Probing Students Misconceptions results from Concept Inventory and Their Understanding in Science Learning

Horace Crogman[1,2]
Raul Peters[3]
Maryam TrebeauCrogman[4]

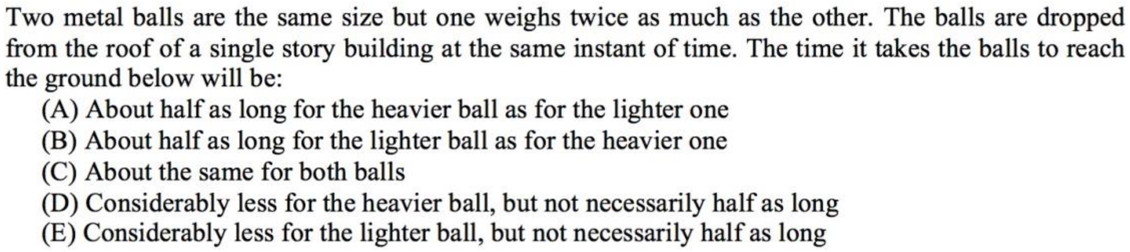
[1] hcrogman@gmail.com
Department of Physics, San Bernardino Valley College
CA, 92410, United States
[2]Research and Development, The Institute of Effective Thinking, Riverside, CA, United States
[3] Department of Physics, Paine College, 1235 15th St. Augusta, GA, United States
[4] Department of Psychology, California State University Stanislaus, Turlock, CA 95382, United States

(Received: 17. 08. 2018, Accepted: 17. 10. 2018)

Abstract
Concept inventory (CI) tests are typically used to measure students' general knowledge before and after instruction. We find issue with the current format of these tests, which some researchers claim, measure students' misconceptions, since the answers choices given to students do not consider their prior knowledge. We particularly analyze Force Concept Inventory (FCI) tests results to reflect on what CI tests are measuring in general. Also, students' choices on FCIs are more associated with their natural experiences than their knowledge of the Newtonian signals. Thus, we propose some modifications to the FCI format and show how this change helps to parse out what of students' answers are simple misunderstanding or true misconceptions to focus on building instruction. Results show that: 1/ Concepts are very disorganized in students' minds, 2/ despite some improvement at post-test, students' choices from pre-test to post-test do not stay consistent, 3/ modifying the test helped come up with clearer explanations about students' choices. We found that very little work has been done to assess and rethink FCIs in the past few decades. Our new proposed design opens doors to fairer and more organic testing/assessment practices in college STEM.

Keywords: Misconception, concept inventory, science learning.

INTRODUCTION

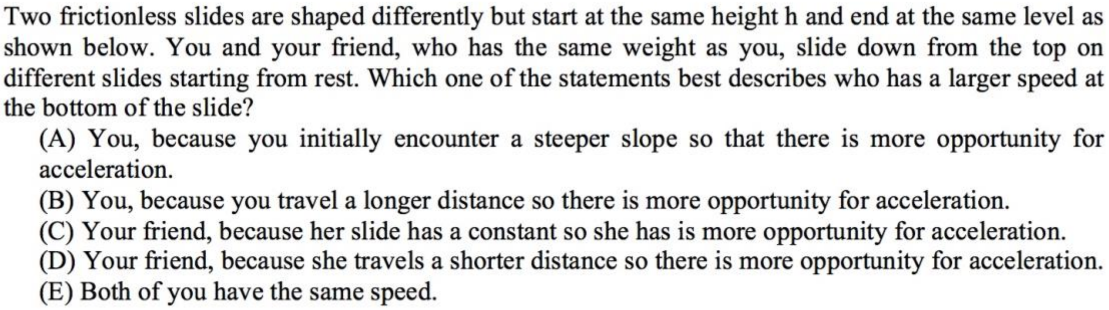

Two metal balls are the same size but one weighs twice as much as the other. The balls are dropped from the roof of a single story building at the same instant of time. The time it takes the balls to reach the ground below will be:

(A) About half as long for the heavier ball as for the lighter one
(B) About half as long for the lighter ball as for the heavier one
(C) About the same for both balls
(D) Considerably less for the heavier ball, but not necessarily half as long
(E) Considerably less for the lighter ball, but not necessarily half as long

Figure 1. *FCI question (Q1 of FCI in our study)*

This type of question is typically found in Physics science books as well as in *Force Concept Inventory* (FCI) tests, which are inventories given to assess students' general concepts understanding at the start of Physics subject instruction (Thornton & Sokoloff, 1998). The expected correct answer to this question is C. It has been reported by many educators that students often get this question incorrect, and are more likely to choose answer A or D. These choices are labeled as misconceptions since the assumption is that FCI measures students' misconceptions. Is option A or D really incorrect for the particular question stated? Is any answer correct at all?

Similarly,



Two frictionless slides are shaped differently but start at the same height h and end at the same level as shown below. You and your friend, who has the same weight as you, slide down from the top on different slides starting from rest. Which one of the statements best describes who has a larger speed at the bottom of the slide?

(A) You, because you initially encounter a steeper slope so that there is more opportunity for acceleration.
(B) You, because you travel a longer distance so there is more opportunity for acceleration.
(C) Your friend, because her slide has a constant so she has is more opportunity for acceleration.
(D) Your friend, because she travels a shorter distance so there is more opportunity for acceleration.
(E) Both of you have the same speed.

Figure 2. *An Energy-Momentum Concept Inventory (EMCI) question.*

This is a question we use as a free response in the classroom (also found in EMCI tests) where most students tend to choose E as their answer. When followed up with the question "Between you and your friend who gets to the bottom first", most students answer that they will get there at the same time. Why? Is it because of some pre-held beliefs or is that their experience is limited with respect to such an event, a lack of knowledge about the effect from the different paths taken? As such, students guess answers that make sense to their experience, i.e. since starting from the same height and weight, students reasoned that if the paths failed to change the speed at the bottom, then their time would be the same. Further, are students' assumptions unreasonable or unscientific?

Learning is the result of experience. It involves an interaction between what Crogman (2017) terms L-Language (learners' language - which cultural impacts should be accounted for), and the I-Language (teachers' language which is formal). In simple terms, learning results from the learner's comprehending the instructors' language (Crogman, 2017; Humphreys-Jones, 1986). Every scientific principle arises because of our experience. Hume (1748) argues that all ideas result from experience and not spontaneously spring into existence, which is essentially an argument of empiricism. For example, can one have an idea about waterfalls if they have no notion about water and how objects behave in the earth's gravity (i.e. 'I must be aware of gravity')? Kant, on the other hand, seems to propose an argument of Rationalism; this is the view that knowledge derives from reason without the aid of the senses (Schunk, 1991), that is, knowledge arises through the mind. We'll agree with Hume here and define learning as the acquisition of knowledge or skills through experience (Hume, 1748; Kolb, 1984, 1985; Lewin, 1951), observation (Chiesa, 1992, 1994; Hall, 2003; Skinner, 1938), change in behavior (Tolman & Honzik, 1930), or instruction (Atkinson, Atkinson, Smith, & Bem, 1993). This involves the learner, the environment, and instructional method and tools.

Experience helps build prior knowledge, which we rely on to generate new ideas and create more knowledge. In physics, or any other subject, every student begins with a well-established system of commonsense beliefs about how the physical world works derived from years of personal experience. This is a primitive form of doing science. When these experiences do not match scientific truth, we say that the students have a misconception or misunderstanding. Demirci (2005) cited a decade of research that suggests that many students enter physics classes with preconceived ideas feeding their misconceptions. This causes students to incorrectly describe the physical world that is consistent with the laws of physics. For instance, the perception that lighter objects tend to float more than heavier ones in fluids.

Educators designed concept inventory (CIs) tests as a measure of students' misconceptions (Madsen, McKagan, and Sayre, 2017; Hestenes, Wells, and Swackhamer, 1992). For example, Hestenes, et al. (1992) designed a CI to assess students' beliefs about force and Newton's laws. Madsen, et al. (2017) argue that CIs are a unique form of assessment speaking to students' unique language and containing students' ideas. They further articulate that students need a sophisticated understanding of the concept(s) to do well on these tests. The inclusion of students' everyday ideas and natural language in the multiple-choice format forces them to select choices that reflect their natural experience, resulting in confusing the learner in their language (Crogman, 2017); that is, when the I-language fails to understand the L-language, misunderstanding occurs. CIs typically contain no choices allowing students to indicate their lack of knowledge or understanding. This results in them guessing, and CIs have nothing allowing to see when this actually happens. Therefore, CIs fall short of addressing the central question of students' misconceptions.

Is a teacher-centered approach enough to overcome students' misconceptions? The literature shows that the general method of teaching most introductory classes is often the traditional didactic pedagogy, in which the lecturer is an expert transmitting knowledge. For example, some researchers have argued that traditional teaching is not effective enough to modify commonsense misconceptions in relation to Newton's laws and motion (Fadaei & Mora, 2015). Students in such a teacher-centered setting may maintain pre-held beliefs difficult to change because instruction only simplifies and minimizes the process through which these beliefs were derived (Crouch & Mazur, 2001). All students' beliefs come from their cultural tradition and conceptual conflicts encountered in class. We suggest that class experiences that create conceptual conflicts cause students to naturally shift their position. This happens through sensory stimuli, which comes through Socratic question asking processes or classroom demonstration and experimentation (Crogman, TrebeauCrogman, Warner, Mustafa, & Peters, 2015; Crogman & TrebeauCrogman, 2016, 2018).

The main question of our essay is to know if there is clear empirical evidence that CIs measure students' understanding of concepts, and study how students' prior knowledge has been considered in the construction of these tests. Based on that model, researchers argued that students' wrong answers in FCIs pre to post-instruction is a measure of their misconceptions (Fadaei & Mora, 2015). Bruun and Brewe (2013), with the same tool, propose that CIs can predict the trajectory of each student's success in the classroom and beyond. It is our argument that when CI predicts a negative student's outcome beyond the classroom then instruction has failed to clarify students' misunderstanding or confront their conflicts. Instruction must be designed to challenge the thinking and prepare them to be better critical thinkers (Crogman &Trebeau, 2016; Crogman, 2018).

Conversely, Huffman and Heller (1992) questioned whether FCIs do actually measure students' misconceptions as what Hestenes et al. (1992) claimed. They observed that items on the FCI are loosely related to each other and that students' understanding of concepts are vague and undifferentiated. Hestenes and Halloun (1995) have made a very strong case against the objections of Huffman and Heller, and the literature overwhelmingly has sided with them arguing for FCIs as a true measure of students' understanding (Fadaei and Mora, 2013; Hake, 1994; Madsen, et al., 2017). Griffith (1997) concludes that, "good students might thereby be misled into making the wrong choice", fearing also that FCIs may force teachers to teach to the test, leading to artificial high FCI scores.

Crogman and Trebeau (2018) join Griffith, and Huffman and Heller in arguing that FCIs can be versatile tools and are better suited to guide teachers' prep than designed to measure misconception. For example, the pre-FCI results are a good classroom tool to build groups to work on concepts throughout instruction (Crogman et al., 2015). FCIs were also used to evaluate the successfulness of peer instruction (Crouch & Mazur, 2001).

Hestenes (1998) argues that the most important feature of the FCI is that it sets a *minimal standard* for effectiveness of instruction in Newtonian mechanics and separates scientific concepts from commonsense physical knowledge. Like Griffiths and others, we argue that FCIs should not be used as a minimal standard in a first course, but as a guide in how to prep instruction. Nonetheless, since Huffman and Heller's (1992), and Griffith' (1997) objections, very little work has been done to clarify what FCIs do and do not actually measure. We are proposing that CIs allow instructors to evaluate the effectiveness of their teaching over time, and across instructors and institutions, instead of measuring misconceptions to determine how students will perform in class. We detail our perspective next by analyzing student's response patterns and proposing new interpretations about CIs' purpose and results.

METHOD

Participants

Participants were recruited from four Fall and Summer Introductory Physics classes.

Table 1. *Demographics of the participants*

	Min	Max
GPA	1.8	4.8
AGE	15	49
Income	<$10,000/year	>$80,000/year
Gender distribution	Male	32
	Female	48

Majors		*Ethnicity*	
Physics	2	Hispanic/Latino	38
Engineering, Computer science, Mathematics	13	East and West Asian	17
Biology, Chemistry, Biochemistry	18	Black	4
Pharmacy, Med, Pre-med, Nursing, Occupational Therapy, Kinesiology, Radiology, health science	34	White	6
Law, Business, Accounting	4	Arabic, Middle Eastern	2
Trade, Agriculture, mechanics	2	Pacific Islander	1
Liberal arts, fine arts, Psychology, Letter, History	6	Multiple Races	11
Undecided	3	Other	3

Study Design

Two community college Physics classes' pre and post-FCI answers are examined. We had 98 participants, 48 reported have never taken physics, and 25 reported taking it in high school. Seventy-nine students took the pre-test, and 87 the post-test. Further, 18 students were missing biographical data.

FCIs are typically built with a large list of multiple-choice questions covering a wide range of concepts to be studied during subsequent instruction. The scores obtained are assumed to reflect their true understanding of each concept assessed. Based on our argumentation, we consider that FCIs are not built to truly reflect students' experiences and background. Thus, we proposed a new type of FCI as follows: a question is asked on a number of general concepts; students must pick an answer among multiple choices. Then the question with its answer is followed by a list of possibly corresponding concepts related to the question at hand. Students then must also pick a matching concept. Thus, students are tested on their understanding of a question as well as their authentic comprehension of the concept attached to it. Table 2 shows the possible combinations that can arise from matching correctly or incorrectly each question and concept answers (examples in appendix).

Table 2. *New distinction to be taken in account between answer types when building FCI tests*

		Answer		**Concept**			**Student's Knowledge State**	
		Corr.	**Incorr.**	**Corr.**	**Incorr.**	**IDK**	**Pre**	**Post**
Answers Scenarios	**A**	X		X			CU	CU
	B		X		X		LPK	MU/LPK/MC
	C	X			X		G/MC	G/MC
	D		X	X			MC/G	MC/G
	E	X				X	LPK/G	G/LPK
	F		X			X	LPK	MU/LPK

LPK- Lack of Prior Knowledge; MC- Misconception; CU- Conceptual Understanding; G-Guessing; IDK- I Do Not Know; MU-misunderstanding

FCIs were administered on a computer in class, at the start and the end of each semester. Students also filled out questionnaires covering Physics knowledge background, age, gender, major, future career plan, SES and family information, prior history of learning disability, and preferred type of activities while young. These were intended to differentiate participants inclined to understand and have experienced the behavior of objects in space, and natural inclination for STEM.

Computation Modeling and Statistics

FCI Items Clarity. Prior to our study analyses, we carried out a factor analysis (FA) to see what FCI items actually measure based on our concerns.

We selected a much larger (n=200) set of FCI score of both community colleges. The analysis was conducted on students' answer patterns. Provided the inventory actually measures the right conceptual dimensions, then under FA, the items that measure each of these dimensions should fairly well cluster together under clear factors; however, if items do not cluster well, then conclusions about what these items measure may not be straightforward anymore.

Scores analyses. For an optimized understanding of the scores obtained, we calculated gain scores between pre and post instruction scores. This shows how students have evolved after formal instruction and modified their understanding of prior natural experience. The *differences* in scores between pre- and post-test (e.g. raw gain, normalized gain or effect size) are used to determine the effectiveness of teaching (Madsen, et al., 2017), which is at the heart of our hypothesis of what FCIs do measure. Additionally, FCIs (normalized) gains are also calculated as the ratio between the actual change and the greatest possible change in one's score for both groups:

$$G = \frac{< \%postscore > - < \%pre\ score >}{100 - < \%prescore >}$$

The FCIs' effective size is calculated as the ratio between the actual change and pooled standard deviation ($stdev$) for both groups:

$$d = \frac{< postscore > - < pre\ score >}{stdev}$$

Data analyses were conducted using SPSS, Version 24 (IBM® SPSS® Statistics, 1989). Primary analyses were conducted by extracting general means and using ANOVAs to compare groups' means and pre/post-test trends on the basis of having taken physics prior to this class. Secondary analyses were conducted based on Group/school membership, gender, and SES.

RESULTS

Factor Analysis

The following tables (Table 3 & 4) detail the factor loadings resulting from our FCI answers FA analysis. The loadings patterns show 11 factors (ignoring negative factor loadings). Variances were no higher than 11% suggesting that choices were scattered and unorganized in students' minds, and that not much of the clustering of item questions makes sense under our 11 factors (see discussion for interpretation).

Table 3. *Conceptual dimensions proposed by FCI measures compared to what factor analysis reveals*

Conceptual Dimension	*Item Number*	*Factor Analysis Matched*
Kinematics	7, 20,21,23,24,25	none
Newton 1st Law	4,6,8,10,18,26,27,28	1,2,4
Newton 2nd Law	6,7,24,25	2
Newton 3rd Law	2,11,13,14	1,3
Super position	9, 18.19, 28	none
Kind of Force	1,3, 5,12,16,17,18,22,23	1,2,4

Table 4. Factorial analysis of FCI clustering under general concepts

Item Number	Factor Loadings	Conceptual Dimension	Problem Situation
FACTOR 1 (11% of variance)			
10	0.333	Newton 1st Law	Hockey puck sliding at constant speed
17	0.667	Kind of Force	Elevator lifted up at constant speed
22	0.619	Kind of Force	Path of rocket drifting sideways in space
23	0.615	Kind of Force	Path of rocket drifting sideways in space
28*	-0.398	Newton 3rd Law	Student pushing off a chair
FACTOR 2 (7% of variance)			
5	0.488	Kind of Force	Ball moving at high speed in a circular channel
18	0.662	Kind of Force	Boy swinging on a swing
25	0.326	Newton 2nd Law	Constant Force on a box
26*	-0.379	Newton 2nd Law	Constant Force on a box
30	0.370	Newton 2nd Law	Hitting ball against the wind
FACTOR 3 (6% of variance)			
10*	-0.344	Newton 1st Law	Hockey puck sliding at constant speed
14	0.643	Newton 3rd Law	Path of falling bowling ball
21	0.426	Kinematics	Path of rocket drifting sideways in space
24*	-0.344	Kinematics	Path of rocket drifting sideways in space
27	0.358	Newton 1st Law	Constant Force on a box
FACTOR 4 (5% of variance)			
4*	-0.546	Newton 1st Law	Balls falling
15	0.632	Kind of Force	Car pushing truck
16	0.437	Kind of Force	Car pushing truck

FCI

Scores Analyses

Of a total of 98 students who turned in information, 64 completed both pre and post-FCI tests (match data). There was a strong positive correlation between pre and post-FCIs, which was statistically significant (*r*=.458, n=165, *p*=.01). Performance improved from pre-FCI (n=64, M=8.38, SD=5.61), to post-FCI (n=64, M=15.43, SD=7.08). There was also a statistically significant difference between pre and post test scores (F(1,165)=43.387, *p* <.001). There was a normalized gain reflecting a medium gain, whereas the effective size suggests a large gain (Table 5). Note that there was a retention rate in the classes of about 95%.

***Table 5**. Changes represented from the FCI match data*

Raw gain	7
Normalized gain	0.32
Effective size	0.99

Scores spread about the mean for each PRE (SD: 5.61) and POST (SD: 7.08) tests showing that the data is widely distributed. On average students' scores improved of about 7 points at post-test albeit not by very large margins. Additionally, students' performance was about the same whether or not they took physics prior; there was no statistically significant difference in FCI means between students who did and did not take physics prior, at both PRE (*p*>.001) and POST (*p*>.001) instruction tests. The same was true for SES, and gender, but there was a significant difference between the two institutions at Pre-test (*p*=.002), difference which disappeared at post-test.

In the scope of our intent to analyze closely students' response and changes, below are represented the changes from pre to post on questions on which students had right answers (Figure 3), and also the changes in wrong answers (Figure 4). Note that in majority, over the 30 questions students improved their right answers and decreased in wrong answers across the different concepts. Depending on the type of questions the difference was wider. This is a good tool for instructors to assess which questions were harder to understand and could contribute to misunderstanding.

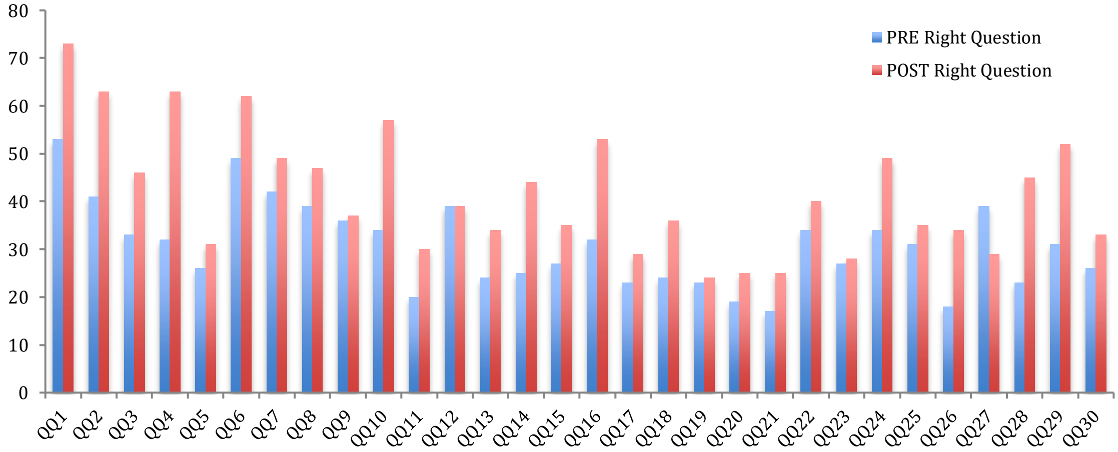


Figure 3. *Changes in frequency of Right answers from Pre to Post instruction*

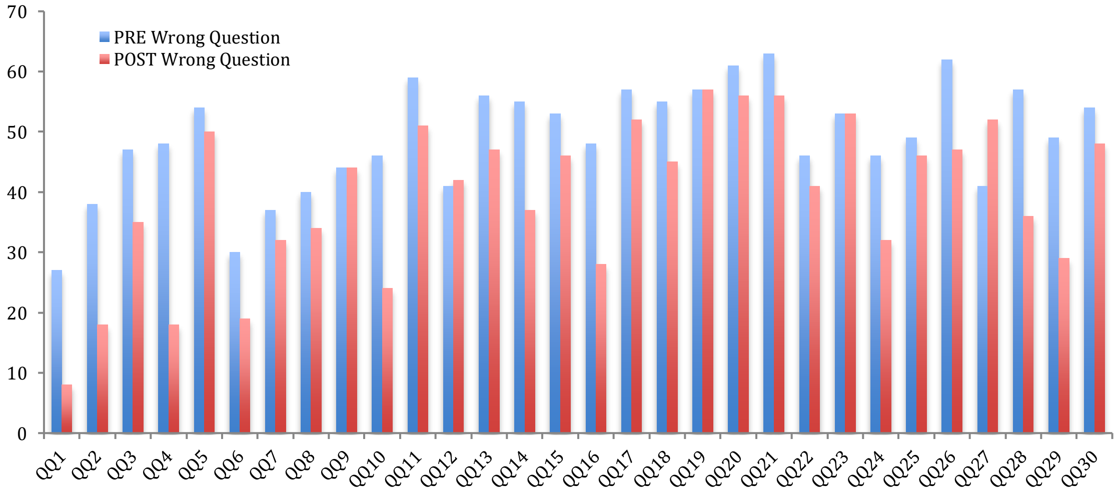


Figure 4. Changes in frequency of Wrong answers from Pre to Post instruction

Our discussion on FCIs started with students' answers on Question 1 of our modified FCI. Tables 6 and 7 detail also three additional questions of similar concept, to help understand students' choices based on their different matching patterns in various categories of the same question. Both tables illustrate the 6 different matching choices students have which we coded (see columns). We see a general tendency to choose the RA/RC match both at pre and post-FCIs for these four questions, which is interesting given that students may or may not have had prior experience with these concepts. We see an increase in this choice from pre to post-test. We studied the pattern in Table 7 this time separating students by whether they formally had had exposure to physics teaching, and we found the same tendencies.

Table 6. Total PRE and POST answers per match type per question.

Questions	*FCI Test*	*Wrong answer Wrong Concept (WA/WC)*	*Right answer Wrong concept (RA/WC)*	*Wrong answer Right Concept (WA/RC)*	*Right answer Right concept (RA/RC)*	*Right answer IDK concept (RA/IDK)*	*Wrong answer IDK concept (WA/IDK)*	*TOTAL*
Q1	PRE	12	3	11	46	5	3	80
	POST	2	5	5	65	3	0	80
Q2	PRE	7	11	25	23	6	8	80
	POST	4	12	14	44	6	0	80
Q3	PRE	6	2	37	31	1	3	80
	POST	5	7	26	39	2	1	80
Q13	PRE	12	3	30	18	3	14	80
	POST	14	8	30	23	3	2	80

***Table 7**. Answers match types by questions by physics background.*

Taken Phys.	Questions	FCI Test	WA/WC	RA/WC	WA/RC	RA/RC	RA/IDK	WA/IDK	TOTAL
Y	Q1	PRE	4	2	5	13	1	0	25
		POST	0	1	0	18	1	0	20
N		PRE	6	1	6	28	4	3	48
		POST	2	3	3	39	2	0	49
Y	Q2	PRE	2	4	8	8	1	2	25
		POST	0	1	4	14	2	0	21
N		PRE	5	5	15	13	4	6	48
		POST	3	8	9	25	4	0	49
Y	Q3	PRE	3	0	13	9	0	0	25
		POST	1	1	7	10	1	0	20
N		PRE	2	2	22	19	0	3	48
		POST	3	4	17	24	1	0	49
Y	Q13	PRE	5	0	11	6	1	2	25
		POST	5	2	7	5	1	0	20
N		PRE	6	2	15	12	1	12	48
		POST	9	2	21	15	1	1	49

Thus, we have seen a change from choosing wrong answers and making wrong answer/concept matches to choosing the right match. However, the question remained, where exactly do students tend to move, in their matches, from pre to posttest because some might go from WA/WC to RA/WC still and not necessarily to the RA/RC column. We believe that understanding these patterns can shed light on our misconception vs. misunderstanding debate.

For this we have illustrated the data both in Table 6 and in Figures 5 to 9 showing how students' choices moved across the 6 matching choices they had. Figure 5 shows a clear drop in the "I don't know" choice from pre to post instruction, which suggests that students felt more confident about their knowledge. We also see an interaction between pre and post-FCI on the total of right vs. wrong matches overall showing that students made the right match more often at post-test. This may indicate that instruction has shifted conceptual understanding and that misconceptions were in general reduced (see discussion).

Table 8. Total count of answers selections types in the 6 categories.

	PRE		POST	
	# of answers	# of students per categories	# of answers	# of students per categories
WA/WC	592	20	566	19
RA/WC	293	10	451	15
WA/RC	479	16	509	17
RA/RC	467	16	619	21
RA/IDK	172	6	168	6
WA/IDK	398	13	117	4

Note: WA: Wrong Answer, WC: Wrong Concept, RA: Right Answer, RC: Right Concept, IDK: I Don't Know

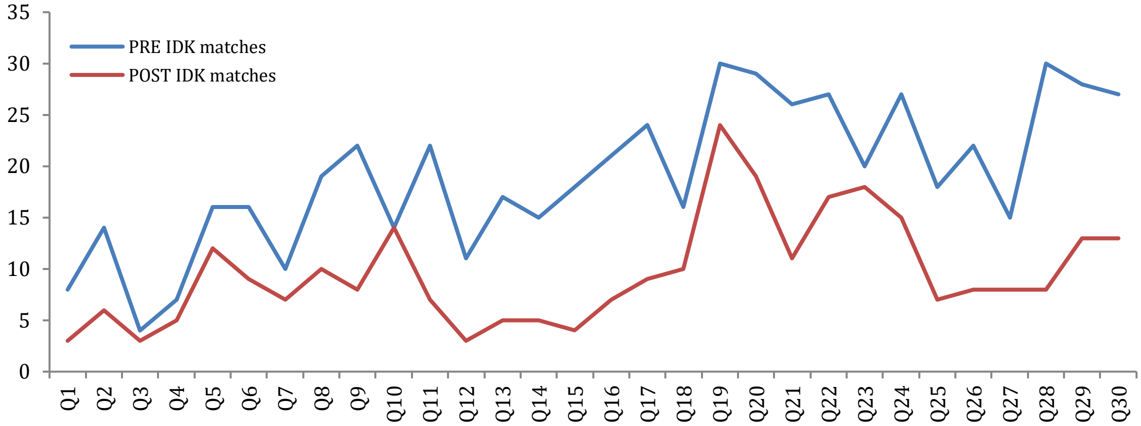


Figure 5. *Decrease of IDK and possibly guessing responses of students on each FCI questions between pre-FCI (blue) and post-FCI (red).*

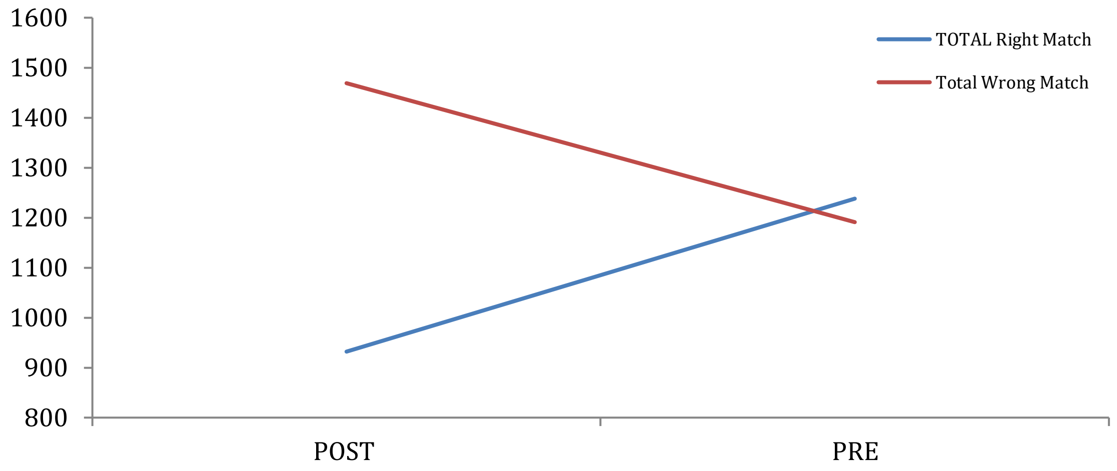


Figure 6. *PRE to POST instruction RA/RC matching.*

DISCUSSION

Factor Analyses and FCIs Dimensions

We performed an FA on more than 200 students' FCI answers. Cluster loading results seem to put in question the way these tests are built and what FCIs actually measure for introductory physics students. Our findings echo that of Huffman and Heller (1995) who also showed that the FCI items they analyzed were only loosely related. Such analyses have not been done in enough volume to push educators and researchers to question FCIs. Yet, FCIs are still widely used in academia. That being said, we do not consider that this result shows FCIs failure to test its six concepts dimensions, but we argue that they do not account for students' prior knowledge, and thus, are fundamentally flawed in both their construction and scores interpretation by educators and researchers.

The six conceptual dimensions proposed by Hestenes, et al. (1992) are logical categories, but are viewed very differently by students, especially those who have never taken physics prior to taking this test. For example, Huffman and Heller showed that students grouped items #6 and #7 together on one factor, while items #24 and #25 grouped together on a different factor; yet, these four items are testing the same concept. Table 4 shows similar findings. Thus, in response to Huffman and Heller, Hestenes and Halloun (1995) raised the issue that their Newtonian signals had many false positives, and a factor analysis on the group of students scoring 60% and above would cluster students' choices into the FCI's six conceptual dimensions. It is for this reason that some instructors conclude that FCIs are good predictors of students' performance. In a similar way, Fadaei and Mora (2015) created eight conceptual dimensions but the problem faced was that students did not categorize their questions as expected. This can be seen also in the data gathered in our modified FCI test, where students were selecting concepts. Although most students got Q1 correctly they failed to select the right match for Q2, Q3 and Q13, and failed to do so at the same high level of correct selection even though they all treat the same concept. Crogman (2017) has argued that the way in which students formulate language and interact with instructors' language will impact their answers selections. Additionally, both Huffman and Hestenes' findings suggest that students' beliefs about physics are loosely organized, incoherent, ill-defined and contextdependent. These are important aspects to consider when revamping FCIs.

Further, we contend that the question, *how much students are doing*, is not addressed by any prior research. Again, in our factor analysis, the items grouped onto a wide variety of ambiguous factors that did not seem to cluster on any clear concepts. This failure suggests that students do not have the prior knowledge needed, so the probability of them choosing answers would naturally be scattered. Based on our findings, we claim that FCIs do not measure what is claimed, corroborating Huffman and Heller's (1992) position.

Modifying the FCI to Distinguish Misconceptions vs. Misunderstandings

More than 65% of students never took physics before, thus we can assume that a majority's prior physics knowledge was constructed from natural experience. For that reason, there is no clear way to claim that pre-FCI tests are a measure of students' misconceptions since the students' knowledge state or abilities is unknown at pre-test. We can argue that there are levels of misconceptions for those who took physics and performed poorly. Yet, we must also account for the amount of time passed between that formal experience and the pre-FCI test. Thus, their performance may also largely rely on memory, which can be subject to a Dunning-Kruger Effect (DKE) (Kruger & Dunning, 1999).

Looking at the typical FCI format options that students choose from, we see that when unsure, they are forced to choose one of five answers anyway, which results in revealing something untrue about the student's knowledge. Adding an "*I Don't Know*" option to the possible answers allows clarifying students' guesses from what they an aware that they do not know, and also what they actually remember. In Figure 6, a larger amount of the students selected "I don't know" at pre-test, which allows us not to assign them erroneous knowledge bases. Further, the DKE might be at play where student is guessing trying to avoid any judgment about not knowing, which can be masked as misconceptions. From Table 6, 44 students were choosing RA/RC at pre-test compared to 61 at post-test, which indicates that instruction helped students learn. Further this improvement is shown on most of the FCI questions, which corresponds to the high effective gain in the data. Therefore, the claim that FCIs can predict how students will do in a class, as they stand in their current form, is false, on the basis that: 1/ it does not account for prior knowledge by affording students with escape options like "I Don't Know", 2/ it does not account for the power of good instruction which can change the course of students' knowledge, behavior and progress over time.

Looking at Q1 in our FCI, improvement increased of 35% (65 students answered correctly vs. 45 at pre-test). However, improvement was more modest on Q2, Q3, and Q13, which cover the same concept. In other groups of questions testing the same concepts, students' improvement was spread. This suggests that FCIs do not measure what the literature claims. Further, Q3 is interesting to explore because it shows no change in WA/RC from pre -test to post-test. This trend was reflected in several other FCI questions. Another observation about the WA/RC match shows an increase on some questions at post-test, yet overall, the WA/RC count is small. This may suggest that after instruction, some students held on to their reasoning. From table 5 this was true whether they took physics prior or not. From table 4 we have the same reflection due to a misunderstanding of the concept, which results in a misconception. From table 6 there are 1469 WA and 932 RA at pre-test, and 1192 WA and 1238 RA at post-test, which shows improvement in the choice of right concepts. Figure 5 illustrates what happened from pre to pot-test. The data suggests that students' choices are less likely due to misconception but rather to other choice issues at play (lack of knowledge, memory, SES background and time to study…/…). This conclusion is reflected in students' choices from pre to pot-test. Although, instruction allows students to be motivated and more confident in their choices, students' pre-held beliefs are difficult to change because classroom instruction is often illustrating special cases remote from what happens in real life experiences.

Figure 5 shows a decrease from pre to pot-test of IDK choices on concepts. As we said, getting a question incorrect does not automatically mean that it is due to a misconception. It could be just a conceptual misunderstanding as it was not even a belief held before the class. Figure 6 showed that the number of correct answers pre to pot-test increased. This configuration allows us to better see where the misconceptions are. Unfortunately, the opportunity to correct these conceptions is missed since the post-test is often given at the end of the semester.

Further, our results indicated that, despite their general improvement, students who tended to score low stayed low, and those who scored high stayed high, this might explain the large standard deviation spread in the data. However, there are pre to post-test improvements of students who scored at the mid order. This could be a reflection on memory retention since students were not alerted ahead of time about the post-FCI, and the tests were given at the beginning and end of the semester.

FCI Choice Patterns and Meaning for Misconceptions

Figures 7 thru 10 show students' selection from pre to post-test on Q1, Q2, Q3, and Q13. In Q1, at post-test, most students stuck to the answers they gave at pre-test as shown in Figure 7. Further, of 46 students that selected RA/RC only 3 changed their answers (8 did not take the post-test). Seven of 12 students who selected WA/WC changed their answers at post-test to RA/RC; 1 student to RA/WC (3 students were missing). However, only 1 student kept their answers from pre-test. When students hold on to their incorrect position, there is likely misconception. We could not say that a student that changed their answer from RA/RC to WA/WC had a misconception, but instead we would see it as a misunderstanding that may have occurred through instruction. Whereas, a change from RA/RC to WA/RC is likely due to an ambiguity of Q1 where the question did not stipulate that air resistance was being ignored for example.

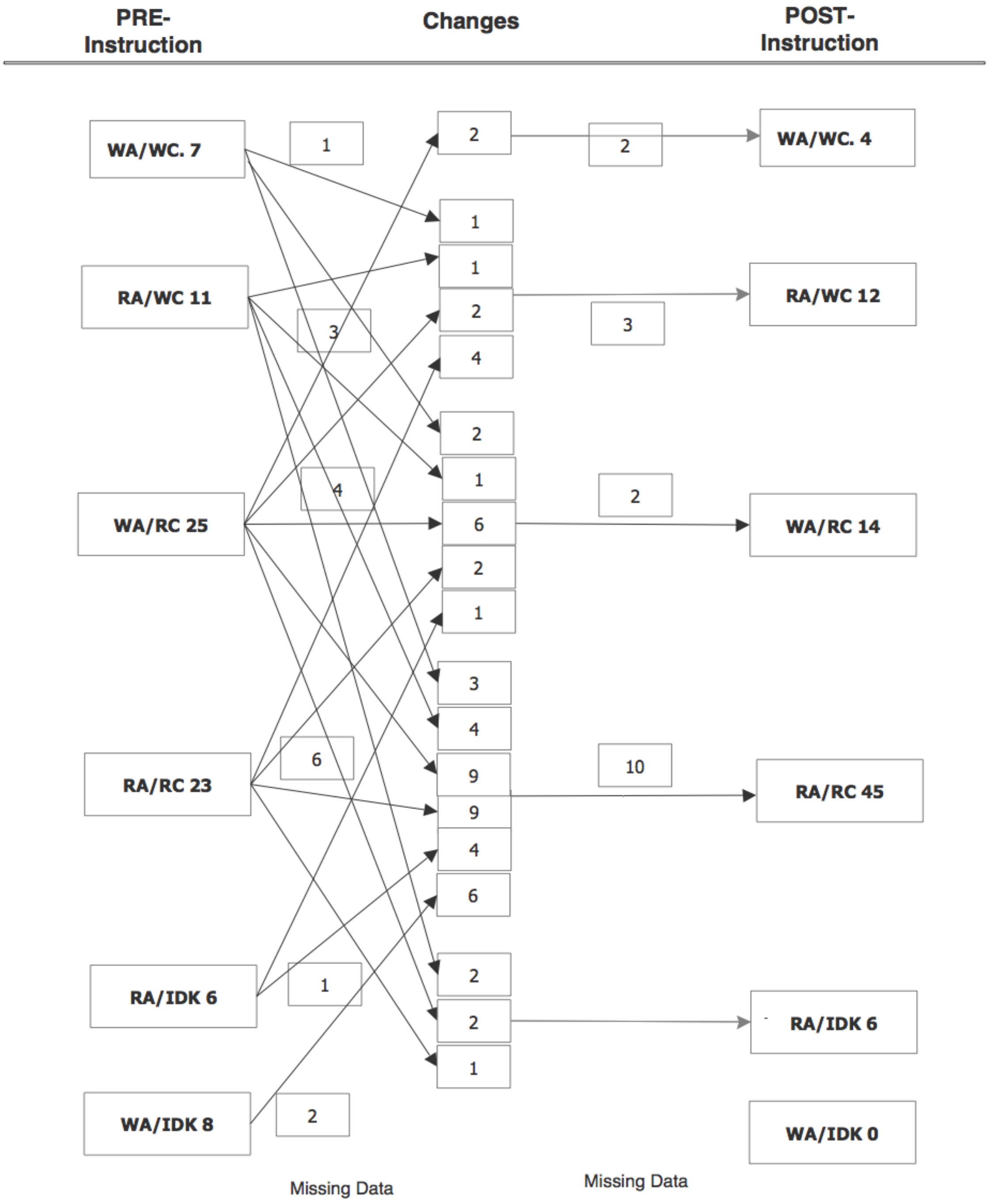


Figure 7. *Students' answers and how they changed from pre to post-test in Q1.*

Figures 7-10 reinforce our discussion that FCIs do not measure what is claimed. We can see a clear pattern where most of the students' choices are not a result of misconception as evident by how they changed their answers. It could be that instruction may have had somehow a negative impact on student understanding.

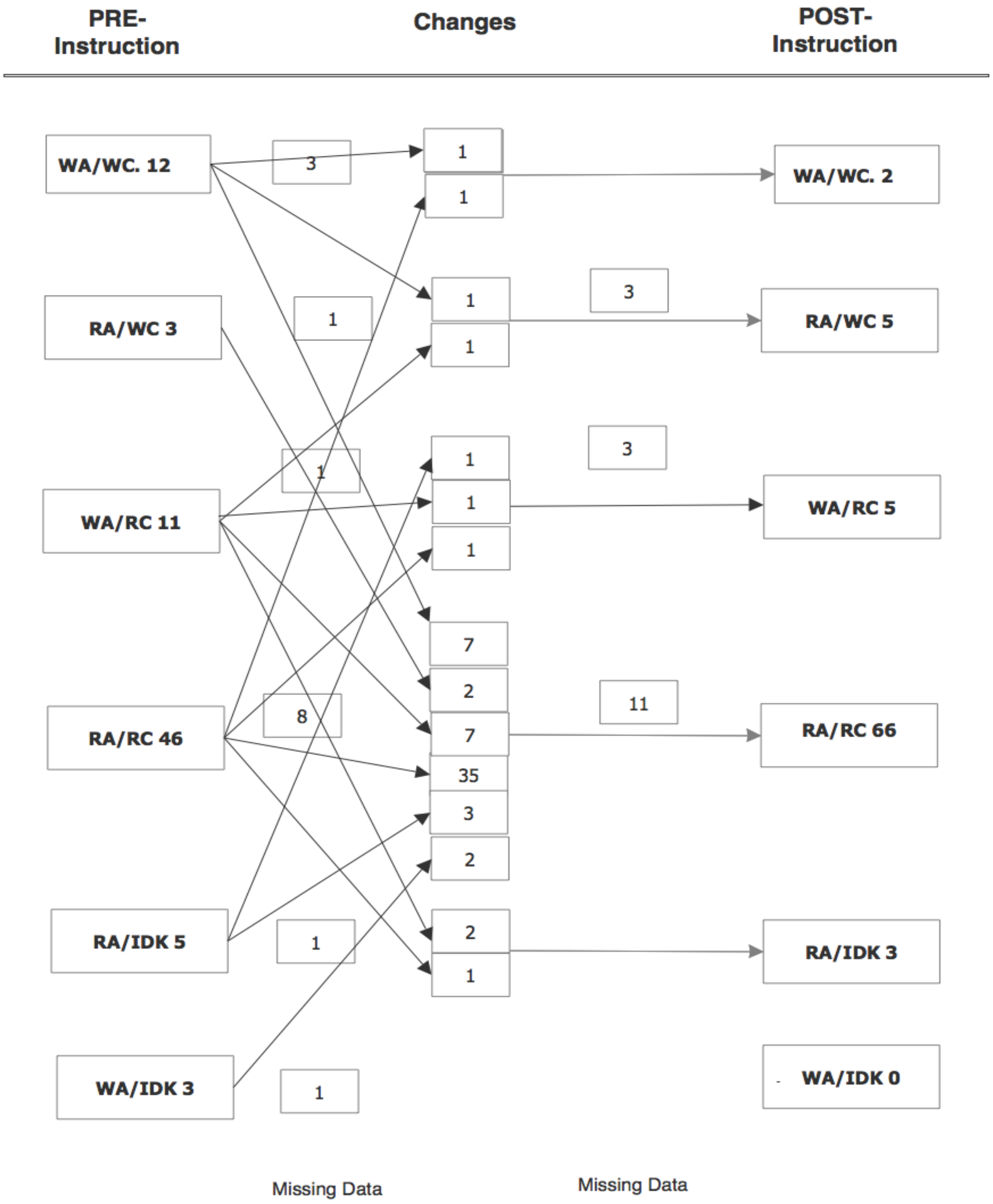


Figure 8. *Students' answers and how they changed from pre to post test in Q2.*

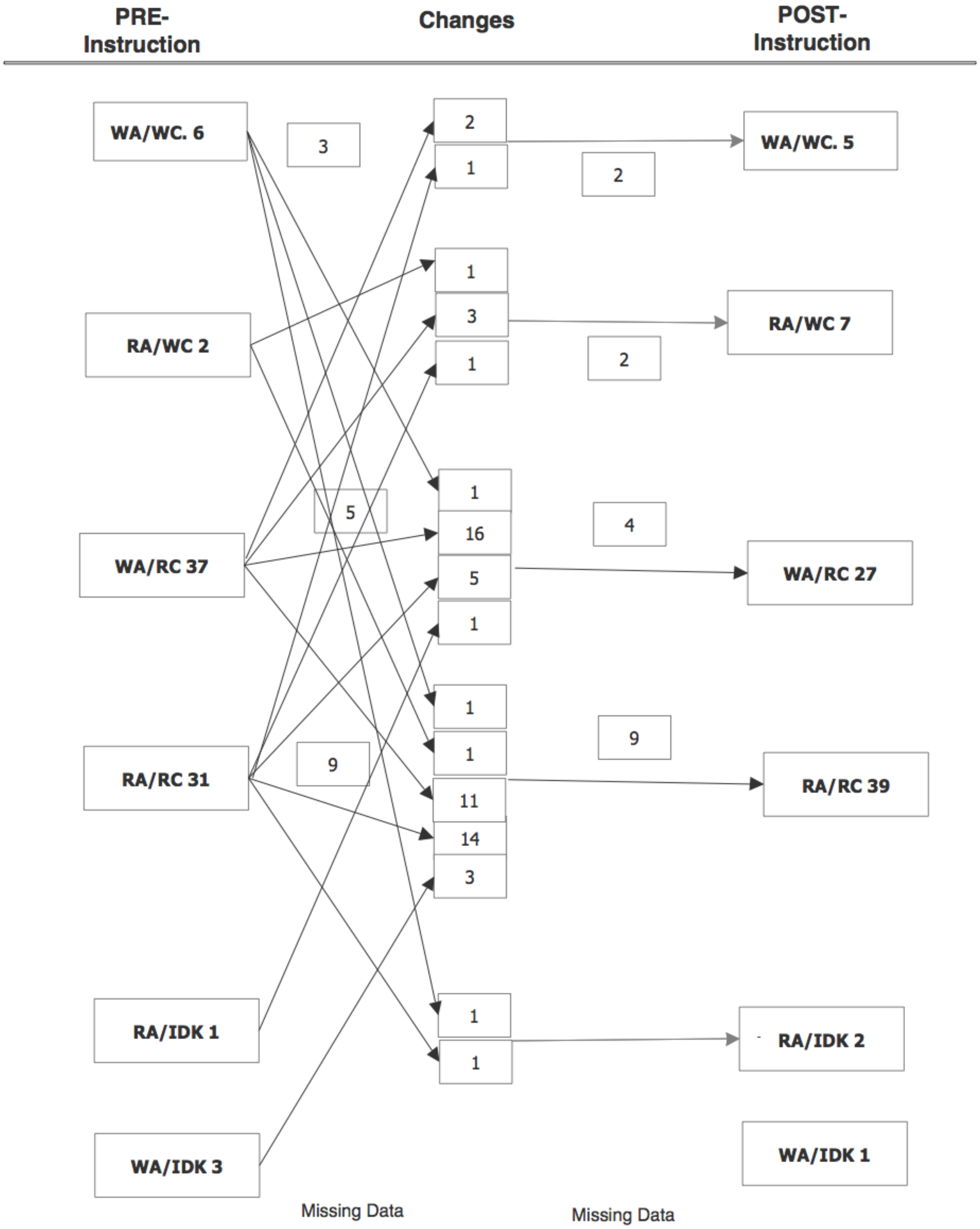


Figure 9. *Students' answers and how they changed from pre to post test in Q3.*

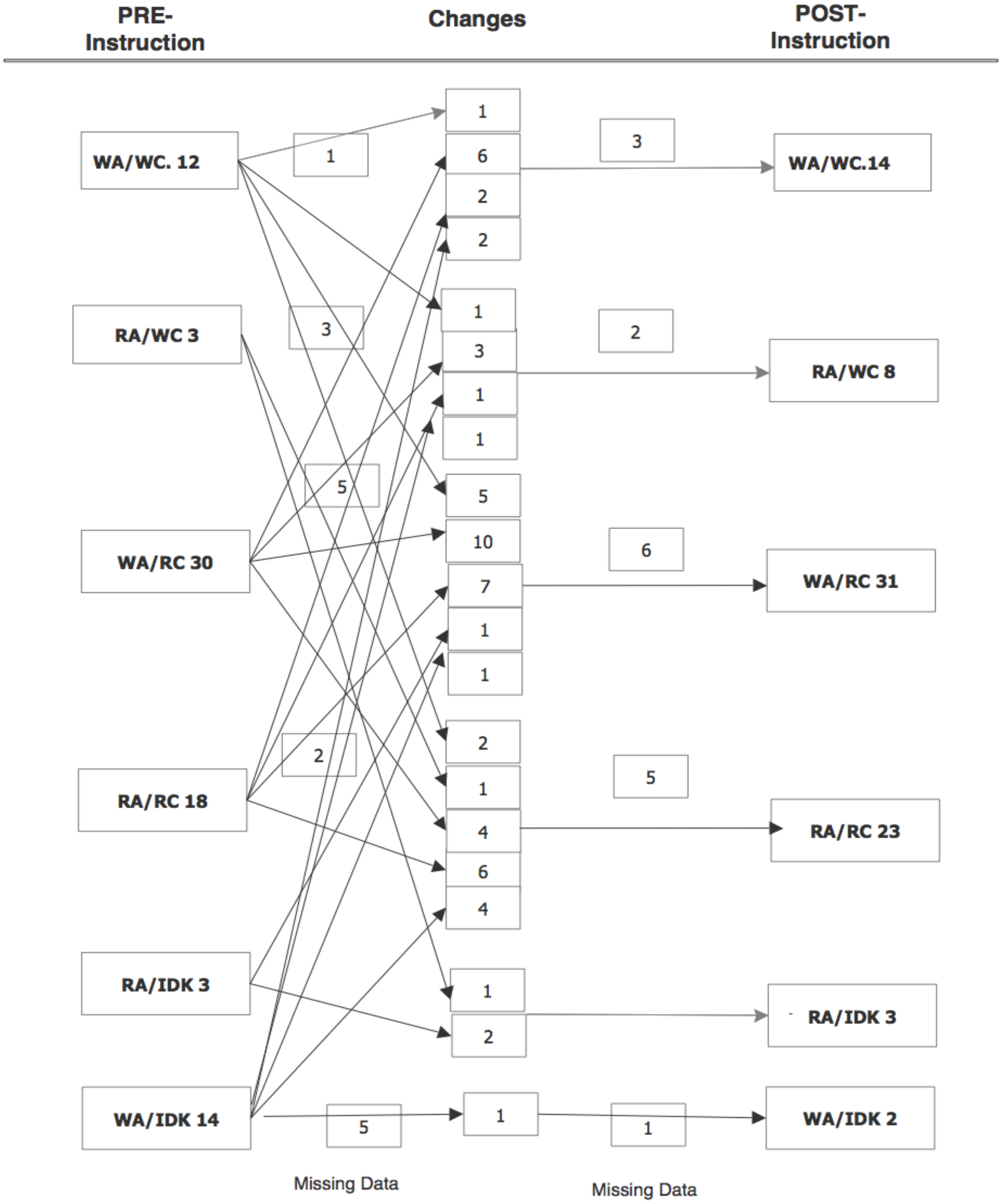


***Figure 10**. Students' answers and how they changed from pre to post test in Q13.*

In results presented by Fadaei and Mora students had difficulty with Q13, and here students' selections show that there are large misunderstandings as well. However, more than 50% of students understood the concept at pre-test. Thus, in our sample, there is a misunderstanding that caused 7 students to go from RA/RC to WA/RC, which deserves more investigation. Nevertheless, these figures reveal that students have a willingness to change their answers, which leads to interpret the results as misunderstandings rather than misconceptions. These patterns contradict Hestenes, and Halloun's (1995) objection in which they say: "the FCI data show that students are not easily induced to discard their misconceptions in favor of Newtonian concepts". Student's choices examined more closely shows that their answers and reasoning change, suggesting that they are not holding on to any preconceived beliefs. Further, only 16 students scored above 55% at pre-test versus 12 at post-test. For Q1, 13 students selected RA/RC versus 9 at post-test, however, 8 stuck with their pre-test choices. In Q2, Q3, and Q13, this was not the case. From pre-test to post-test the students changed largely; For example, in Q2, 8 students selected RA/RC, but only 1 did so at post-test. This is consistent with the finding of our factor analysis. Students' concepts are loosely connected with the six conceptual dimensions extracted. One would have to consider students getting at least 21 questions correctly to have a consistent selection of 60% which would give the result speculated by Hestenes and Halloun (1995).

We suggest that if the students had the option to choose "*I don't know*", a large number would select this at least at pre-test. We observed that when we gave students that choice, they chose IDK more often at pre-test. This data also shows that Hestenes, and Halloun's (1995) conclusion is incorrect, and students have a willingness to change. However, misunderstandings plagued students perhaps due to the I-language and L-language gap (Crogman, 2017). Sayer (2013) argues that, "the multiplicity of tasks in the comprehension process casts heavy unconscious burden on the comprehended, which renders comprehension potentially risky and liable for interpretive errors. Such errors may preclude extracting the intended meaning behind a piece of discourse causing misunderstanding." Thus, FCI measure of misconception may be a false positive. Therefore, like Huffman and Heller (1992), we too conclude that FCIs do not measure well students' misconceptions or predict their performance because of its ill-conception.

Limitations

Our power was sometimes limited by a sample with missing data. Also, the FCI data were from classes with the same instructor, which could induce bias. Future studies should consider broader groups from diverse classes and subjects with modified pre to post-test FCIs.

CONCLUSION

Most students lack formal scientific knowledge to construct concepts. Through their exploration, they formulate knowledge that may be entrenched in their cultural traditions which is not easily shaken without the introduction of conceptual conflicts in class. Conceptual conflicts cause learners to reexamine the reasons for their commonsense beliefs. Since CIs do not create conceptual conflicts there is no way for these tests to measure students' misconceptions. In this study students' selections from pre-test to post-test seem to be very scattered. Thus, our findings about students' correct answers does not stay consistent from pre-test to post-test but their overall performance does improve. Does FCI data "present a highly consistent picture, showing statistically reliable and discriminating measures of minimal performance in mechanics"? Hake's results (1997) were interpreted to suggest that students' pre-held beliefs were incompatible with Newtonian concepts, yet these determine students' performance. This is just a matter of interpretation, because students' choices are constantly shifting which is what we found, and what Griffith suspected. FCIs do not necessarily measure students' minimal performance or their misconception. Students' choices are the result of how well they understood the language of the instruction. Our findings show that students have a strong willingness to change their answers after instruction, suggesting that at pretest the students' lack of knowledge and ambiguities in the test are factors contributing to them guessing. The failure of traditional instruction is not that it overlooks the crucial influence of students' personal beliefs on what they learn, as Hestenes and Halloun (1995) perceived, but typically lacks the language for effective communication, which hampers students' ability to ask questions, and explore concepts in a hands-on fashion. Students' concepts are not well defined enough in their minds for FCIs to make any absolute conclusion about students' choices or misconceptions.

A literature review revealed that very little work has been done to assess and rethink FCIs in the past few decades. Yet, a few researchers have demonstrated that as currently built, these tests do not measure what they claim. We propose a new perspective on how to build FCIs based on our study, which allowed FCI takers to express more clearly their thoughts. Our proposal for an improved FCI version will allow instructors and researchers to more clearly parse out students' misunderstandings, from misconceptions, and commonsense beliefs. This new design opens doors to fairer and more organic testing/assessment practices in college STEM, and we hope will be the subject of further investigation.

REFERENCES

Atkinson, R. L., Atkinson, R. C., Smith, E. E., & Bem, D. J. (1993). *Introduction to Psychology* (11th ed.). Fort Worth, TX: Harcourt Brace Jovanovich.

Bruun, J., & Brewe, E. (2013). Talking and learning physics: Predicting future grades from network measures and Force Concept Inventory pre-test scores. *Physical Review Special Topics-Physics Education Research*, *9*(2), 020109.

Chi, M. T. H. (2008). Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In S. Vosniadou (Ed.), Handbook of research on conceptual change (pp. 61–82). Hillsdale, NJ: Erlbaum)

Chiesa, M. (1992). Radical behaviorism and scientific frameworks: From mechanistic to relational accounts. *American Psychologist, 47*, 1287–1299. http://dx.doi.org/10.1037/0003-066X.47.11.1287

Chiesa, M. (1994). *Radical behaviorism: The philosophy and the science*. Boston, MA: Authors' Cooperative.

Crogman, T. H. (2017, December). Grasping the interplay between the Verbal Cultural diversity and Critical thinking, and their Consequences for African American education. In *Frontiers in Education,* 2 (p. 64). Frontiers.

Crogman, T. H. & TrebeauCrogman, M. (2018). Modified generated question learning, and its classroom implementation and assessment. *Cogent Education, 5*(1), 1459340.

Crogman, T. H. & TrebeauCrogman, M. (2016). Generated questions learning model (GQLM): Beyond learning styles. *Cogent Education*, *3*(1), 1202460.

Crogman, T.H., TrebeauCrogman, A.M., Warner, L., Mustafa, A.,& Peters, R.(2015).Developing a new teaching paradigm for the 21st century learners in the context of Socratic methodologies. *British Journal of Education, Society & Behavioral Science, 9*, 62–95.

Crouch, C., & Mazur, E. (2001). Peer instruction: Ten years of experience and results. *American Journal of Physics*, *69*(9), 970-977.

Demirci, N. (2005). A Study about Students' Misconceptions in Force and Motion Concepts by Incorporating a Web-Assisted Physics Program. *Turkish Online Journal of Educational Technology-TOJET*, *4*(3), 40-48.

Fadaei, A. S., & Mora, C. (2015). An investigation about misconceptions in force and motion in high school. *US-China Education Review*, *5*(1), 38-45.

Griffiths. D, (1997). Millikan Lecture 1997: Is there a text in this class? *Am. J. Phys.* 65: 1141-1143.

Hall, G. (2003). *The psychology of learning*. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (Vol. 2, pp. 837–845). London: Nature Publishing Group.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64-74.

Hestenes, D. (1998). Who needs physics education research!? Am J Phys. 66: 465–7.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The physics teacher*, *30*(3), 141-158.

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *The Physics Teacher*, *33*(8), 502- 502.

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher*, *33*(3), 138-143.

Hume, D. (1748). *An inquiry concerning human understanding, 1955*. Indianapolis, IN: Bobbs-Merrill.

Humphreys-Jones, C. (1986). Make, make do and mend: The role of the hearer in misunderstandings. In G. McGregor (Ed.) Language for Hearers(pp. 105-126). Oxford, England: Pergamon Press.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development* Englewood Cli., NJ: Prentice-Hall.

Kolb, D. (1985). *Learning style inventory: Self scoring inventory and interpretation booklet*. Boston, MA: McBer & Company.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, *77*(6), 1121

Lewin, K. (1951). *Field theory in social science*. New York, NY: Harper & Row.

Leith, D. (1987). Drag on non-spherical objects. *Aerosol science and technology*, *6*(2), 153-161.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2017). Best practices for administering concept inventories. *The Physics Teacher*, *55*(9), 530-536.

Sayer, I. M. (2013). Misunderstanding and language comprehension. *Procedia-Social and Behavioral Sciences*, *70*, 738-748.

Schunk, D. H. (1991). *Learning theories: An educational perspective*. New York, NY: Macmillan.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York, NY: Appleton-Century.

Thornton R. K, Sokoloff D. R. (1998). Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula. *American Journal of Physics*, *66* (4): 338.

Tolman, E. C., & Honzik, C. H. (1930). *"Insight" in rats*, Vol. 4, pp. 215–232. Berkeley, CA: University of California, Publications in Psychology.

APPENDIX A

Proposed modifications for FCI

1. Two metal balls are the same size, but one weighs twice as much as the other. The balls are dropped from the roof of a single-story building at the same instant of time. The time it takes the balls to reach the ground below will be:

A. *About half as long for the heavier ball as for the lighter one.*
About half as long for the lighter ball as for the heavier one.
B. *About the same for both balls.*
C. *Considerably less for the heavier ball, but not necessarily half as long.*
D. *Considerably less for the lighter ball, but not necessarily half as long.*
E. *I do not know**

1.1. What concept was used in the above question?

A. *Projectile motion*
B. *Newton 2nd Law*
C. *Energy*
D. *Conservation of Momentum*
E. *I do not Know**

2. The two metal balls of the previous problem roll off a horizontal table with the same speed. In this situation:

A. *Both balls hit the floor at the same horizontal distance from the base of the table.*
B. *The heavier ball hits the floor at about half the horizontal distance from the base of the table than does the lighter ball.*
C. *The lighter ball hits the floor at about half the horizontal distance from the base of the table than does the heavier ball.*
D. *The heavier ball hits the floor closer to the base of the table than the lighter ball, but not necessarily at half the horizontal distance.*
E. *The lighter ball hits the floor considerably closer to the base of the table than the heavier ball, but not necessarily at half the horizontal distance.*
F. *I do not know**

2.2. What concept was used in the above question?

A. Free Fall
B. Vector Addition
C. Newton's 3rd Law
D. Superposition principle
E. I do not Know*

3. A stone dropped from the roof of a single-story building to the surface of the earth:

A. Reaches a maximum speed quite soon after release and then falls at a constant speed thereafter.
B. Speeds up as it falls because the gravitational attraction gets considerably stronger as the stone gets closer to the earth.
C. Speeds up because of an almost constant force of gravity acting upon it.
D. Falls because of the natural tendency of all objects to rest on the surface of the earth.
E. Falls because of the combine effects of the force of gravity pushing it downward and the force of the air pushing it downward.
*F. I do not know**

3.1. What concept was used in the above question?

A. Projectile motion
B. Gravity
C. Energy
D. Velocity
*E. I do not Know**

** FCI Statement modification introduced in our pilot test.*