Examining evolving performance on the Force Concept Inventory using factor analysis

M. R. Semak,[1] R. D. Dietz,[1] R. H. Pearson,[2] and C. W. Willis[1]
[1]Department of Physics and Astronomy, University of Northern Colorado, Greeley, Colorado 80639, USA
[2]Department of Statistics, Grand Valley State University, Allendale, Michigan 49401, USA
(Received 21 October 2015; revised manuscript received 1 September 2016; published 10 January 2017; publisher error corrected 19 January 2017)

The application of factor analysis to the *Force Concept Inventory* (FCI) has proven to be problematic. Some studies have suggested that factor analysis of test results serves as a helpful tool in assessing the recognition of Newtonian concepts by students. Other work has produced at best ambiguous results. For the FCI administered as a pre- and post-test, we see factor analysis as a tool by which the changes in conceptual associations made by our students may be gauged given the evolution of their response patterns. This analysis allows us to identify and track conceptual linkages, affording us insight as to how our students have matured due to instruction. We report on our analysis of 427 pre- and post-tests. The factor models for the pre- and post-tests are explored and compared along with the methodology by which these models were fit to the data. The post-test factor pattern is more aligned with an expert's interpretation of the questions' content, as it allows for a more readily identifiable relationship between factors and physical concepts. We discuss this evolution in the context of approaching the characteristics of an expert with force concepts. Also, we find that certain test items do not significantly contribute to the pre- or post-test factor models and attempt explanations as to why this is so. This may suggest that such questions may not be effective in probing the conceptual understanding of our students.

DOI: 10.1103/PhysRevPhysEducRes.13.010103

I. INTRODUCTION

The Force Concept Inventory (FCI) is a multiple choice test developed to assess students' understanding of fundamental Newtonian force concepts [1]. It is usually administered before and after instruction with the hope of gauging, in particular, the effect of such instruction. Still, several researchers have wondered how effective the FCI is in making such an assessment and what else can be learned from its results.

The FCI and factor analysis share a long history. Factor analysis is a statistical technique used to find a set of patterns, or factors, in a collection of data. Three years after the publication of the FCI, Huffman and Heller [2] used factor analysis to investigate the results of administering the FCI to almost 1000 students. They concluded that the few identifiable factors did not correspond well with the conceptual dimensions the authors of the FCI had proposed. There ensued a lively exchange [3,4] in which the applicability of factor analysis to the FCI was both called into question and defended. Nothing conclusive was determined.

Scott, Schumayer, and Gray [5] applied factor analysis to investigate how over 2000 students answered the FCI questions after they had completed the mechanics section of an introductory physics course. They determined that five factors could be identified.

Here we concentrate on comparing student performance on the FCI before and after instruction in mechanics, and the consequent evolution of the factor pattern. In doing so, we explore the use of factor analysis in assessing our students' ability to assimilate Newtonian concepts. However, we mainly examine the use of factor analysis in gauging the effect of instruction on the associations among physical ideas that can be made by students. Each FCI question can be seen to deal with a specific mechanical concept(s) (as interpreted by an expert). Also, certain subsets of questions refer to the same concept. The students' response patterns to these questions can tell of associations made between and among certain questions, thus ideas. Moreover, if a question's response pattern is such that little, if any, association with other test items is found, it may be that this question is unique in its conceptual content. On the other hand, a question may simply not be clear and, in turn, it is misinterpreted. This could influence the student response pattern to deviate from what would be expected given the question's actual conceptual content. Such an outcome leads us to examine the question as to its effectiveness in testing a student's understanding of a particular concept.

To be clear, whether students answer FCI questions correctly or not, factor analysis of their test responses can reveal patterns. We will interpret any of these patterns as some type of association. The evolution (pre- to post-instruction) of these patterns may give us insight as to how

the students' process of grasping mechanical ideas develops.

II. METHODOLOGY

The FCI data set consists of the results of 427 pre- and post-tests collected over a four-year period from students in algebra and calculus-based introductory physics. The populations taking the pre- and post-tests consisted of the same students. Newtonian mechanics was the subject of study over most of the semester for all students. The variables to be analyzed consist of a string of zero's and one's for each of the 30 questions (test items) in which a one (zero) indicates that a student answered correctly (incorrectly). The length of each string is equal to the sample size for the corresponding test (427).

Exploratory factor analysis is a statistical technique that attempts to achieve a parsimonious explanation of observed data, or variables [6]. This method tries to describe the correlations among these variables in terms of an underlying structure, a set of factors. In developing a factor model, a correlation matrix for the observed variables is constructed. Variables which form highly correlated groups will tend to form a factor. Particular eigenvalues of the correlation matrix correspond to these factors and indicate the degree to which the corresponding variables are related [6]. Moreover, the analysis reveals factor loadings, which are the correlations (positive or negative) of a factor with its associated variables. In this way, factor analysis attempts to find a group of factors onto which subsets of variables "load." Overall, the researcher wishes to find the minimum number of factors that can sufficiently explain the correlations among the observed data [7].

Ultimately, however, it is desirable that these factors will be amenable to a coherent conceptual interpretation; i.e., the factors will have a useful meaning to the researcher. A rotation strategy can be helpful in this regard. Geometrically, the factors can be seen as axes with which the associated variables (questions), seen as vectors, tend to align, thus defining the factor. Initially in the analysis, the factor axes are orthogonal demonstrating that no intercorrelations exist among the factors. These axes may be rotated to optimize loadings so as to facilitate an easier interpretation of the factors. A rotation can be orthogonal which leaves the factors uncorrelated, or correlations can be allowed by way of an oblique rotation [6]. In the latter case, a set of orthogonal factors, seen as axes, is transformed such that the factors are no longer necessarily mutually orthogonal and the projections they make on one another are indicative of their correlations [6]. (The factors are oriented in the sense of a nonrectangular Cartesian coordinate system in which the axes are not necessarily orthogonal to one another.) The correlations among the variables are still well defined [6].

As mentioned, in performing such an analysis with the FCI, the variables are the questions on the FCI. What is observed is each student's success with each question (whether the student answers the question correctly or not). So, the data we examine are dichotomous. With such variables, the Pearson [7] correlations used when handling continuous data with factor analysis are inappropriate. A matrix of tetrachoric correlations is considered proper [8–10]. If one assumes that the dichotomous measurements are based on continuous variables that are not directly observable, a tetrachoric correlation is an estimate of the values of the correlations among these underlying latent variables [10].

Using Mplus [11], a software package capable of calculating tetrachoric correlations, a factor model was sought for both our pre- and post-test FCI data using the weighted least squares with mean and variance adjustment (WLSMV) estimator [11]. We allowed for correlations among the factors by employing an oblique (Quartimin [6]) rotation. We used Mplus to generate a set of models for between 1 and some maximum number of factors (nine for our study) along with their respective fit statistics. An n-factor model uses the factors corresponding to the n largest eigenvalues of the correlation matrix. In determining which is the optimum model (determining the number of factors to retain), researchers have considered keeping only the factors corresponding to eigenvalues greater than 1 [12]. The fit statistics have also been used in determining the number of factors for one's model [9]. We used the method, parallel analysis, along with the fit statistics for this task [9,13–15].

Parallel analysis is a technique which can be used to determine the number of factors to be retained in exploratory factor analysis [13]. With this method, a correlation matrix is computed from a random set of data using the same number of variables and observations as the original sample. Comparing the eigenvalues of the correlation matrices for the randomized and original data set, one is led to the appropriate number of factors for one's model. If the eigenvalues generated via the randomized data are larger than some subset of those for the original data set, the factors corresponding to this subset are considered insignificant [16] [see Figs. 2(a) and 2(b)]. We performed a parallel analysis on our data using R, an open-source statistical programming language capable of performing parallel analysis for data requiring that tetrachoric correlations be used [17,18].

Our factor model was then developed using these computational tools. When the analysis was first performed, questions 2, 3, and 29 had comparatively small loading values for both the pre- and post-test cases. As a result, these were omitted from the study. A subsequent analysis led to the selection of five factors for the pretest model and six factors for that of the post-test.

It should be noted that for both the pre- and postmodels, the sample tetrachoric correlation matrices were not positive definite, having some (small) negative eigenvalues. Figures 1(a) and 1(b) show scree plots for these

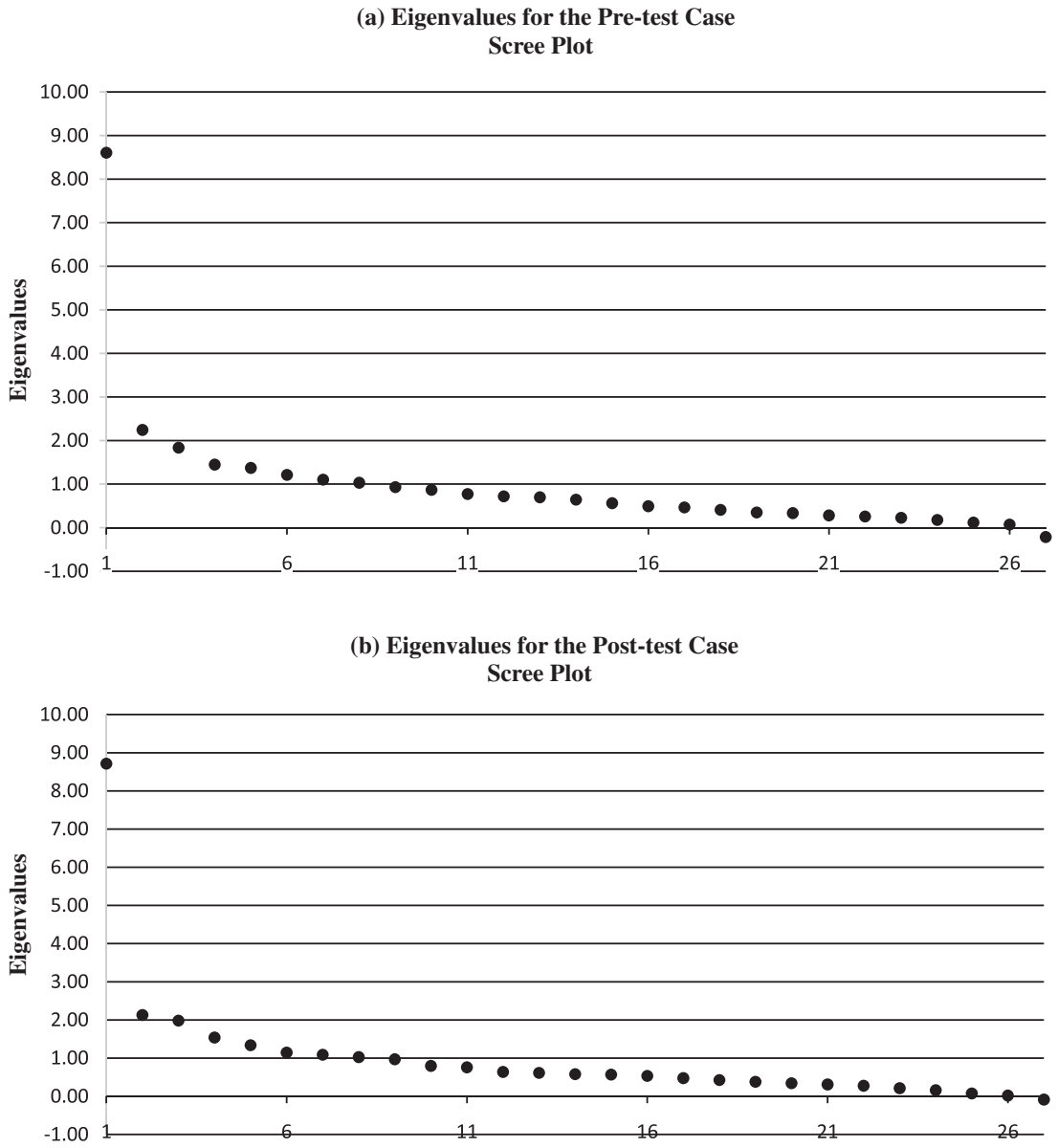

FIG. 1. The scree plots shown for the (a) pretest and (b) post-test cases.

eigenvalues. At first, one might be concerned that this invalidates these models given the eigenvalues' interpretation as correlation measures. However, given the way in which these matrices are constructed using Mplus, this can happen [19]. The methods used for estimation do not assume a positive-definite sample correlation matrix when considering dichotomous data [19]. Moreover, that there are small negative eigenvalues is not significant in that the correlation matrices fit the samples well [19].

III. FACTOR MODEL FIT

One of the fit statistics we considered is the "chi-square test for model fit" that compares the fit of a model with a certain number of factors to that of a model which perfectly reproduces all of the correlations that are observed and has as many factors as variables. The latter model is considered to be saturated [20]. Such a model does not serve the cause of parsimony, however. One requires a model in which subgroups of correlated variables are identified, thus revealing a smaller number of factors than variables while adequately explaining the correlations among the variables. So, each of a set of models with between 1 and some maximum number of factors is tested via the chi-square test for fit to the saturated model. A parsimonious model with a satisfactory fit to the saturated model is sought. The null hypothesis is that the model for a given number of factors is sufficient in explaining the correlations among the variables. A lower chi-square value indicates a better fit. The test's corresponding p value is, of course, the probability of rejecting the null hypothesis when it is true [21]. So, if the

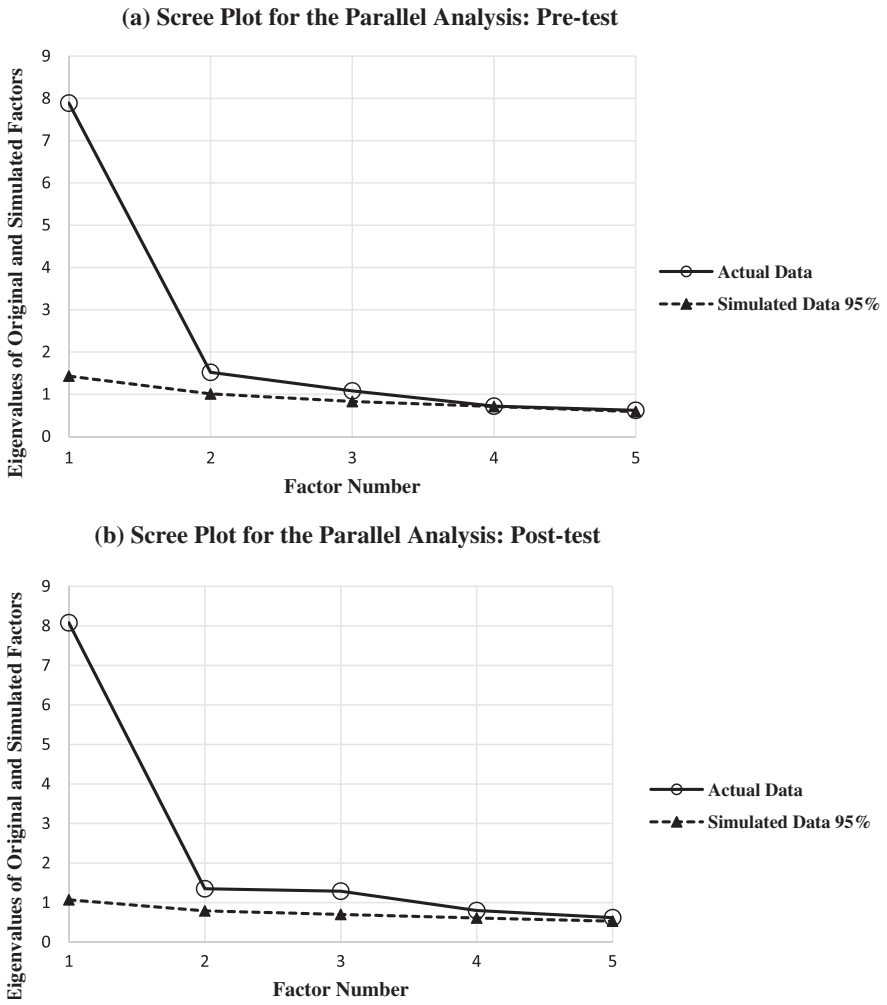

FIG. 2. The (partial) scree plots resulting from the parallel analysis for the pretest (a) and post-test (b) cases. As shown, for each case, five eigenvalues from the original data's correlation matrix were found to be larger than the eigenvalues of the correlation matrix for the simulated data. This suggests that there are five factors for each case, pre- and post-test. For the post-test case, we actually decided to use six factors (please see text for explanation). The other eigenvalues for the original data take on values lower than those for the simulated data. Each eigenvalue for the simulated data is calculated at a 95% confidence level.

chi-square test is insignificant ($p > 0.05$) for a model with a certain number of factors, the model is seen to adequately explain the correlations among the variables [11,22]. We also consider a baseline model for both the pre- and post-test in which the variables have random correlations. This allows us to determine a bad chi-square test value. The baseline models will have large chi-square test values indicating a poor fit to systems with strong correlations among the variables.

For our FCI pretest data, parallel analysis suggests five factors be used. This model has a chi-square value of 236.16 (with 226 degrees of freedom and a p value of 0.31) as compared to the baseline model's value of 2432.88 (with 351 degrees of freedom and negligible p value). The five factor model's test is considered insignificant [9,22]. For our FCI post-test data, parallel analysis, again, suggested the use of five factors. However, a five factor model yields a significant chi-square test ($p = 0.01$). A six factor model has a chi-square value of 234.48 (with 207 degrees of freedom and a p value of 0.07) as compared to the baseline model's value of 3695.53 (with 351 degrees of freedom and negligible p value). The six factor post-test model's chi-square test, as in the case for the pretest data, is considered insignificant [9,22]. Given this, and the greater interpretability of the six factor model, we choose six factors for the post-test case. (Moreover, Cho *et al.* found that when in error, the polychoric-based parallel analysis method mostly underestimated the number of factors [23]).

Next, the pair of fit indices, the comparative fit index (CFI) [14,22] and the Tucker-Lewis index (TLI) [14,22] compares our factor model for a certain number of factors to the corresponding baseline model. Indeed, these indices effectively quantify the differences in the chi-square values for our factor model and the baseline model while taking the number of degrees of freedom for each into account. The indices' values range from 0 to 1 with values over

0.900 indicative of a reasonably good fit [14,22]. Both the CFI and TLI depend on the average size of the correlations in the data. The higher the correlations, the closer these values are to 1. The five factor model for our pretest data has a CFI of 0.996 and a TLI of 0.998. As for the six factor post-test model, the CFI is 0.991 and the TLI is 0.984.

IV. DISCUSSION

A. Test statistics and items 2, 3, and 29

First, Table I gives some basic quantities describing our pre- and post-test results. We present this table for completeness. As mentioned, data have been collected for four years with students diverse in backgrounds in algebra and calculus-based introductory physics courses. The instruction tends to be consistently lecture oriented for both classes.

Next, as mentioned, test items 2, 3, and 29 did not have loadings high enough in either the pre- or post-test case to warrant their inclusion in the rest of the study. Could it be that these questions are not very challenging or not challenging enough, thereby generating trivial strings of 0's or 1's as variables which would not correlate significantly to other variables with less trivial strings? As a simple measure, looking at the percentage of students answering a question correctly given the test's mean and standard deviation (determining if a percentage falls within, at least, 1 standard deviation from the mean) may suggest the question's level of difficulty. For the pretest, 23%, 38%, and 34% of the students, answered questions 2, 3, and 29, respectively, correctly. On the post-test, 46%, 65%, and 59% of the students, answered questions 2, 3, and 29, respectively, correctly. With the mean score on the pretest being 34% (with a standard deviation of 17%) and that of the post-test being 50% (with a standard deviation of 19%), we feel this suggests that the difficulty level of these items should not lead to an issue. Given this rather simple explanation, we chose to explore other reasons as to the relatively small loadings. Still, as part of an extended study, we would like to examine the items' challenge levels more carefully (controlling for student ability) for our population.

Question 2 is related to the test's first question. Question 1 asks about the time it will take two balls having different masses to hit the ground when dropped at the same time. Question 2 concerns the same two balls rolling off of a horizontal table and asks about where the balls land in relation to one another as they hit the floor. In the case of the pretest, one may answer question 1 correctly given access to popular and/or high school science without some depth of understanding of mass, acceleration, force, and freely falling objects. (The correct answer may be memorized.) Then, one's performance on question 1 may have no bearing on one's performance on question 2. Moreover, question 2 also involves the horizontal component of the balls' motions possibly lending more confusion to one's preinstruction experience. It could be argued that the other FCI items have a weak, if any, relation to the nuances that could be found in question 2. So, that the second item has little to no correlation to any other test item is not surprising. However, in the post-test case, this weak correlation persists for question 2. We feel that our instruction did not give enough time to carefully cover the concept of mass and free fall motion (we mostly spent time with freely falling objects in one and two dimensions without discussing the question of mass directly). Student confusion could have persisted and accounted for the weak loading values of this item postinstruction.

TABLE I. Pre- and post-test statistics, $N = 427$.

	Pretest	Post-test
Mean score	0.342	0.495
Standard deviation	0.174	0.194
Skewness	0.977	0.501
Kurtosis	0.905	−0.439

As for questions 3 and 29, whether considering the pre- or post-test case, there may have been confusion concerning if and how the force due to a mass of air is to enter the problems. Question 3 mentions the force of the air in a distractor and question 29 addresses it in the problem statement. Question 3 asks what is true about a stone falling from a roof to the surface of Earth. The last response claims the stone falls because of gravity and the force of air acting downward on it. The other responses do not mention a force due to air. Even if this response is incorrect, a student may wonder if the effect of a force due to the air should be considered. If so, how will it affect the interpretation of the other responses? Will the stone reach a terminal speed rather than continue to accelerate downward? Both situations are given as possible answers. In a pretest situation, a student may, indeed, ask whether such a force should be considered and if so, how. Postinstruction, students may feel that the question intends for one to consider only the "textbook" freely falling case in which any effect of air is assumed negligible. Certain other test items in which the force of the air could be considered do not mention it. Moreover, our students had not yet been formally exposed to the concept of buoyancy and/or air resistance. So, question 3 may have poor correlation to any other questions for such reasons. A similar discussion can be held for question 29. This item asks about which forces are acting on an empty office chair at rest on the floor. A "net downward force exerted by the air" is one of the possibilities to be considered [1]. In our past (unpublished) research, students who took the FCI (pre- and post-tests) were interviewed one on one to see what might be confusing about the test items. Certain students did claim that question 29 was somewhat confusing given their lack of experience with buoyancy along with the wording of the response mentioned above (a distinction between "downward" and "net downward" seemed problematic—does this mean that the net force due to the air is downward or that

there is an overall downward force due to the air in addition to the force the air from other directions?). We are still attempting to interpret these results.

As has been suggested to us, it may be interesting and, indeed, more instructive, to study the actual student responses to these questions rather than, simply, whether the questions were answered correctly or not. Looking at more detailed response patterns along with think-aloud interviews concerning these questions, is certainly a future interest for us.

Concerning the items that were used in our factor analysis, we chose to consider loadings larger than or equal to 0.340 in absolute value for both the pre- and post-test models. This establishes a "noise floor," above which questions load onto factors in a way amenable to physical interpretation. Choosing a cutoff value in this way is typical in factor analysis and is based on the researcher's judgment [9].

Also, as mentioned concerning questions 2, 3, and 29, it may be that certain test items correlate significantly or not due to trivial response patterns due to, say, a question being too challenging or not challenging enough. Again, question difficulty may be suggested by the percentage of students answering a question correctly or incorrectly given a test's mean and standard deviation. For the pretest (with mean 34% and standard deviation 17%), over 60% of our students answered items 1, 6, 7, 12, 16, 17, 24, and 27 correctly. For the pretest factor model, questions 6, 7, 12, 24 and 27 did load on one factor (however, without distinction) with other items and items 1 and 16 loaded together with one other item. Question 17 loaded with other items. Also, for the pretest, questions 5, 17, and 18 were answered correctly by under 10% of our population and loaded together (without great distinction) along with other items. For the post-test (with mean 50% and standard deviation 19%), over 70% of our students answered items 1, 6, 7, 12, 16, and 24 correctly. In the post-test model, questions 1, 6, and 7 formed their own factor while items 16 and 24 loaded on separate factors (without distinction). Question 12 did not have a loading significant for this model. Moreover, for the post-test case, less than 28% of our population answered questions 5, 17, 18, 25, 26, and 30 correctly. Questions 5, 18, 26, and 30 load onto a factor with other items while questions 17 and 25 form their own factor.

We see that items on each test that can be grouped as difficult or not difficult can load together onto a factor. Yet, none of these groups in their entirety load onto the same factor. Moreover, other items (not in these groups) often share the same factor with members from these groups. Items from these groups which do load onto the same factor do not necessarily do so with comparable strengths. Overall, we feel the need for a further explanation as to the placement of the items of these groups in the factor models. (However, these results do need further study.) With that, we will proceed with the explanations below.

Again, to be clear, for what follows, all test items except questions 2, 3, and 29 are considered.

B. Factor model for the pretest

In the case of the pretest, all of the questions met the cutoff criterion. Table II shows the absolute values of the loadings for this model before filtering out values below the threshold of 0.340. Table III shows questions having loadings with absolute values above this threshold and ordered so as to best display the factor structure. The five factor model contains factors that, in general, are not simply identified with a specific introductory physics concept.

The first factor has a set of questions asking about forces (including centripetal), which are suddenly applied or discontinued, resulting speeds and trajectories due to such changes, two-dimensional motion in the presence of a constant force, and identifying forces being applied to an object. Possibly, these questions are correlated due to most

TABLE II. The factor model results are shown for the FCI pretest analysis using Mplus with WLSMV estimation on tetrachoric correlations. A QUARTIMIN rotation was effected. Here the absolute values of the loadings are shown. The italicized values are those below the threshold of 0.340.

	FACTORS				
Item	1	2	3	4	5
1	0.134	0.350	0.051	0.268	0.129
4	0.042	0.003	0.252	0.627	0.028
5	0.252	0.007	0.491	0.010	0.338
6	0.383	0.197	0.083	0.024	0.032
7	0.421	0.108	0.012	0.018	0.062
8	0.583	0.134	0.014	0.050	0.078
9	0.269	0.437	0.065	0.013	0.218
10	0.664	0.062	0.065	0.004	0.208
11	0.212	0.279	0.518	0.044	0.022
12	0.508	0.260	0.068	0.117	0.009
13	0.360	0.140	0.532	0.030	0.074
14	0.632	0.052	0.010	0.081	0.091
15	0.060	0.055	0.198	0.891	0.168
16	0.020	0.343	0.083	0.329	0.176
17	0.060	0.099	0.589	0.336	0.018
18	0.307	0.058	0.432	0.130	0.106
19	0.081	0.013	0.068	0.005	0.673
20	0.068	0.035	0.045	0.043	0.673
21	0.407	0.215	0.249	0.030	0.037
22	0.637	0.304	0.222	0.001	0.030
23	0.769	0.001	0.089	0.021	0.007
24	0.550	0.057	0.124	0.108	0.066
25	0.279	0.071	0.709	0.024	0.171
26	0.054	0.067	0.714	0.013	0.494
27	0.536	0.209	0.051	0.041	0.052
28	0.008	0.053	0.036	0.881	0.260
30	0.513	0.002	0.306	0.116	0.030

TABLE III. The absolute values of the loadings 0.340 and higher are shown for the model of Table II. Also, the item ordering reflects the ordering of the loading values. An identification of the factors is attempted. Only one of the five factors is simply identified with specific force concepts. The others concern a mixture of Newtonian ideas.

	FACTORS				
	Dynamics including projectile motion, identify forces	Uncertain	First law, dynamics, identify forces	Third law	Kinematics, dynamics
Item	1	2	3	4	5
23	0.769				
10	0.664				
22	0.637				
14	0.632				
8	0.583				
24	0.550				
27	0.536				
30	0.513				
12	0.508				
7	0.421				
21	0.407				
6	0.383				
9		0.437			
1		0.350			
16		0.343			
26			0.714		0.494
25			0.709		
17			0.589		
13	0.360		0.532		
11			0.518		
5			0.491		
18			0.432		
15				0.891	
28				0.881	
4				0.627	
19					0.673
20					0.673

of them dealing with sudden changes in force and the consequences of these changes (dynamics). Also, many involve diagrams concerning resulting trajectories. Overall, these are not necessarily the associations that an expert would make given the physical nuances of these items.

The second factor involves question 9, which seeks a description of the resulting speed of an object initially moving with a constant velocity that is suddenly kicked in a specific direction. The student is asked to compare the resulting speed to the object's original speed and the speed that would be obtained if only acted on by the kick. (We see this question as rather nontrivial for someone who has not yet been instructed.) This second factor also has question 1, alluded to earlier, which asks about two objects of different masses freely falling. Also, there is question 16, which would be seen by an expert to be about the third law. It asks about a car pushing a truck after the car has reached a constant speed (as opposed to question 15, which asks about the same car pushing the same truck while the car is speeding up). We find it difficult to see why these items are correlated (indeed, their loadings values are relatively weak).

The third factor involves seven items. The two loading most strongly, items 25 and 26, ask about the same scenario, a box being pushed with a constant horizontal force across a horizontal floor. However, in the first of these questions, the box moves with a constant velocity as it is pushed. The next question asks about the motion if the force applied in the previous question is doubled. These questions are related by situation and that the net force determines how the box moves (dynamics). Yet, a more nuanced view may distinguish between aspects of the first and second law. The next item, question 17, asks about the forces applied to an object moving with a constant velocity. A connection with the first two items (especially item 25) can be argued here. However, the rest of this factor's items, as do certain items in factor 1, ask the student to identify the forces acting on objects moving with various trajectories (the net force on each object is nonzero in all but one of these questions). Question 13, which involves identifying forces acting on an object, also loads onto factor 1 with a smaller loading value than on factor 3. This double loading toughens the challenge one finds when attempting to precisely interpret these factors. Factor 3 appears to lack the item correlations one might find with expert test takers.

Factor 4 has items concerning the third law (items 4, 15, and 28). Here we do find a rather simple relation to mechanical concepts indicative of expert respondents. However, it is interesting to note that question 16 (described above) does not significantly load onto this factor when it seems clear that this item is more strongly related to the third law than to any of the concepts of factor 2 (the factor onto which question 16 does load significantly).

Finally, factor 5 contains two kinematics questions (19 and 20) somewhat strongly loading along with item 26 (loading nominally) described above. The kinematics items do stand out in that the way these questions are presented is not necessarily familiar to our students (for the pre- or postcase). (Yet, the item score was above average for item 19. They scored below average for item 20.) Still, these questions certainly do have a conceptual association. Also, question 26 has a dynamical character in that the forces resulting in an object's motion are important to this question, and it double loads onto factor 3, thus lending confusion to the process of interpretation. That item 26 loads onto this factor is puzzling. Yet, the character of the kinematics questions lends sense to this factor's structure.

C. Factor model for the post-test

First, for the post-test situation, three items, 8, 9, 12, and 14, did not have loading values above the threshold of

0.340. This raises questions as to how well students identified the conceptual content of these questions. Questions 8, 12, and 14 ask the student to identify which trajectory an object follows given some initial state. A selection of trajectories are presented from which to choose. Next, the characteristics of the speed of the object referred to in question 8 is the subject of question 9. The style of item 9 may be argued to present a challenge to the student before and/or after instruction. The one-on-one interviews we conducted (alluded to previously) revealed students to be confused by all of these questions after instruction. Students claimed to have had issues with the diagrams used for items 8, 12, and 14 and the wording of item 9. Also, questions 12 and 14 deal with projectile motion in which, again, students may have the same concerns about air resistance as they may have had with questions 3 and 29.

It is interesting that all of the loadings within the noise occurred in the case of the post-test. Could it be that, given instruction, students are aware of more details that can be considered and get confused? Could it be that the questions, themselves, have certain shortcomings? In any case, the factor analysis can suggest which test items may be problematic by indicating the failure of questions to significantly correlate with other questions concerning the same concepts. For example, items 12 and 14 share content with item 2 (projectile motion) and item 8 can be seen, at least, to be conceptually similar to several other FCI items which ask about the effect of a suddenly applied force.

Table IV shows the absolute values of all of the loadings for the six factor model of the post-test. Table V shows test items having loadings with absolute values of 0.340 and above.

The first factor for the post-test case has three items, two of which (6 and 7) load noticeably stronger than the third (1). Questions 6 and 7 are quite similar in that both concern centripetal force, a sudden loss of the latter, and the trajectory followed by the object in question after this loss. One could claim that these two questions ask about the first law if the trajectory is restricted to the plane parallel to that of the original circular path. Item 1 as described is the third item to load on this factor, and we are not certain why it does so. Its loading value is nominally smaller than the other items. Perhaps, for the post-test case, these questions have characteristics that allow for them to be answered through little reasoning or by memorization. Still, that items 6 and 7 have considerably higher loading values than does item 1 seems to indicate that students find association between 6 and 7 given the features of these two questions alone. With this, we will consider this factor to involve centripetal force or the first law.

Factor 2 appears to have two conceptual themes: force identification and the effect on the motion of an object by a temporarily or constantly applied force (dynamics). In the case of the pretest, all of the items of this factor were divided between two factors each pertaining to other concepts as well. Now, they are united under one factor. Possibly the students associate identifying forces with determining their effect on an object. Whatever the case may be, the sake of the model's conceptual clarity is better served by this arrangement of these items than that of the pretest model. [Also, it should be noted that three of these items (10, 24, and 27) have relatively weak loading values.]

The third law is the main concept explored in the items of factor 3. All items of the FCI directly related to the third law (4, 15, 16, and 28) load prominently on this factor. This includes question 16 which, in the case of the pretest factor model, loaded rather weakly onto factor 2 which we found difficult to interpret in terms of a single concept. For the post-test model, question 16, although having the lowest loading value for this factor, would be found by an expert to be associated with the third law items even though nuances arise in this question (pushing a truck after reaching a constant speed rather than during acceleration—as mentioned). This association may be indicative of some level of

TABLE IV. The factor model results are shown for the FCI post-test analysis using Mplus with WLSMV estimation on tetrachoric correlations. A QUARTIMIN rotation was effected. Here the absolute values of all of the loadings are shown. The italicized values are those below the threshold of 0.340.

	FACTORS					
Item	1	2	3	4	5	6
1	0.352	0.016	0.016	0.214	0.378	0.127
4	0.046	0.088	0.782	0.086	0.001	0.196
5	0.101	0.768	0.016	0.062	0.195	0.051
6	0.956	0.047	0.020	0.011	0.166	0.033
7	0.657	0.031	0.046	0.021	0.040	0.050
8	0.056	0.309	0.079	0.221	0.141	0.021
9	0.021	0.333	0.109	0.156	0.288	0.026
10	0.136	0.399	0.214	0.210	0.010	0.211
11	0.067	0.680	0.022	0.160	0.123	0.001
12	0.263	0.203	0.141	0.012	0.294	0.004
13	0.018	0.804	0.070	0.010	0.004	0.125
14	0.059	0.306	0.219	0.067	0.177	0.082
15	0.335	0.040	0.819	0.055	0.280	0.018
16	0.126	0.061	0.546	0.031	0.215	0.183
17	0.095	0.067	0.070	0.106	0.077	0.606
18	0.059	0.852	0.022	0.031	0.228	0.007
19	0.019	0.001	0.032	0.737	0.208	0.055
20	0.004	0.156	0.117	0.647	0.030	0.038
21	0.078	0.253	0.024	0.018	0.344	0.034
22	0.006	0.673	0.044	0.234	0.260	0.165
23	0.093	0.534	0.100	0.028	0.268	0.004
24	0.217	0.352	0.304	0.018	0.072	0.232
25	0.076	0.208	0.002	0.015	0.001	0.781
26	0.285	0.539	0.019	0.049	0.037	0.436
27	0.050	0.360	0.150	0.062	0.320	0.012
28	0.033	0.059	0.591	0.211	0.013	0.276
30	0.037	0.570	0.031	0.252	0.127	0.036

TABLE V. Here the absolute values of the loadings 0.340 and higher are shown for the model of Table IV. Three questions (8, 9, 12, and 14) had loadings below this value. Also, the question ordering reflects the ordering of the loading values. Here one sees an identification of all but one of the six factors with Newtonian concepts. Note that the third law factor includes all FCI questions that an expert would consider to be associated.

	FACTORS					
	Centripetal force/First law	Identify forces, dynamics	Third law	Kinematics	Uncertain	First law given constant velocity
Item	1	2	3	4	5	6
6	0.956					
7	0.657					
18		0.852				
13		0.804				
5		0.768				
11		0.680				
22		0.673				
30		0.570				
26		0.539				0.436
23		0.534				
10		0.399				
27		0.360				
24		0.352				
15			0.819			
4			0.782			
28			0.591			
16			0.546			
19				0.737		
20				0.647		
1	0.352				0.378	
21					0.344	
25						0.781
17						0.606

maturing on the part of the students when it comes to the third law.

The fourth factor just has the two kinematics items, 19 and 20, loading. As said in describing the pretest case, our students are not necessarily familiar with the form of presentation of these questions. More students did answer correctly for each of these questions than in the case of the pretest (about the item average for both questions). Still, it is difficult to say that the association between these questions is based on an understanding of kinematics. As no other items load onto this factor (as did not happen for the case of the pretest), we do see a positive development in the factor model as these two questions are unique in content and presentation style for the FCI.

Factor 5 has items 1 and 21 with comparatively weak loading values when considering the item loadings of the pre- and post-test factor models. Question 1, which concerns two freely falling objects of unequal mass, also loads (with a lower value) onto the first factor as we have discussed. Question 21 asks about the trajectory of a rocket having its initial flight path being affected by the application of a force over a finite period of time. Both items ask about an object under the influence of constant force. Beyond this, we do not see why there is any significant association between the two, which would suggest any maturation in physical thinking. There are other FCI questions with which one could argue that, at least, question 21 can find a more physically meaningful association.

Items 17, 25, and 26 form the sixth factor and were described. These three items did load together (among other items which ask the student to identify the forces acting on an object) on a single factor for the pretest model. Also, in the pretest case, recall that question 26 double loaded with a factor having only kinematics questions. For the post-test model, questions 25 and 17 load more heavily than does question 26 and both concern the first law in a particular way. In both questions, an object is moving with a constant velocity and the student must determine how the forces being applied against and along the direction of motion relate. Question 26 is a follow-up to question 25 and asks about the speed of the object of question 25 if one of the applied forces is doubled. Besides their connection in this simple way (which may explain the weak correlation), questions 25 and 26 are conceptually different in that, as an expert could argue, question 25 pertains to the first law and question 26 to the second law. Given the relatively weak

loading value of question 26, we see this factor as primarily related to the first law. In addition, question 26 loads (more strongly) onto the second factor described above, where it can be seen to better fit conceptually. It should be noted that for the pre- and post-test cases, the majority of students did answer the three items of factor 6 incorrectly (yet performing better on the post-test).

This is a good place to stress again that the factor analysis is not about how many items students answer correctly, but about the correlations among the items found in any patterns generated through student responses. So, here we again find patterns suggestive of conceptual associations students made whether or not they responded correctly to the questions. We feel that students, as they build their physical understanding, may develop associations among physical concepts indicative of a maturation process even if their understanding is not yet fully developed. It is these associations we hope to detect through the factor analysis.

D. Factor pattern evolution

Compared to that of the pretest case, the factor pattern of the post-test case has developed a form more easily interpreted in terms of Newtonian ideas. By this, we claim that, after instruction, several factors pertain to a more specific mechanical concept or set of concepts which suggests some level of maturing of the students' problem solving process. Correlations made among the test items may imply that more general physical ideas common to the items are being recognized. Such a development may be indicative of thinking more like that of an expert with force concepts [24]. In approaching a problem, experts can be seen to "abstract physics principles" as opposed to novices who tend to base their solution process on a problem's "literal features" [24].

In the factor model for the pretest, as mentioned, the first factor has many items for which a trajectory to be followed by some object is a common feature (including diagrammatically). Several also deal with a sudden change in force. Yet, these problems present different fundamental mechanical considerations for determining the correct trajectory or the effect of changing a force. In this way, one might claim that students are engaged with the literal features of these problems. Our one-on-one interview results support this. Two of this factor's items, 6 and 7, with strong loadings (along with item 1, which loads weakly), form the first factor of the post-test model. Although these problems are quite similar in presentation style, they both concern the same specific principles which an expert would associate. Moreover, for the pretest model, these problems loaded rather weakly among questions with the characteristics explained earlier. The prominent loadings these items take for the post-test model seem to indicate a recognition of these items' deeper mechanical connection. One might suppose that these questions are superficially related through repeated exposure during instruction, however, the one-on-one interview results suggest otherwise.

A similar discussion can apply for items 17, 25, and 26, which are associated in a rather superficial sense with other questions in the pretest model. Yet, they form their own factor in the post-test case. As described, items 17 and 25 relate to the first law and item 26 to the second law with item 26 being a follow up to item 25. Here the relation between items 17 and 25 can be seen to be one based on a specific principle, and item 26 is closely associated in principle to item 25.

Questions related through the third law, 4, 15, and 28, make up their own factor in the case of the pretest. Item 16 should also be correlated to them given an expert view. This does occur in the post-test model, and these items form a single factor. An expert would see these four as the only problems forming this factor for the FCI. As mentioned, items 15 and 16 differ in nuance but would still be seen as (fundamentally) third law problems.

At this point, note that we did not confirm the findings of Scott *et al.* [5] in which question 16 is found to load onto both a first and third law factor, suggesting that students may be confusing these two concepts. Recall, this question concerns a compact car moving at a constant speed on a level road and pushing a large truck that has broken down. It asks what is true about the force the car applies to the truck and that which the truck applies to the car during the push. Given our past work, our students do not appear to have the issues with this question as seen by those authors.

Problems 19 and 20 loaded together with problem 26 in the pretest model. It is interesting that problems 19 and 20 form their own factor in the post-test case. These two kinematics questions are certainly unique in presentation and would be associated by an expert. Given this, it is difficult to tell if students recognize the physical content they share or if the features of their presentation prompt the correlation.

Finally, besides factor 5, which we find difficult to interpret, the post-test model has factor 2 concerning force identification and dynamics. The associations involving identifying forces could be seen as aligned with an expert view, yet considering all the items of this factor, a novice character appears to persist.

V. CONCLUSION

Exploratory factor analysis has given us insight as to the possible conceptual associations made by students as a result of instruction. We do not necessarily see this as a tool to determine gains or losses in conceptual understanding. Simply, the relations found among FCI questions of clearly identifiable physical content offer insight into student thinking. Certainly, the path to becoming an expert is difficult to explore [25]. Still, the results of our analysis can be seen to suggest an evolution in the students' problem solving process towards that of an expert. We ask whether the associations made by students are indicative of a mind

on the path to grasping many of the invaluable subtleties so important to deepening one's physical understanding. Also, given certain loading or correlation values or lack of interpretability, this analysis can help justify examining certain questions as to their effectiveness in testing students' abilities in identifying certain physical concepts.

Also, for the future, we'd like to perform a gender specific analysis of this kind. We wonder if this might grant us some insight into the well-known gender gap concerning the FCI [26].

ACKNOWLEDGMENTS

The authors offer great thanks to Susan Hutchinson, Professor of Statistics, Applied Statistics, and Research at the University of Northern Colorado, and Cynthia Galovich, Chair of the Department of Physics and Astronomy also at the University of Northern Colorado, for many thoughtful discussions concerning this project. Also, we are grateful to the reviewers for an extensive critique and extremely helpful suggestions.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[2] D. Huffman and P. Heller, What does the force concept inventory actually measure?, Phys. Teach. **33**, 138 (1995).
[3] D. Hestenes and I. Halloun, Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller, Phys. Teach. **33**, 502 (1995).
[4] P. Heller and D. Huffman, Interpreting the force concept inventory: A reply to Hestenes and Halloun, Phys. Teach. **33**, 503 (1995).
[5] T. F. Scott, D. Schumayer, and A. R. Gray, Exploratory factor analysis of a Force Concept Inventory data set, Phys. Rev. ST Phys. Educ. Res. **8**, 020105 (2012).
[6] H. H. Harman, *Modern Factor Analysis*, 2nd ed. (The University of Chicago Press, Chicago, Illinois, 1976).
[7] B. O. Muthén, Exploratory Factor Analysis. Retrieved from http://www.statmodel.com/discussion/messages/8/8.html (2001).
[8] R. L. Gorsuch, *Factor Analysis* (Psychology Press, East Sussex, 2008), pp 295–297.
[9] S. Hutchinson (private communication).
[10] J. S. Uebersax, Introduction to the Tetrachoric and Polychoric Correlation Coefficients, and references therein (2015). Retrieved from http://www.john-uebersax.com/stat/tetra.htm#ex1.
[11] L. K. Muthén and B. O. Muthén, *Mplus User's Guide*, 6th ed. (Muthén & Muthén, Los Angeles, CA, 1998).
[12] H. F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. **20**, 141 (1960).
[13] J. L. Horn, A rationale and test for the number of factors in factor analysis, Psychometrika **30**, 179 (1965).
[14] L. E. Garrido, F. J. Abad, and V. Ponsoda, Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation, Psychol. Methods **21**, 93 (2016.
[15] R. D. Ledesma and P. Valero-Mora, Determining the Number of Factors to Retain in EFA: an easy-to use computer program for carrying out Parallel Analysis, Pract. Assess. Res. Eval. **12**, 2 (2007).
[16] A. A. Cota, R. S. Longman, R. R. Holden, G. C. Fekken, and S. Xinaris, Interpolating 95th percentile eigenvalues from random data: An empirical example, Educ. Psychol. Meas. **53**, 585 (1993).
[17] R Core, Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (2013).
[18] A. A. Beaujean, Factor Analysis using R, Pract. Assess. Res. Eval. **18**, 1 (2013).
[19] B. O. Muthen, Non-positive-definite tetrachoric correlation matrix, www.statmodel.com/discussion/messages/23/104.html?1006875516 (2001).
[20] Stata FAQ: What are the saturated and baseline models in sem? UCLA: Statistical Consulting Group. Retrieved from http://www.ats.ucla.edu/stat/stata/faq/sem_baseline.htm (2015).
[21] M. F. Triola, *Elementary Statistics*, 12th ed. (Pearson Education, Inc., Boston, MA, 2014).
[22] D. Hooper, J. Coughlan, and M. Mullen, Structural equation modelling: guidelines for determining model fit, Electron. J. Bus. Res. Meth. **6**, 53 (2008), and references therein.
[23] S. Cho, F. Li, and D. Bandalos, Accuracy of the parallel analysis procedure with polychoric correlations, Educ. Psychol. Meas. **69**, 748 (2009).
[24] M. T. H. Chi, P. J. Feltovich, and R. Glaser, Categorization and representation of physics problems by experts and novices, Cogn. Sci. **5**, 121 (1981).
[25] C. Wieman, The "Curse of Knowledge," or why intuition about teaching often fails, *The Back Page* (APS News, College Park, MD, 2007), Vol. 16, https://www.aps.org/publications/apsnews/200711/backpage.cfm.
[26] J. Docktor and K. Heller, Gender differences in both force concept inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).