly negatively correlated with test performance ($r=-0.60$ and $r=-0.45$). This may have a number of explanations. First, Table I shows that students' performance on graphical questions is superior to their performance on mathematical or language-based problems. If relatively more of a student's presentation goes toward solving problems with a higher average, then this may indicate he or she cannot solve the problems with a lower average. Secondly, it may indicate that drawing as a primary mode of expression is not effective for the presentation and solution of physics problems.

Language elements are positively correlated with performance on tests, posttests, and normalized gain. This consistent positive correlation sheds light on the negative correlations found in Table IV. Since all variables are scaled by the Total Elements, the noncomposite variables must add to one. Therefore, if, for example, the number of symbols found in a student's work increases, while the number of words is fixed, the Symbol Ratio increases while the Word Ratio decreases. The negative correlations indicate that if the preponderance of a student's work was in mathematical elements or graphical elements, then the student on average performed more weakly on tests and on the conceptual exam.

The scaled variables were regressed upon the test average, the posttest score, and the normalized gain as shown in Table V. Both the scaled subgroup totals and the full set of scaled variables were used. The scaled variables explained some-

TABLE IV. Scaled Total Test Presentation—All variables in this table were scaled by dividing by Total Elements. r_{test} is the correlation with test average, r_{gain} is the correlation with the normalized gain on the CSEM, and r_{post} is the correlation with the posttest score. For fall 2006, $N=87$ students were included in the study. A test average was calculated for all students, 81 students completed the posttest, and a normalized gain was calculated for 74 students. For fall 2007, $N=135$ students were included in the study. A test average was calculated for all students, 127 students completed the posttest, and a normalized gain was calculated for 121 students.

	Fall 2006			Fall 2007		
Scaled variable	r_{test}	r_{gain}	r_{post}	r_{test}	r_{gain}	r_{post}
Words ratio	0.33[b]	0.19	0.31[b]	0.28[b]	0.34[b]	0.32[b]
Sentences ratio	0.24[b]	0.20	0.26[b]	0.33[b]	0.35[a]	0.39[a]
Total language elements ratio	0.33[b]	0.19	0.31[b]	0.29[b]	0.35[a]	0.33[b]
Numbers ratio	−0.34[b]	−0.20	−0.30[b]	−0.17[b]	−0.16	−0.17
Symbols ratio	0.18	0.05	0.00	−0.08	−0.16	−0.19[b]
Relations ratio	−0.11	−0.14	−0.17	−0.22[b]	−0.19[b]	−0.19[b]
Operators ratio	−0.02	0.11	0.02	0.18[b]	0.05	0.09
Total mathematical elements ratio	−0.11	−0.07	−0.18	−0.10	−0.17	−0.18[b]
Graphed objects ratio	−0.60[a]	−0.34[b]	−0.39[b]	−0.45[a]	−0.25[b]	−0.26[b]
Graphed symbols ratio	−0.15	−0.10	−0.12	−0.20[b]	−0.24[b]	−0.19[b]
Graphed words ratio	0.09	0.04	0.13	0.11	−0.08	−0.05
Total graphical elements ratio	−0.41[a]	−0.23[b]	−0.26[b]	−0.31[b]	−0.29[b]	−0.25[b]

[a]Correlation with $p<0.0001$.
[b]Correlation with $p<0.05$.

what less variance of the test average than the unscaled variables, but explained equal variance for the posttest score and normalized gain. The R^2 for models involving scaled subset totals was substantially less than that of the full set of scaled variables indicating that it is the details of the presentation that are important in explaining performance. As was found with the regressions of the unscaled variables, no general set of variables was present in the minimum R^2_{adj} model for all regressions in Table V; therefore, no subset of the presentation variables emerges as the key to explaining performance.

TABLE V. Regression with Scaled Variables: All variables were scaled by dividing by the total elements. The "*R*" postfix represents a scaled variable, a variable that has been divided by Total Elements. Subset models may contain only the variables Language Ratio, Graph Ratio, or Mathematics Ratio.

	N	R^2	R^2_{adj}	Variables	Model
Fall 06 Test	87	0.20[a]	0.18	Subset	Language Ratio, Graph Ratio
	87	0.46[a]	0.43	All	NumberR, RelationsR, GraphedSymbolR, GraphedWordsR, GraphedObjectsR
Fall 07 Test	135	0.14[a]	0.12	Subset	Language Ratio, Graph Ratio
	135	0.33[a]	0.30	All	WordsR, SentenceR, OperatorsR, GraphedWordsR, GraphedObjectsR
Fall 06 Gain	74	0.05[b]	0.04	Subset	Graph Ratio
	74	0.13[b]	0.11	All	NumbersR, GraphedObjectsR
Fall 07 Gain	121	0.16[a]	0.15	Subset	Language Ratio, Graph Ratio
	121	0.28[a]	0.24	All	WordsR, SentencesR, SymbolsR, OperatorsR, NumbersR, GraphedObjectsR
Fall 06 Post-test	81	0.12[b]	0.09	Subset	Language Ratio, Math Ratio
	81	0.22[b]	0.19	All	NumbersR, GraphedWordsR, GraphedObjectsR
Fall 07 Post-test	127	0.13[b]	0.12	Subset	Graph Ratio, Math Ratio
	127	0.29[a]	0.26	All	WordsR, SentencesR, SymbolsR, OperatorsR, GraphedWordsR

[a]Model that is significant with $p<0.0001$.
[b]Model that is significant with $p<0.05$.

V. DISCUSSION

The main features that emerged from the analysis were, first, that the number of presentation elements found in students' solution of test problems explained from 32–36 % of the variance in test performance but only 9–15 % of the variance in the conceptual posttest. The inclusion of all nine measured variables, their subgroup totals, and the overall total increased this to 44–49 % for test average and 22–28 % for posttest score. The variance explained decreased slightly to 46% and 33% for test average and to 22% and 29% for posttest score when the variables were scaled by the total number of presentation elements. When placed among the research studies presented in the Introduction, the measured counts of student writing and the scaled variables accounted for more of the variance in student test performance than any other single reported measure including logical reasoning ability [16,17], mathematical reasoning ability [17,18,24], physics pretest [19], and time-on-task [23]. This is especially impressive since most of the studies reviewed in the introduction used regressors that contained some graded measure of student performance, such as a pretest score, while this study used no regressor based on a student score.

The amount of language, sentences and words, found in the student's test solutions was consistently correlated with performance both on the tests and on the conceptual posttest. This correlation with test average decreased slightly when the variables were scaled, but the correlation with posttest score was consistent. The total amount of language used and the fraction of writing that went into language use emerged as the most consistently positively correlated variables with conceptual performance.

The amount of mathematics use of any kind was strongly correlated with performance. These strong correlations largely vanished when the variables were scaled. The most significant correlation of the scaled mathematics variables with test average was a negative correlation ($r=-0.34$ and $r=-0.17$) with Number Ratio=Numbers/Total Elements. One of most common suggestions for the improvement of physics writing is to carry out more of the calculation with symbols, not numbers. The observation of a negative correlation with Number Ratio indicates that a relatively higher use of numbers is a feature of poorly performing student writing, supporting the standard advice.

The scaled graphics variables were strongly negatively correlated with test performance ($r=-0.41$ and $r=-0.31$), and to a somewhat lesser extent with posttest and normalized gain. Drawing as the primary mode of expression was not a feature of high performing students.

This study measured only the writing on tests in two semesters of the same course at one institution. The experiment should be expanded to other topics and different classes before the results can be viewed as universal. The measurement technique should be applicable to most problem solutions at any level in physics. It would be interesting to investigate the evolution of writing behavior from introductory to advanced classes and to determine if any features of writing that emerge as a student advances are strongly correlated with performance.

VI. IMPLICATIONS FOR INSTRUCTION

This project was initiated because of strong anecdotal evidence at the University of Arkansas that improving student presentation of physics solutions improves learning. It has been our experience, however, that efforts to improve presentation are very unpopular with students and very expensive in terms of instructor and grader time. This project sought to establish a quantitative measure of the effect of the quantity and quantitative features of student writing in the solution of physics problems on the student's performance in the class. If no such connection could be found, it would be difficult to support the continuation of expensive efforts to improve writing. The R^2 values observed show that the measured features of student writing explain as much variance in test performance as logical reasoning ability, pretest score, or many other performance-based measures. As such, this study provides support for the continuation of efforts to get students to express their physics solutions more completely and to use more linguistic description in their solutions. Also, the strong correlation found suggests that caution should be exercised in relying solely on multiple-choice or online problems for homework and tests. Naturally, correlation does not imply causation, so simply making students write more may not improve learning. Writing-to-learn interventions have generally produced small effect sizes [14]. However, for the class studied, more writing and writing containing more language and fewer numbers was on average a feature of better-performing students. This suggests a program that asks a students to explain their work more thoroughly in words and to carry out more of their mathematical reasoning symbolically could be an aid to learning.

VII. FUTURE

The strength of the result suggests that more research is warranted. Additional measurements of writing in other physics topics and at different institutions should be carried out. Further, an experiment is planned that implements a strong policy encouraging good writing behavior for some sections of the class in this study, while other sections are used as a control. This future study could test the efficacy of writing policies and further test the usefulness of writing measures in explaining performance.

VIII. CONCLUSION

This study sought to answer two questions: (1) to what extent is the variance in student performance explained by the quantitative features of student writing and (2) what features of student writing are most important for performance? The total amount of writing in nine categories explained more variance in performance than any other single reported measure. The details of the allocation of student writing among the measured categories, as characterized by the

scaled variables, also explained as much variance in test performance as performance-based measures examined in other studies. The regression analysis did not identify a subset of the observations that were key to increased student performance; however, the amount of language was consistently and significantly positively correlated with test average and posttest score.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation as part of the evaluation of improved learning for the Physics Teacher Education Coalition, Grant No. PHY-0108787.

[1] T. F. Slater, Engaging Student Learning in Science Through Writing Tasks, Phys. Teach. **46**, 123 (2008).
[2] W. J. Mullin, Writing in Physics, Phys. Teach. **27**, 342 (1989).
[3] C. Bazerman, J. Little, L. Bethel, T. Chavkin, D. Fouquette, and J. Garufis, *Reference Guide to Writing Across the Curriculum* (Parlor Press, West Lafayette, IN, 2005).
[4] C. A. Carlson, A Simple Approach to Improving Student Writing, J. Coll. Sci. Teach. **36**, 48 (2007).
[5] R. E. Rice, Scientific Writing—A Course to Improve the Writing of Science Students, J. Coll. Sci. Teach. **27**, 267 (1998).
[6] C. L. Jerde and M. L. Taper, Preparing Undergraduates for Professional Writing, J. Coll. Sci. Teach. **33**, 34 (2004).
[7] M. E. Walvoord, M. H. Hoefnagels, D. D. Gaffin, M. M. Chumchal, and D. A. Long, An Analysis of Calibrated Peer Review (CPR) in a Science Lecture Classroom, J. Coll. Sci. Teach. **37**, 66 (2008).
[8] T. L. Hein, Using Writing to Confront Student Misconceptions in Physics, Eur. J. Phys. **20**, 137 (1999).
[9] C. Kalman, M. W. Aulls, S. Rohar, and J. Godley, Students' Perceptions of Reflective Writing as a Tool for Exploring an Introductory Textbook, J. Coll. Sci. Teach. **37**, 74 (2008).
[10] C. S. Kalman, S. Rohar, and D. Wells, Enhancing Conceptual Change Using Argumentative Essays, Am. J. Phys. **72**, 715 (2004).
[11] J. Kalman and C. Kalman, Writing to Learn, Am. J. Phys. **64**, 954 (1996).
[12] V. Prain, B. Hand, and S. Kay, Writing for Learning in Physics, Phys. Teach. **35**, 40 (1997).
[13] L. P. Rivard, A Review of Writing to Learn in Science: Implications for Practice and Research, J. Res. Sci. Teach. **31**, 969 (1994).
[14] R. L. Bangert-Drowns, M. M. Hurley, and B. Wilkinson, The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis, Rev. Educ. Res. **74**, 29 (2004).
[15] H. Stadler, R. Duit, and G. Benke, Do Boys and Girls Understand Physics Differently?, Phys. Educ. **35**, 417 (2000).
[16] D. Liberman and H. Hudson, Correlation Between Logical Abilities and Success in Physics, Am. J. Phys. **47**, 784 (1979).
[17] H. Hudson and D. Liberman, The Combined Effect of Mathematics Skills and Formal Operational Reasoning on Student Performance in the General Physics Course, Am. J. Phys. **50**, 1117 (1982).
[18] W. Wollman and F. Lawrenz, Identifying Potential "Dropouts" from College Physics Classes, J. Res. Sci. Teach. **21**, 385 (1984).
[19] I. A. Halloun and D. Hestenes, The Initial Knowledge State of College Physics Students, Am. J. Phys. **53**, 1043 (1985).
[20] D. E. Meltzer, The Relationship Between Mathematics Preparation and Conceptual Learning Gains in Physics: A Possible "Hidden Variable" in Diagnostic Pretest Scores, Am. J. Phys. **70**, 1259 (2002).
[21] R. M. Schmidt, Who Maximizes What? A Study of Student Time Allocation, Am. Econ. Rev. **73**, 23 (1983).
[22] W. Admiraal, T. Wubbels, and A. Pilot, College Teaching in Legal Education: Teaching Method, Students' Time-On-Task, and Achievement, Res. Higher Educ. **40**, 687 (1999).
[23] J. Stewart, J. McGee, and G. Stewart, Using Time-On-Task Measurements to Understand an Introductory Science Class, presented at Summer 2006 AAPT meeting., URL http://www.uark.edu/depts/physinfo/phystec/research/summer2006apstalkJCS.pdf
[24] A. B. Champagne, L. E. Klopfer, and J. H. Anderson, Factors Affecting the Learning of Classical Mechanics, Am. J. Phys. **48**, 1074 (1980).
[25] P. M. Sadler and R. H. Tai, Success in Introductory College Physics: The Role of High School Preparation, Sci. Educ. **85**, 111 (2001).
[26] D. P. Maloney, T. L. O'Kuma, C. Hieggelke, and A. V. Huevelen, Surveying Students' Conceptual Knowledge of Electricity and Magnetism, Am. J. Phys. **69**, S12 (2001).
[27] D. Scarl, *How to Solve Problems for Success in Freshman Physics Engineering and Beyond*, 6th ed. (Dosoris Press, Glen Cove, NY, 2003).
[28] W. DeBuvitz, Answering Essay Questions, Phys. Teach. **46**, 165 (2008).