A multi-faceted approach to measuring student understanding

Ian T. Griffin, Kyle J. Louis, Ryan Moyer, Nicholas J. Wright, and Trevor I. Smith

*Department of Physics & Astronomy and Department of STEAM Education,
Rowan University, 201 Mullica Hill Rd., Glassboro, NJ 08028, USA*

Data from the FMCE suggest that there may be inconsistencies in students' understanding of forces when various types of motion are presented. These inconsistencies are highly evident when testing for students' understanding using graphs. In the current study we assume that measurements of student understanding of a particular topic depend both on the student and on the instrument used to make the measurement. Multiple measurements are needed to build a more complete picture of what the student believes to be true. We compare individual student responses to 12 questions from three FMCE question clusters. Using various visualizations of the data including model analysis plots, contingency tables of student responses, and consistency plots, we identify trends that are not evident from traditional normalized-gain-based analyses. Statistical results from the χ^2 test of independence and one-way analysis of variance provide support for our findings.

I. INTRODUCTION

Previous results have shown that student learning on isomorphic questions from the Force and Motion Conceptual Evaluation (FMCE) [1] depend on the details of the question setup. Smith, Wittmann, and Carter found that learning trends differed on the Force Sled question cluster (in which answer choices are presented as descriptions of forces) and the Force Graphs cluster (answers as graphs of force vs. time) [2]; however, the cause of these differences is yet unknown. Do students struggle to interpret graphs of forces as is well documented for graphs of kinematic quantities [3, 4]? How might these results inform instructional strategies?

In order to begin answering these questions, we have identified four questions in each of three clusters — Force Sled (FS), Force Graphs (FG), and Acceleration Graphs (AG) [5] — that present students with nearly identical descriptions of motion (velocities); we use these to define the four cases shown in Table I, each of which contains three questions that are joined by a common type of motion. We focus on individual students' responses to questions within a given case to measure within-student coherence (does an individual give three answers that convey the same understanding?) as well as between-students consistency (do answer patterns match across multiple students?). Identifying matching responses for each of the questions in a case and interpreting each set of responses as corresponding to a particular mental model (Table II) allow us to represent and analyze our data in a variety

TABLE I. Isomorphic question groups from the Force Sled (FS), Force Graphs (FG), and Acceleration Graphs (AG) question clusters.

Case	Described Motion	Question		
		FS	FG	AG
1	moving right, speeding up	1	16	22
2	moving right, steady speed	2	14	26
3	moving right, slowing down	3	18	23
4	moving left, speeding up	4	19	25

TABLE II. Definitions of models consistent with responses to the FMCE. Not all models are evident on all questions. Other models were defined that only relate to Cases 2 and 3.

Model Name	Model Description	Cases
Correct, $F \propto dv/dt$	Consistent with Newton's second law: net force is proportional to the rate of change of velocity.	1–4
Common, $F \propto v$	Net force is proportional to velocity.	1–4
Graph as Picture $F \propto dv/dt $	Graphs can be read as literal pictures of the situation. Similar to the correct model, but ignoring sign/direction.	1–4 3,4
$F \propto v $; Graph Read Left	Net force is proportional to speed; reading the graph in the direction of motion.	4

of ways including model analysis [6], contingency tables [7], and consistency plots [8]. The majority of this paper is devoted to exploring the affordances and constraints of each of these representations. We are particularly interested in documenting cases in which students have a relatively high tendency to exist in a superposition or mixed-model state (selecting incoherent responses within a case) [6], and identifying possible hierarchies in student learning for various questions [7]. To highlight various aspects of our results we present examples from three different schools, two of which were previously analyzed [2]. All courses used research-validated instructional materials. For brevity we present results from Cases 1 & 4; analyses of Cases 2 & 3 show similar trends.

II. CONTINGENCY TABLES

Both within-student coherence and between-students consistency for a question pair may be efficiently represented using a contingency table in which each cell shows the number of students who gave a particular pair of answers to the questions (Table III). We restrict our analyses to examining com-

TABLE III. Contingency tables for Case 1: moving right and speeding up, FS (question 1) vs. FG (question 16). Cohen's w measures the strength of the within-student coherence: $w < 0.1$ is weak, $w \approx 0.3$ is moderate, and $w > 0.5$ is strong.

School 1		pretest		post-test	
$N = 195$		Force Graphs		Force Graphs	
Force Sled	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto v$	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto v$
	$F \propto v$	4	9	39	15
		1	181	17	124
		$w = 0.48$		$w = 0.60$	
School 2		pretest		post-test	
$N = 180$		Force Graphs		Force Graphs	
Force Sled	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto v$	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto v$
	$F \propto v$	8	16	103	5
		10	146	35	37
		$w = 0.31$		$w = 0.54$	
School 3		pretest		post-test	
$N = 340$		Force Graphs		Force Graphs	
Force Sled	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto v$	$F \propto \frac{\Delta v}{\Delta t}$	$F \propto v$
	$F \propto v$	19	23	224	15
		12	286	52	49
		$w = 0.47$		$w = 0.49$	

parisons between FS and FG questions (both asking about force but differing in presentation of answer choices) and between FG and AG questions (asking about different quantities but all answers as graphs) [9]. We chose to only consider answer choices with corresponding options in all clusters and omitted answer choices chosen by fewer than 5% of students at all three schools. Limiting our data sets in these ways allows us to make claims about student coherence without overemphasizing relatively unlikely responses.

Table III shows the pre- and post-test contingency tables from all three schools for Case 1: FS (question 1) vs. FG (question 16). The upper left cell in all tables indicates correct answers on both questions, and on-diagonal cells show students who coherently chose either the correct or most common incorrect answer [10]. Cohen's w provides a measure of this coherence and is equivalent to the correlation coefficient with $w < 0.1$ considered a weak correlation, $w \approx 0.3$ being moderate, and $w > 0.5$ strong [11]. Results indicate that individual student responses are moderately to strongly coherent both before and after instruction and that this relationship either increases (Schools 1 and 2) or stays the same (School 3). Table III also suggests that a hierarchy may exist at Schools 2 and 3 after instruction: students who chose the correct FS answer were very likely to also choose the correct FG answer, but not vice versa [7]. This implies that students may need to understand FG questions before being able to understand FS questions. This trend is not observed at School 1, but these students mostly choose both common incorrect answers after instruction.

(a) School 1

(b) School 2

(c) School 3

FIG. 1. Consistency plots for Case 1: moving right and speeding up, FS (question 1) vs. FG (question 16). Corresponding answer choices are shown in parentheses after the name of the model.

III. CONSISTENCY PLOTS

Contingency tables show us that individual student responses on the FS and FG questions are fairly strongly correlated after instruction, but students may display their learning on the FG before the FS. However, we cannot see how students transition from the pre- to the post-test tables. To represent these transitions we use consistency plots in which circles within the plot represent how students answered on the pretest, triangles represent how students answered on the

FIG. 2. Consistency plot for School 3, Case 4: moving left and speeding up, FS (question 4) vs. FG (question 19).

post-test (forming an “arrow” from the circle), and squares represent students that did not change answers from pre- to post-test [8]. The scaling of the shape sizes, line thickness, and font size make common trends apparent.

We can clearly see in Fig. 1 that the majority of students at School 1 started and ended the course by choosing the most common incorrect answer for both questions, and that the most consistent transition at both Schools 2 and 3 is from the incorrect response on both to the correct answer on both (arrows pointing up and to the left). These plots support the contingency table results that many students are internally coherent as shown by the largest numbers being in either the upper left or lower right cells. We can also see the higher learning trends on the FG questions at Schools 2 and 3 as shown by larger arrows going left than going up from the lower right cell at both schools (30 vs. 5 and 49 vs. 14, respectively).

Consistency plots make many trends more salient. The right column of Fig. 1(b) shows five students each going up and down; this cyclic transition shows up as a zero net change on the contingency tables. Moreover, consistency plots help us visually interpret rich data sets from more complex cases like the plot for School 3, Case 4 (moving left, speeding up) shown in Fig. 2. Perhaps the most salient feature of this plot is the number of different transitions that students make from pre- to postinstruction. It is immediately clear that students who start the semester giving a particular pair of answers will not necessarily give the same answers as their peers after instruction: “beginning state” + “instruction” $\neq$ “ending state”. Another aspect of Fig. 2 is that the strongest attractor is choosing the correct answers for both questions with most students ending the semester in the upper left corner after starting in many different cells.

Two of the most interesting features of Fig. 2 occur in the right two columns. The right-most column acts as an attractor, with several students ending the semester by choosing this answer, while no one started there. The second-right-most column, bottom row shows what Wittmann and Black call a “starburst,” with a strong majority of students leaving that response (in various directions), and no one coming in [8]. Attractors and starbursts may also be seen by comparing

pre/post contingency tables, but the tables cannot show the variety of transitions students make from pre to post.

The incorrect attractor and starburst suggest that some incorrect responses may be considered more or less sophisticated than others, which is consistent with previous results [12]. Answer “H” on FG question 19 may be considered the most naïve with most students abandoning this choice, and answer “A” may be considered the most sophisticated with students only choosing this after instruction. In fact, “A” is consistent with choosing a graph of the magnitude of the correct force but ignoring the negative sign of the force, consistent with documented graphing difficulties [3]. Choice “H” is consistent with the idea that the net force is proportional to speed, but (because the object is moving to the left) reading the graph from right to left. This matches a common incorrect conceptual reasoning, but pairs with a strong misunderstanding of how to read graphs, which supports the notion that “H” is a very naïve option.

IV. QUANTITATIVE COMPARISONS BETWEEN SCHOOLS

Based on the contingency tables (Table III) and consistency plots (Fig. 1) for Case 1, it appears as though students at Schools 2 and 3 show similar trends, and that students at School 1 learned significantly less overall. We choose three different methods to quantitatively compare the schools and confirm (or refute) these claims: statistical analyses of individual students’ normalized gains, model analysis with error bars derived from the standard error, and a χ^2 test of independence based on consistency plot transitions.

Using a one-way ANOVA to compare students’ normalized gains allows us to check for statistically significant differences between schools, and calculating Cohen’s d provides an interpretation of the size of any differences. Table IV shows the results of statistical analyses for learning gains on the entire FMCE, the 12 questions included in Cases 1–4, and Case 1 alone. Not all data sets include the energy questions, thus we have excluded them from our analyses. Our threshold

TABLE IV. Statistical Results: Average normalized gains from Schools 1–3 as well as main effect p -values and pairwise effect size and statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Average g (%)			ANOVA Main Effect (p)	Cohen's d		
	S1	S2	S3		1v2	1v3	2v3
Full Test	28.7	59.7	69.1	< 0.001	0.9***	1.3***	0.3**
Cases 1–4	22.9	61.1	71.4	< 0.001	0.9***	1.1***	0.2*
Case 1	30.8	65.3	77.3	< 0.001	0.7***	1.1***	0.3**

FIG. 3. Case 1 model plot. Error bars derived from the standard error of the response distributions [2, 13].

for significance is $p < 0.05$, and we used Tukey's HSD for *post hoc* pairwise comparisons. Table IV shows that all three schools are statistically significantly different with School 3 showing the highest learning gains and School 1 the lowest; however, the difference between Schools 2 and 3 show only small effect sizes. This is somewhat consistent with our results that Schools 2 and 3 appear to be similar on contingency tables and consistency plots.

Using the definitions in Table II one may use model analy-

sis to compare the schools. Figure 3 shows the Case 1 model plot for the correct and most common incorrect models with error bars derived from the standard error of the response distributions [13]. As with the ANOVA results, all three schools are shown to be statistically different, with no overlapping error bars; however, both Schools 2 and 3 end the semester firmly in the correct-model region, with School 1 staying in the incorrect region. This reflects the effect sizes in Table IV.

Finally, we created a data set for each school based on the number of students making each transition on the consistency plots and compared the schools using the χ^2 test of independence (threshold at $p < 0.05$), using Fisher's exact test and Bonferroni's correction for *post hoc* pairwise comparisons ($p < 0.017$). For all plots in all four cases, comparisons between Schools 2 and 3 had weak to moderate effect sizes and were not found to be statistically different (e.g., $p = 0.20$ for Case 1). Comparisons between School 1 and either School 2 or School 3 showed large effect sizes and statistically significant differences ($w > 0.5$; $p < 0.001$). These results support our original interpretations of the contingency tables and consistency plots and suggest that similarities exist between Schools 2 and 3.

V. SUMMARY

As we have shown, different approaches to analyzing our data reveal various similarities and differences between the three schools. Model analysis explicitly treats students as having a certain probability of existing in a superposition state of mental models that can be measured differently based on the question being asked and the answer choices provided. Using contingency tables and consistency plots with appropriate statistical tests allows us to measure the extent to which a class's responses to two questions are correlated and show how these responses change over time. All of these are useful (and may be necessary) for gaining a more complete picture of the effects of instruction. Future work will include synthesizing results across multiple cases and incorporating all three questions within each case.

[1] R. K. Thornton and D. R. Sokoloff, *Am. J. Phys.* **66**, 338 (1998).
[2] T. I. Smith, M. C. Wittmann, and T. Carter, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020102 (2014).
[3] L. C. McDermott, M. L. Rosenquist, and E. H. van Zee, *Am. J. Phys.* **55**, 503 (1987).
[4] R. J. Beichner, *Am. J. Phys.* **62**, 750 (1994).
[5] T. I. Smith and M. C. Wittmann, *Phys. Rev. ST Phys. Educ. Res.* **4**, 020101 (2008).
[6] L. Bao and E. F. Redish, *Phys. Rev. ST Phys. Educ. Res.* **2**, 010103 (2006).

[7] R. Rosenblatt and A. F. Heckler, *Phys. Rev. ST Phys. Educ. Res.* **7**, 020112 (2011).
[8] M. C. Wittmann and K. E. Black, *Phys. Rev. ST Phys. Educ. Res.* **10**, 010114 (2014).
[9] We omit FS and AG question comparisons because these differ both in the quantity examined and the answer presentation.
[10] Case 1 is shown as an example due to its simplicity; Cases 2–4 include additional rows and columns.
[11] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, 1988).
[12] R. K. Thornton, *AIP Conf. Proc.* **399**, 241 (1997).
[13] T. I. Smith, *Phys. Rev. Phys. Educ. Res.* **12**, 023102 (2016).