Personality types and student performance in an introductory physics course: part 2

Jason J.B. Harlow,[1] David M. Harrison,[1] Michael Justason,[2]
Andrew Meyertholen,[1] Carolyn Sealfon,[1] and Brian Wilson[1]

[1]*Dept. of Physics, Univ. of Toronto, Toronto, ON M5S 1A6, Canada*
[2]*Faculty of Engineering, McMaster Univ., Hamilton, ON L8S 0A3, Canada*

We previously measured large correlations between personality type and student success in a first-year physics course for life science students. We now report results from a follow-up study carried out in the fall of 2017. We found similar relationships between personality type and success as measured by test and exam performance and scores on the Force Concept Inventory (FCI). Furthermore, we saw that performance did not depend on whether student learning teams were heterogeneous or homogenous in terms of personality type. Lastly, there were significant correlations between measured personality type and student motivation, gender, and attitudes towards physics.

I. INTRODUCTION

In the Fall of 2016 we measured the personality type of the students in a 800-student calculus-based introductory course primarily intended for students in the life sciences using the True Colors personality test, and looked for correlations with student performance as measured by test and exam performance and scores on the Force Concept Inventory (FCI).[1] Some of us were initially skeptical that we would observe any correlation between personality type and student performance, and were therefore surprised that we saw large correlations. Two personalities in particular, Green (characterized as a philosophical idea person) and Blue (characterized as a passionate mediator), had statistically significant differences of about five standard deviations in most course assessments. Perhaps most compelling was the fact that there were also differences in learning assessment gains on the Force Concept Inventory (FCI) . Furthermore, we found physics faculty and graduate students to have a very different personality profile from these students, leaning towards green. Our conclusion was that it's possible the different research-based learning techniques we are using may not be as effective for some particular students.

That study was for the Fall 2016 session. Here we report on the Fall 2017 session of the same course. The main goal of this follow-up study was to better understand and hopefully develop methods to reduce the performance gap that was observed. One hypothesis is that students with similar personalities could benefit from interacting more with students with different personality types. To avoid needless duplication, below we assume some familiarity with the concepts and methods of Ref. 1.

Students spend 2 hours per week in classes, in which Peer Instruction and Interactive Lecture Demonstrations are used extensively. In addition they spend 2 hours per week in a combined tutorial/laboratory which we call *Practicals*, in which they work in teams of 4 on activities based on a guided-discovery model of instruction. Each Practical group consisted of up to 9 such teams.

To test our hypothesis we assigned students to Practical teams according to personality. For one-half of the groups we formed each Practical team so that it consisted of students with roughly the same measured personality type ("sorted"), while for the other half each team consisted of a mixture of different measured personality types ("spread"). We then looked for correlations in performance with what type of team the students were in. We also examined satisfaction as measured by the end of term evaluation and whether it was different for different types of teams.

II. METHODS

As already mentioned, Ref. 1 discusses results from the Fall 2016 session of the course, and here we concentrate on the Fall 2017 session. The structure and pedagogy were similar in the two sessions but with a few changes. First, in the Fall 2016 session BW did the classes while in the Fall 2017 session BW did the first half and JJBH did the second half of the classes. Second, in the earlier session the Peer Instruction component of the classes used conventional clickers and therefore all ConcepTests were in multiple-choice format. In the later session we changed to Pearson's *Learning Catalytics*, which uses smart phones or tablets or laptops and allows other formats for the ConceptTests.[2] Third, as discussed in sub-section D below, we formed the teams in the Practicals in two different ways based on the students' color type. Finally, in the classes we implemented some of the changes in content and presentation discussed in Ref. 1; these changes involved only a small fraction of the total class time.

In the work of Ref. 1, we used the 15-question half FCI for both the pre-course and post-course assessment. Here we used the full 30-question FCI instrument. As already mentioned, here we used the True Colors instrument in both the pre-course and post-course. The final part of the assessment was some questions about the students, their background, motivation for taking the course, and some questions about their attitudes about physics; the questions were somewhat different on the pre-course and the post-course assessment and are given along with the student responses in the Results section. In addition, inserted in the middle of the assessment was a question to insure that the

students were actually reading the questions: this reduced the number of students in the pre-course sample by 3% on the pre-course and 5% on the post-course.

Sub-section A below describes a new method of measuring and classifying student personality types. Sub-section B discusses some measures of the validity and reliability of the True Colors test instrument. Sub-section C describes how we attempted to measure team effectiveness for different types of teams.

A. Measuring and Using Personality Type

The personality assessment was based on the same five groups of word clusters used in Ref. 1. However, the way in which the multiple-choice questions were formed was different than in that study. For example, for Group I, the questions are now:

Question 31: For Group I, which Set is **most** like you?
A B C D
Question 32: For Group I, which Set is **a lot** like you?
A B C D
Question 33: For Group I, which Set is **somewhat** like you?
A B C D
Question 34: For Group I, which Set is **least** like you?
A B C D

Even if two sets are roughly equal in terms of how well they describe you, please make your answers to Questions 31 – 34 all different. Similarly, please make your answers different for Groups II – V below.

There are similar questions for the other four groups of word clusters.

As discussed in Ref. 1, there were problems with the original form of the multiple-choice questions. For example, the total of the four color scores should be 50, which was true for just under one-half of the students. The new format does much better. For example, on the pre-course assessment 22 students of 822, ~3%, did not have total color scores of 50.

In Ref. 1 we assigned the four quadrants of an *x*-*y* plot to the four colors in the order Blue – Gold – Green – Orange, and discussed why this assignment was consistent with our observations. Then for each student we calculated a centroid of his or her color scores defined by:

$$x_C = \frac{\text{Blue+Orange-Gold-Green}}{4}$$
$$y_C = \frac{\text{Blue+Gold-Green-Orange}}{4} \qquad (1)$$

and used the quadrant where the centroid was located to assign a color type. About 10% of our students had a centroid that was on one or both of the axes so did not have an assigned color type.

We have shown that for all assessment instruments the Green-Blue difference correlated strongly with student performance. Therefore, we have formed a coordinate system of β (for "Blue") and γ (for "Gold") axes rotated by θ relative to the *x*-*y* axes, with β pointing from the Green to the Blue quadrant, as shown in Fig. 1.

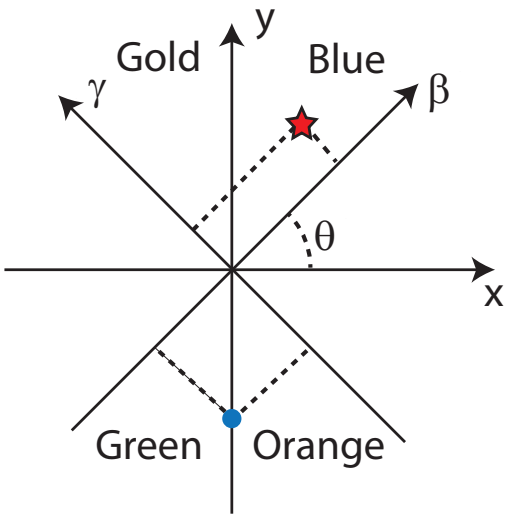

Fig. 1. Defining the β–γ coordinates

Then the coordinates of a centroid in the rotated coordinate system are given by:

$$\begin{aligned} \beta_C &= x_C \cos\theta + y_C \sin\theta \\ \gamma_C &= -x_C \sin\theta + y_C \cos\theta \end{aligned} \qquad (2)$$

Fig. 1 illustrates for two values of the centroid. The centroid of the student represented by the red star is in the Blue quadrant, and has a positive β component. The centroid of the student represented by the blue circle lies on the –*y* axis and previously was assigned to no color type. But this student has a negative β_C so we now assign the student to Green. One advantage of this method is that for all students we now have a numerical value for the strength of their Green-Blue values.

We examined the Spearman correlation coefficient[3] between the values of β_C and the pre-course FCI for different values of θ. The data, which are not shown, did not show significant differences for values between 40^o and 50^o. Therefore, we chose $\theta = 45°$.

With this choice, $\sin\theta = \cos\theta = 1/\sqrt{2}$ and the β_C coordinate is:

$$\beta_C = \frac{x_C + y_C}{\sqrt{2}} \tag{3}$$

From Eqn. 1, Eqn. 3 becomes:

$$\beta_C = \frac{\text{Blue} - \text{Green}}{2\sqrt{2}} \tag{4}$$

This does not depend on the Gold and Orange scores. For any angle other than 45° this is not true. This is the value that will be used below. In the literature β_C is sometimes referred to as the Thinking-Feeling or TF axis.[4]

Fig. 2 illustrates for a real student whose color scores were (Blue, Gold, Green, Orange) = (5, 10, 15, 20). The scores are the blue diamonds, and the centroid is the red star. The centroid lies on the –*y* axis with y_C = -5.0.

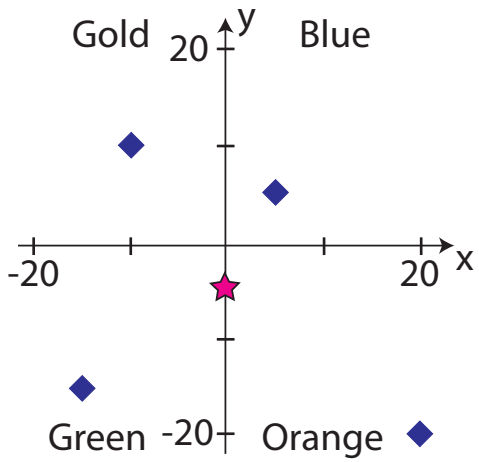

Fig. 2. Colors scores, the blue diamonds, and the centroid, the red star, for a student.

The β_C component of the centroid is -3.54, so the student is Green. This seems completely reasonable. The minimum possible color score is 5, so this student could not be less Blue.

For analysis we defined "cut points" of values of β_C so the class was divided into three groups. The lower cut point was -1.25 and the higher one was +1.25. The members of the group with $\beta_C \geq +1.25$ are the most strongly Blue students, and below we call this group of students Blue$_\beta$; the group with $\beta_C \leq -1.25$ we similarly call Green$_\beta$. Since in an RGB additive color scheme green + blue = cyan, the students with $-1.25 < \beta_c < +1.25$ we call Cyan$_\beta$. The cut point values were chosen to attempt to balance two conflicting goals. On the one hand we want all three groups to have about the same number of students; this would require cut points of -1.77 and +1.06. On the other hand, we want the cut points to be symmetric about $\beta_C = 0$.

Table I shows the number and percentage of students in each of the 3 groups for the pre-course. Because of rounding errors the percentages do not add up to exactly 100.

Table I. The number and percentage of students in each group for the pre-course assessment

$Blue_{\beta}$	$Cyan_{\beta}$	$Green_{\beta}$
247 (30%)	253 (32%)	322 (39%)

Figure 3 is a histogram of the values of β_C. The vertical red dashed lines are the cut points.

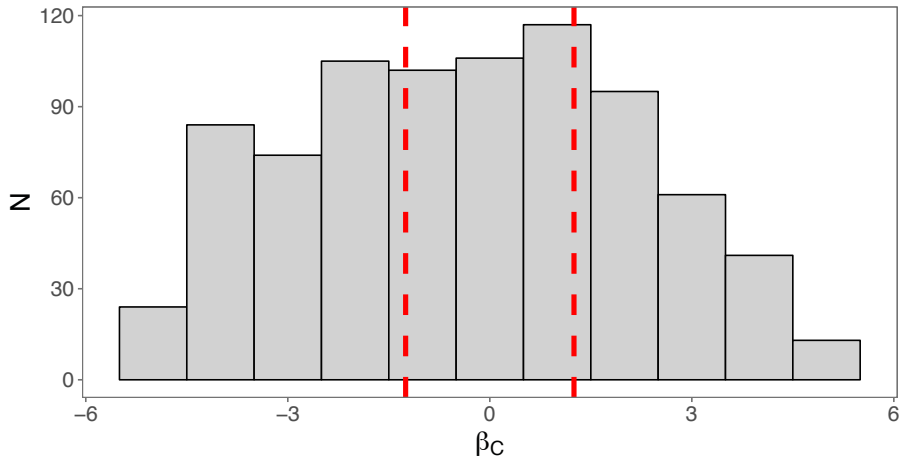

Figure 3. Values of β_C for the pre-course assessment. The vertical red dashed lines are the cut points.

As discussed, the color assessment was given twice, once at the beginning of the term and again at the end of the term. All analysis except for that in the next sub-section is based on the pre-course assessment.

B. Reliability and Validity of the True Colors Instrument

Although as discussed in Ref. 1, the fact that the True Colors assessment correlates so strongly with student performance is the central fact, some analysis of the reliability and validity of the test instrument is possible from the data. First, as in Ref. 1, we calculate the Cronbach α for each of the five measurements of the four colors separately. The results are shown in Table II. The values are comparable to those in Ref. 1.

Table II. Cronbach α for the Pre-Course and the Post-Course Assessment

Color	Pre-Course	Post-Course
Blue	0.76	0.77
Gold	0.70	0.60
Green	0.73	0.66
Orange	0.71	0.68

These values are less than those reported for the full MBTI instrument.[5] The True Colors instrument that we used consists of five ranking tasks, so it is reasonable to take the number of questions to be 5 although there were actually 20 questions on this part of the assessment. By comparison, the full MBTI consists of 93 questions. In general measures of reliability go up with the number of questions although if the number of questions becomes too large, fatigue begins to become important. The relation between the number of questions and the reliability can be estimated by the Spearman-Brown prophecy formula:[6]

$$\alpha_{\text{new}} = \frac{m\alpha_{\text{old}}}{1+(m-1)\alpha_{\text{old}}} \tag{5}$$

where m is the length of the new test divided by the length of the old test. This can be re-written as:

$$m = \frac{\alpha_{\text{new}}(\alpha_{\text{old}}-1)}{\alpha_{\text{old}}(\alpha_{\text{new}}-1)} \tag{6}$$

For example, to raise the reliability from 0.75 to 0.90, which is comparable to the values for the full MBTI instrument, gives $m = 3$ so would require about 15 different sets of 4 clusters of words instead of the 5 sets in the True Colors assessment.

The pre-course and post-course assessments were given 11 weeks apart. Comparing color results for these two assessments can give some indication of the reliability of our form of the True Colors assessment. However, 70% of our students are in their first year of university, so are typically 18 or 19 years old. A further 14% of the students are in their second year and are typically one year older. As discussed more fully in Ref. 1, we believe that the measured personality type is in part due to the student's self-image, which they may have acquired from what they have been told by their parents, peers, or former teachers. Physiologically, the pre-frontal cortex in the brain has a large role to play in the development of personality, and is still developing until age 25 or so.[7]

Therefore we might expect that at this relatively young age, in the 11 weeks between the test and the re-test at least some of our students will have changed their measured personality type independent of the reliability of the test instrument. Table III shows the Spearman rank correlation coefficients for the pre-course and post-course assessment for the 610 students who did both the pre-course and the post-course assessment. Correlations > 0.90 are heuristically often called "very high", correlations between 0.70 and 0.90 are "high", 0.50 to 0.70 are "moderate", 0.30 to 0.50 are "low", and values < 0.30 are "negligible".[8]

Table III. Pre- and Post-Course Spearman Correlations

Compared	Correlation
Total Blue Score	0.66
Total Gold Score	0.68
Total Green Score	0.70
Total Orange Score	0.72
β_C	0.71

Looking at the assignment of students into one of three colors, Blue_β, Cyan_β, and Green_β, 377 = 62% of the 610 students did not change their color from the pre-course to the post-course assessment. Table IV shows the changes for students who did change their color type.

Table IV. Number of students who change their color type from the pre-course to the post-course assessment.

Change	Number
Blue_β to Cyan_β	41
Blue_β to Green_β	6
Cyan_β to Green_β	52
Green_β to Cyan_β	59
Green_β to Blue_β	18
Cyan_β to Blue_β	57

Only 6 students changed from Blue_β to Green_β and 18 from Green_β to Blue_β. The remainder of the changes were from one color type to its nearest neighbor.

Table V shows the color assignments based on the post-course assessment. The percentages are almost identical to those from the pre-course assessment shown in Table I in the previous sub-section.

Table V. The number and percentage of students in each group based on the post-course assessment

Blue_β	Cyan_β	Green_β
189 (31%)	184 (30%)	237 (39%)

C. The Effectiveness of Different Kinds of Teams

As mentioned in the Introduction, in the Practicals each group consists of about 36 students, divided into 9 teams of 4 students each. We formed the teams in two different ways, which we call "spread" and "sorted." For the spread method, which we used for about one-half of the groups, we assigned team numbers, 1, 2, 3, 4, 5, 6, 7, 8 and 9 to the nine most Blue students based on their β_C value. Then the next nine students were also

assigned to teams 1 though 9, and so on. So these teams contain students with a mixture of different personality types. For the sorted method, used for the other half of the groups, we assigned students with the largest four β_C values to team number 1, then the next four to team number 2, and so on; all four students with the lowest β_C values we assigned to team number 9. These teams have roughly the same personality type. In order to avoid biases, we did not inform either the students or their TA Instructors that we had constructed the teams this way. This method of assigning teams is similar to a previous study we did on team effectiveness, except that there we assigned students to teams based on the Pre-Course FCI scores.[9]

We measured student performance by looking at the scores on the pre-course FCI, the two term tests, the post-course FCI, the final examination, and the normalised gain on the FCI for the different types of teams. As in Ref. 1, the normalised gain is:

$$< g >_{\text{median}} = \frac{< \text{Post-Course FCI \%} > - < \text{Pre-Course FCI \%} >}{100 - < \text{Pre-Course FCI \%} >} \qquad (7)$$

where the angle brackets on the right hand side indicate medians. We only calculated the gain for "matched" students who had taken both the pre-course and the post-course FCI.

We also looked at student satisfaction as measured by an end of term evaluation of the Practicals, which was separate from the post-course assessment.

As discussed more fully in Ref. 9, the students meet for 2 hours per week in the Practicals in teams, and also meet for 2 hours per week in the classes. This could weaken any measured correlations of team type with student performance since they could be spending half of their time with students who were not in their team. Therefore for the session studied here we urged the students to sit with their Practical teammates during the classes, and in the post-course assessment asked them whether they had done what we suggested. As seen in the Results section, most students ignored our suggestion.

III. RESULTS

In sub-section A below we report on measures of student performance and correlations with color type. Then in sub-section B we give the results of the final part of the pre-course and post-course assessments, which is about the students, their background, motivation for taking the course, and attitudes towards physics. Sub-section C discusses student performance as a function of the type of team they were in. Sub-section D discusses an end-of-term evaluation of the Practicals..

A. Measures of Student Performance

Here we report on student performance on the pre-course FCI, the two terms tests, the post-Course FCI, the final examination, and the normalised gain on the FCI.

Analyzing the data using the same methods as in Ref. 1 gave similar results. For example, Table VI compares the mean values of the centroids as reported in Ref. 1 and the values for the session studied here. Both are from the pre-course assessment. The uncertainties are $\sigma_m = \sigma / \sqrt{N}$. Figure 4 is the scatter plot of the centroids for the session studied here, which looks very similar to Fig 3 of Ref. 1.[10] The open circle is the mean of the centroids. Points to the left and below the black dashed line have a negative β_C value, and points to the right and above have a positive one. The blue dotted lines are the cut points used to define the 3 groups of students.

Table VI. The mean values of the centroids

	$\overline{x}_C$	$\overline{y}_C$
Session of Ref. 1	-0.97 ± 0.06	0.65 ± 0.06
Session studied here	-1.04 ± 0.08	0.57 ± 0.08

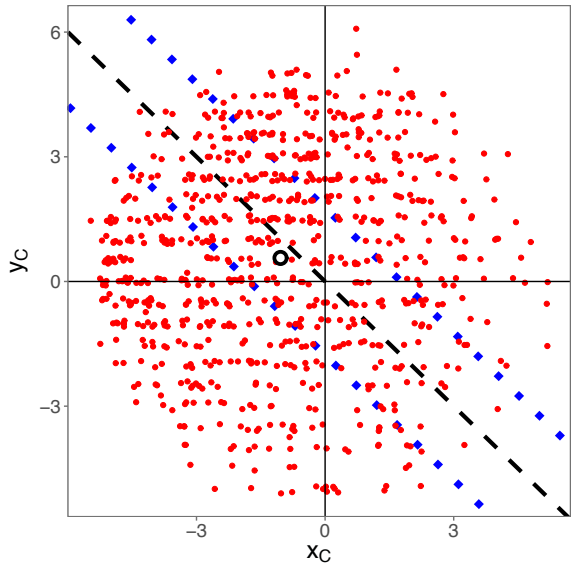

Figure 4. The centroids of the students. The open circle is the mean of the centroids.

Another example of the methods of Ref. 1 involves the median of the scores on the pre-course FCI in percent, where the student color type is determined by the quadrant of the location of the centroid. Table VII compares the results for the earlier session to the one studied here. Recall that the earlier session used the 15-question half FCI, while the later one used the full 30-question FCI.

Table VII. Median Pre-Course FCI Scores in Percent Using the Method of Assignment of Color Type Used in Ref. 1

	Session of Ref. 1	Session studied here
Blue	35.7 ± 4.4	40.0 ± 4.0
Gold	50.0 ± 2.8	50.0 ± 3.3
Green	57.1 ± 4.8	60.0 ± 5.2
Orange	50.0 ± 5.4	46.7 ± 6.7

We do not present the other analysis that we did using the methods of Ref. 1, in part because, as already discussed, the pedagogy has changed somewhat. Instead in Table VIII we present assessment results using the 3 color groups based on the students' β_C score. All values are in percent. For the pre-course and post-course FCI, since the distributions are not Gaussian we report the median instead of the mean values, with an uncertainty of $1.58 \times \mathrm{IQR} / \sqrt{N}$, where IQR is the interquartile range and N is the number of students in the sample. This is roughly a measure of the 95% confidence interval.

Figure 5 is a histogram of the grades on Test 1. It is clearly not Gaussian either so, as with the FCI scores, the median is more appropriate than the mean in characterizing the results, and those are what are reported in Table VIII. For consistency, then, all other assessments are similarly reported using the median; it can be argued that this measure is preferred to the mean even for bell shaped distributions.[11]

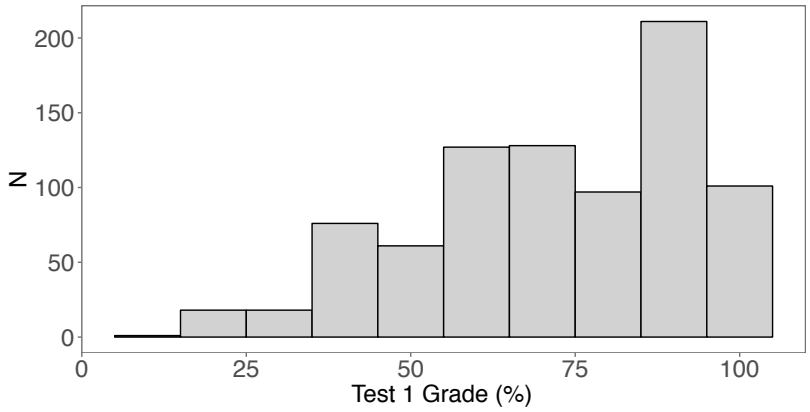

Figure 5. Test 1 Grades

The $\langle g \rangle_{\text{median}}$ is for "matched" students who wrote both the pre-course and the post-course FCI.

Table VIII. Scores on the FCI and grades on the tests and final exam.

	N	All	$Blue_\beta$	$Cyan_\beta$	$Green_\beta$	$Green_\beta$ - $Blue_\beta$
Pre-Course FCI	822	50.0 ± 2.2	40.0 ± 3.0	46.7 ± 4.0	60 ± 3.8	20.0 ± 4.8
Test 1	739	75.0 ± 1.7	71.4 ± 3.5	75.0 ± 3.0	85.7 ± 2.3	14.3 ± 4.2
Test 2	683	60.7 ± 1.4	57.1 ± 2.5	60.7 ± 2.7	67.9 ± 2.4	10.8 ± 3.5
Post-Course FCI	668	73.3 ± 2.1	60.0 ± 3.7	73.3 ± 3.0	80.0 ± 3.0	20.0 ± 4.8
Final Exam	659	59.5 ± 1.5	52.4 ± 2.8	59.5 ± 2.3	66.7 ± 2.3	14.3 ± 3.6
Normalised FCI Gain $\langle g \rangle_{median}$	610	0.43 ± 0.05	0.29 ± 0.09	0.47 ± 0.08	0.45 ± 0.10	0.16 ± 0.13

Recalling that the uncertainty in the final column of Table VIII corresponds roughly to a 95% confidence interval, i.e. equivalent to $2 \times \sigma_m$ for a normal distribution, as in Ref. 1 we characterize the uncertainty in the values of $Green_\beta$ - $Blue_\beta$ by the number of "standard deviations." These are: Pre-Course FCI ~ 8, Test 1 ~ 7, Test 2 ~ 6, Post-Course FCI ~ 8, Final Exam ~ 8, Normalised FCI Gain ~ 2. These values are comparable to but somewhat greater than the values using the previous method of assigning colors in Ref. 1.

Table IX shows the values of Cliff's δ comparing the $Green_\beta$ and $Blue_\beta$ scores. We also give a heuristic characterization of its size, and the 95% confidence interval. Since none of the confidence interval ranges include zero, the differences can be taken as statistically significant.

Table IX. Cliff's δ comparing the $Green_\beta$ and $Blue_\beta$ scores.

	δ	95% CI
Pre-Course FCI	0.299 (small)	0.207 – 0.386
Test 1	0.311 (small)	0.213 – 0.403
Test 2	0.283 (small)	0.180 – 0.380
Post-Course FCI	0.363 (medium)	0.254 – 0.462
Final Examination	0.363 (medium)	0.260 – 0.457

Figure 6 is the boxplot for the Pre-Course FCI scores. The small dot is the mean, and the associated "error bars" are σ_m. Note that because the distribution of scores is not Gaussian, the mean and its uncertainty are not particularly good measures of the distribution. In this case the asymmetry of the upper and lower quartile values compared to the median, the skewness, causes the mean value to be different than the median.

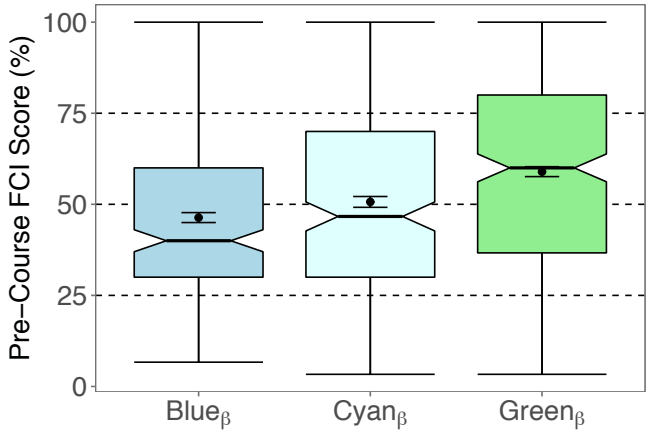

Figure 6. Boxplot of Pre-Course FCI scores for the 3 color groups.

Figure 7 is the boxplot for the results of Test 1. As in Ref. 1, the dots outside the "whiskers" represent datapoints that are considered to be outliers.

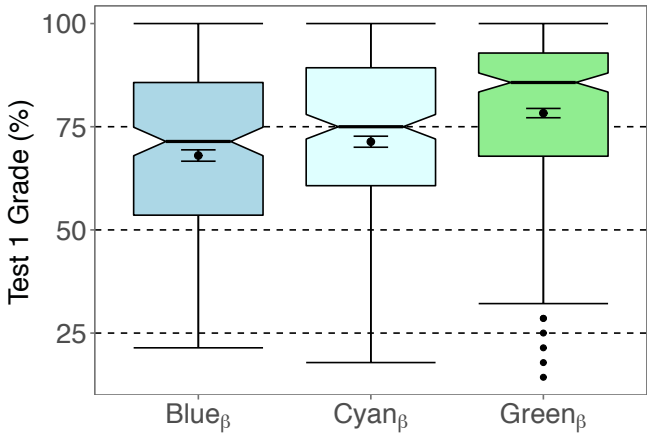

Figure 7. Boxplot of the Test 1 grades for the 3 color groups.

The boxplots for the other assessments did not provide any further insight, and are not shown.

Comparing the results in this sub-section to those of Ref. 1 it is clear that assigning color types based on the centroid projections onto the β axis as here is at least as effective as the previous methodology.

B. Student Background, Motivation, and Attitudes

First we discuss the results of Part 3 of the pre-course assessment for all students and for the three color groups, and then discuss the post-course assessment. Note that for the pre-course assessment for some questions, the total number of students is not equal to the

number of students in the dataset, 822. This is because some students answered the question with a choice that is not included in the question. The stated uncertainties assume that the uncertainty in the number of students for each answer is the square root of that number. Because of rounding errors, not all percentages in all columns add to 100. Question 52, the check of whether the students were reading the questions, is not shown.

Pre-Course Question 51: What is your intended or current Program of Study (PoST)?

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. Life Sciences	$642 = (78 \pm 3)\%$	$193 = (78 \pm 6)\%$	$203 = (80 \pm 6)\%$	$246 = (76 \pm 5)\%$
B. Physical and Mathematical Science	$91 = (11 \pm 1)\%$	$27 = (11 \pm 2)\%$	$26 = (10 \pm 2)\%$	$38 = (12 \pm 2)\%$
C. I haven't decided yet	$28 = (3 \pm 1)\%$	$9 = (4 \pm 1)\%$	$10 = (4 \pm 1)\%$	$9 = (3 \pm 1)\%$
D. Other /NA[12]	$61 = (7 \pm 1)\%$	$18 = (7 \pm 2)\%$	$14 = (6 \pm 1)\%$	$29 = (9 \pm 2)\%$

Pre-Course Question 53: What is the main reason you are taking PHY131?

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. It is required	$263 = (32 \pm 2)\%$	$99 = (40 \pm 4)\%$	$86 = (34 \pm 4)\%$	$78 = (24 \pm 3)\%$
B. For my own interest	$133 = (16 \pm 1)\%$	$38 = (15 \pm 2)\%$	$36 = (14 \pm 2)\%$	$59 = (18 \pm 2)\%$
C. Both because it is required and for my own interest	$379 = (46 \pm 2)\%$	$90 = (36 \pm 4)\%$	$122 = (49 \pm 4)\%$	$167 = (52 \pm 4)\%$
D. Other / NA	$44 = (5.4 \pm 0.8)\%$	$20 = (8 \pm 2)\%$	$6 = (2 \pm 1)\%$	$18 = (6 \pm 1)\%$

Pre-Course Question 54: Did you take Grade 12 Physics or an equivalent course elsewhere?

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. Yes	$602 = (74 \pm 3)\%$	$174 = (70 \pm 5)\%$	$184 = (73 \pm 5)\%$	$244 = (77 \pm 5)\%$
B. No	$215 = (26 \pm 2)\%$	$73 = (30 \pm 4)\%$	$68 = (27 \pm 3)\%$	$74 = (23 \pm 3)\%$

Pre-Course Question 55: What is your gender?

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. Female	485 $=(59\pm3)\%$	166 $=(67\pm5)\%$	162 $=(64\pm5)\%$	157 $=(49\pm4)\%$
B. Male	332 $=(40\pm2)\%$	81 $=(33\pm3)\%$	88 $=(35\pm4)\%$	163 $=(51\pm4)\%$
C. Neither of these are appropriate for me	3 $=(0.4\pm0.2)\%$	$0=0\%$	2 $=(0.8\pm0.6)\%$	1 $=(0.3\pm0.3)\%$

Pre-Course Question 56: Encountering a new type of physics question or concept makes me feel:

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. Excited	114 $=(14\pm1)\%$	24 $=(10\pm2)\%$	38 $=(15\pm2)\%$	52 $=(16\pm2)\%$
B. Curious	405 $=(49\pm2)\%$	97 $=(39\pm4)\%$	115 $=(46\pm4)\%$	193 $=(60\pm4)\%$
C. Indifferent	50 $=(6\pm1)\%$	19 $=(8\pm2)\%$	17 $=(7\pm2)\%$	14 $=(4\pm1)\%$
D. Anxious	240 $=(29\pm2)\%$	103 $=(42\pm4)\%$	78 $=(31\pm4)\%$	59 $=(18\pm2)\%$
E. Irritated	12 $=(1.5\pm0.4)\%$	4 $=(2.0\pm0.8)\%$	5 $=(2.0\pm0.9)\%$	3 $=(0.9\pm0.5)\%$

Pre-Course Question 57: Nearly everyone is capable of understanding physics if they work at it.

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. Strongly disagree	69 $=(9\pm1)\%$	19 $=(8\pm2)\%$	21 $=(8\pm2)\%$	29 $=(9\pm2)\%$
B. Disagree	78 $=(8\pm1)\%$	23 $=(9\pm2)\%$	30 $=(12\pm2)\%$	25 $=(8\pm2)\%$
C. Neutral	179 $=(22\pm2)\%$	61 $=(25\pm3)\%$	51 $=(20\pm3)\%$	67 $=(21\pm3)\%$
D. Agree	355 $=(44\pm2)\%$	106 $=(43\pm4)\%$	117 $=(46\pm4)\%$	132 $=(41\pm4)\%$
E. Strongly agree	136 $=(17\pm1)\%$	36 $=(15\pm3)\%$	33 $=(13\pm2)\%$	67 $=(21\pm3)\%$

There are significant differences by color for the reason for taking the course (Pre-Course Question 53), gender (Pre-Course Question 55), and student reactions to a new physics question or concept (Pre-Course Question 56).

Previously we have found correlations of student performance with their reason for taking our course.[13] We have also found correlations with gender,[14] a result that has been widely duplicated. Perhaps surprising is that here there is no measured correlation of measured personality type with whether the student had taken a senior-level high school physics course, although in Ref. 13 we found correlations in performance for this factor.

The questions in the final part of the post-course assessment were somewhat different than on the pre-course one. The questions about the students, their background, and motivation for taking the course are not repeated. Here are the questions we asked and the results. Post-Course Question 52, the check of whether the students were reading the questions, is not shown. Post-Course Question 51 was Pre-Course Question 56 on the pre-course assessment, and Post-Course Question 53 was Pre-Course Question 57 on the pre-course assessment. Post-Course Question 54 is new.

Post-Course Question 51: Encountering a new type of physics question or concept makes me feel:

	All	Blue_β	Cyan_β	Green_β
A. Excited	$82 = (13 \pm 1)\%$	$17 = (11 \pm 3)\%$	$27 = (14 \pm 3)\%$	$38 = (15 \pm 2)\%$
B. Curious	$233 = (38 \pm 3)\%$	$47 = (29 \pm 4)\%$	$66 = (34 \pm 4)\%$	$120 = (47 \pm 4)\%$
C. Indifferent	$76 = (12 \pm 1)\%$	$20 = (12 \pm 3)\%$	$26 = (13 \pm 3)\%$	$30 = (12 \pm 2)\%$
D. Anxious	$178 = (29 \pm 2)\%$	$56 = (35 \pm 5)\%$	$62 = (32 \pm 4)\%$	$60 = (23 \pm 3)\%$
E. Irritated	$41 = (7 \pm 1)\%$	$21 = (13 \pm 3)\%$	$12 = (6 \pm 2)\%$	$8 = (3 \pm 1)\%$

Post-Course Question 53: Nearly everyone is capable of understanding physics if they work at it.

	All	Blue_β	Cyan_β	Green_β
A. Strongly disagree	$68 = (11 \pm 1)\%$	$18 = (11 \pm 3)\%$	$21 = (11 \pm 2)\%$	$29 = (11 \pm 2)\%$
B. Disagree	$86 = (14 \pm 2)\%$	$24 = (15 \pm 3)\%$	$29 = (15 \pm 3)\%$	$33 = (13 \pm 2)\%$
C. Neutral	$133 = (22 \pm 2)\%$	$37 = (23 \pm 4)\%$	$32 = (17 \pm 3)\%$	$64 = (25 \pm 3)\%$
D. Agree	$233 = (38 \pm 3)\%$	$60 = (37 \pm 5)\%$	$79 = (41 \pm 5)\%$	$94 = (37 \pm 4)\%$
E. Strongly agree	$90 = (15 \pm 2)\%$	$22 = (14 \pm 3)\%$	$32 = (17 \pm 3)\%$	$36 = (14 \pm 2)\%$

Post-Course Question 54: It was suggested that in classes, not the Practicals, that you sit with your teammates from the Practicals. Did you do this?

	All	Blue_{β}	Cyan_{β}	Green_{β}
A. Usually or always	$49 = (8 \pm 1)\%$	$16 = (10 \pm 3)\%$	$19 = (10 \pm 2)\%$	$14 = (6 \pm 1)\%$
B. Sometimes	$43 = (7 \pm 1)\%$	$12 = (8 \pm 2)\%$	$13 = (7 \pm 2)\%$	$18 = (7 \pm 2)\%$
C. Rarely	$44 = (7 \pm 1)\%$	$13 = (8 \pm 2)\%$	$15 = (8 \pm 2)\%$	$16 = (6 \pm 2)\%$
D. Hardly ever or never	468 $= (77 \pm 4)\%$	118 $= (74 \pm 7)\%$	145 $= (76 \pm 6)\%$	205 $= (81 \pm 6)\%$

Note that Question 57 on the pre-course assessment, Question 53 on the post-course one, is from the Colorado Learning Attitudes about Science Survey.[15]

The color correlation with reactions to a new physics question or concept (pre-course Question 56 and post-course Question 51) remained about the same from the beginning to the end of the term.

The suggestion that students sit with their teammates during classes was made after there had been six classes, and most students ignored our advice. Students regularly ignore their professors, and they also can get very attached to the seat in the classroom where they have been sitting. Nonetheless, we were disappointed in this result.

C. Team Types and Student Performance

As discussed, the Practicals consist of 28 groups each with up to 36 students, and each group consists of up to 9 teams of about 4 students. In 14 of the groups students were assigned to teams so that all teams had a mixture of different personality types, the "spread" groups, and for the other 14 groups each team consisted of students with roughly the same personality type, the "sorted" groups. Table X summarises the student performance for these different types of groups. The pre-course FCI scores should be equal since at this time the teams had not yet been formed. When Test 1 was given 80% of the students had worked with their teammates only 3 times, and 20% had worked with their teammates 4 times. Similarly, when Test 2 was given, 80% of the students had met with their teams 7 times, and 20% had met 8 times.

The exact numerical equality of the values in many rows of the table is due in part to the fact that we are reporting medians of percentages, and that the values of the percentages are quantized. For example, the FCI has 30 questions, so the quantisation of FCI scores is $\hbar = 100/30 \cong 3.33\%$. Similarly, both term tests were graded out of 28 points so $\hbar = 100/28 \cong 3.57\%$.

Table X. Student performance for different types of groups.

	Sorted	**Spread**
Pre-Course FCI	50.0±3.1	50.0±3.3
Test 1	75.0±2.4	75.0±2.6
Test 2	60.7±2.0	60.7±2.1
Post-Course FCI	73.3±2.8	76.7±2.9
Final Examination	57.1±2.1	59.5±2.3
Normalised FCI gain $\langle g \rangle_{median}$	0.43±0.08	0.50±0.08

Table XI shows the student performance by student color type for the sorted groups. The final column shows the difference between $Green_{\beta}$ and $Blue_{\beta}$ students for the entire class, taken from Table VIII of sub-section A above. Table XII similarly shows the student performance by color type for the spread groups.

Table XI. Student performance by color type for sorted groups.

	$Blue_{\beta}$	$Cyan_{\beta}$	$Green_{\beta}$	$Green_{\beta}$ - $Blue_{\beta}$	Overall ($Green_{\beta}$ - $Blue_{\beta}$)
Pre-Course FCI	43.3±5.0	45.0±5.7	60.0±5.7	16.7±7.6	20.0±4.8
Test 1	67.9±4.3	75.0±4.1	82.1±3.2	14.2±5.4	14.3±4.2
Test 2	60.7±3.0	60.7±3.2	67.9±3.2	7.2±4.3	10.8±3.5
Post-Course FCI	63.3±5.0	70.0±4.8	83.3±4.0	20.0±6.4	20.0±4.6
Final Exam	52.4±3.9	57.1±2.8	66.7±3.1	14.3±4.6	14.3±3.6
Normalised FCI gain $\langle g \rangle_{median}$	0.31±0.12	0.44±0.11	0.54±0.13	0.23±0.18	0.16±0.13

Table XII. Student performance by color type for spread groups.

	$Blue_{\beta}$	$Cyan_{\beta}$	$Green_{\beta}$	$Green_{\beta}$ - $Blue_{\beta}$	Overall ($Green_{\beta}$ - $Blue_{\beta}$)
Pre-Course FCI	40.0±5.9	50.0±5.9	60.0±5.4	20.0±8.0	20.0±4.8
Test 1	71.4±4.9	75.0±4.8	85.7±3.3	14.3±5.9	14.3±4.2
Test 2	57.1±4.1	64.3±4.0	67.9±3.0	10.8±5.1	10.8±3.5
Post-Course FCI	60.0±5.8	76.7±3.9	80.0±3.7	20.0±6.9	20.0±4.6
Final Exam	52.4±4.0	59.5±3.1	66.7±3.4	14.3±5.2	14.3±3.6
Normalised FCI gain $\langle g \rangle_{median}$	0.29±0.13	0.50±0.11	0.50±0.12	0.21±0.18	0.16±0.13

There are no significant differences in student performance based on the type of team.

D. Practical Evaluation

At the end of the term, an anonymous evaluation of the Practicals was administered. Many of the questions were about the effectiveness of the Teaching Assistants, but five of them were specifically about the effectiveness and satisfaction with the Practicals. All of these questions were answered using a 5-point Likert scale, with 5 the most positive and 1 the least positive. We then averaged these five answers.

Although the evaluation is anonymous, we do know whether the students were in a sorted or spread team. The mean evaluation scores were essentially identical for the two types of teams: 3.74±0.03 for the sorted teams and 3.73±0.04 for the spread teams. Figure 8 shows the density plot[16] for these two types of teams: the distributions are nearly identical.

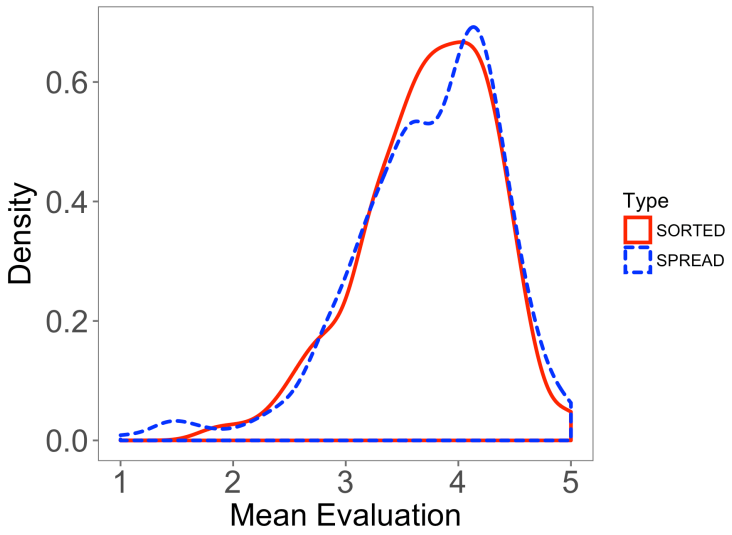

Figure 8. Density plot of the Practical evaluation for different types of teams

IV. DISCUSSION

As discussed in Ref. 9, various theoretical reasons have been proposed that argue in favor of student learning teams with a mix of different types. For personality type, our study does not indicate that this is true for our students. Student performance as measured by the tests, final examination, post-course FCI scores and normalised FCI gains was statistically identical for teams consisting of students with the same personality type as for teams with a mix of personality types. Student satisfaction with the part of the course involving these teams was also independent of the type of team.

In Ref. 9 we similarly found that student performance was independent of whether student learning teams were homogeneous or heterogeneous in terms of pre-course FCI scores,

Our null results are also consistent with a recent study of a general education science course by Pociask, Gross, and Shih.[17] Where we formed teams based only on the measured personality type or the pre-course FCI score, they formed mixed teams based on personality, year, major and gender. They also formed other teams randomly and others for which the students were allowed to choose their teams themselves. There was essentially no difference in performance for all three types of teams.

It was interesting to us to see a report that thinking about personality types using the Myers-Briggs Type Indicator, on which our True Colors instrument is based, can lead to dramatic improvements in classroom effectiveness.[18] However, effectiveness was measured only by student evaluations, which do not correlate with things like normalised FCI gains.[19]

In Physics Education Research there are a number of different diagnostic instruments. They all have somewhat different emphases and test for somewhat different things. For example, in assessing student conceptual understanding of classical mechanics, two of the best known are the FCI, used in this study, and the Force and Motion Conceptual Evaluation.[20]

Similarly, in assessing personality traits there are a number of different instruments. These include the Myers-Briggs Type Inventory discussed in Ref. 1, NEO which is briefly discussed in Ref. 1, the True Colors instrument that we use that is based on the Myers-Briggs one, Enneagram,[21] the DISC Assessment,[22] the Hermann Brain Dominance Instrument,[23] the Hogan Personality Inventory,[24] and the Thomas-Kilmann Conflict Mode Instrument.[25] However, these instruments are often criticized for not being particularly valid or reliable.[26]

One well-known way of characterizing cognitive traits in psychology is called Empathizing–Systemizing (ES) Theory. As described by its originator, Simon Baron-Cohen, in 2002, “empathising involves the attribution of mental states in others” while systemizing is an “entirely different [kind] of process” involving an inductive process to “form a rule about how [an] aspect of [a] system works.”[27] Baron-Cohen predicted a large

correlation between systemizing and math, physics, and engineering. The empathizing quotient has been shown to be strongest for post-secondary humanities students, while the systemizing quotient is larger for science students.[28] The test instrument has been shown to be reasonably reliable and valid.[29] It seems likely to us that the ES scale is measuring something very similar to our β_C one. The course studied here is the first semester of a two-semester sequence, and the second semester includes waves, electricity and magnetism and special relativity. We gave the ES test instrument in this second course and compared results to the True Colors instrument. The correlations with student performance were much greater for the True Colors measurement than for the ES one.[30]

V. CONCLUSIONS

We have refined the way we characterize the measured personality type of students, and the new method shows the same or larger correlations with student performance as Ref. 1.

The most important question is how can we change our pedagogy to reduce the performance gap we observe.

Here, we have investigated whether student learning teams should be heterogeneous or homogenous in the personality type of the team members. There was no measurable difference, either in terms of performance or satisfaction. As discussed, we have previously shown that student performance did not depend on forming teams based on pre-course FCI scores. Perhaps doing the simple thing, forming teams randomly, is the only reasonable strategy. Or perhaps some effect can be found by forming teams based on Gold-Orange personality types, i.e. by the projection of the centroid onto the γ axis.

Thus, measured personality type joins a long list of factors that correlate with student performance, but for which reducing the observed performance gaps is problematic. These factors include gender, pre-course FCI scores, reasons for taking an introductory physics course, cognitive level, and more.

ACKNOWLEDGEMENTS

We thank April Seeley, Dept. of Physics, Univ. of Toronto, who assisted in the collection of the data. Physics undergraduate Emily Knuckey devised the question about the reaction of students to a new physics question or concept used in last part of the pre-course and post-course assessments as part of her summer 2017 undergraduate research project.

REFERENCES

[1] J.J.B. Harlow, D.M. Harrison, M. Justason, A. Meyertholen, and B. Wilson, "Personality types and student performance in an introductory physics course," Phys. Rev. PER **13**, 020124 (2017).
[2] https://www.pearson.com/us/higher-education/products-services-teaching/learning-engagement-tools/learning-catalytics.html (Retrieved September 29, 2017).
[3] C. Spearman, "The proof and measurement of association between two things," Amer. Jour. of Psychology **15**, 72 (1904).
[4] For example C.R. Martin, *Looking at Type: The Fundamentals Using Psychological Type To Understand And Appreciate Ourselves And Others*, (Center for Application of Personality Type, 2001). This characterization actually goes back to Jung's pioneering 1921 analysis in C.G. Jung, *Psychological Types*, R.F.C. Hull, ed., H.G. Baynes, trans., in *The Collected Works of C.G. Jung*, Vol. 6, Bollingen Series XX (Princeton Univ. Press, 1976).
[5] https://www.cpp.com/pdfs/MBTI_FormM_Supp.pdf (Retrieved June 29, 2017)
[6] C. Spearman, "Correlation calculated from faulty data," British Jour. of Psychology **3**, 271 (1910), and W. Brown, "Some experimental results in the correlation of mental abilities," British Jour. of Psychology **3**, 296 (1910).
[7] See, for example, R.M. Sapolsky, *Behave: The Biology of Humans at Our Best and Worst* (Penguin, New York, 2017), especially Chapt. 6, "Adolescence; or, Dude, Where's My Frontal Cortex?"
[8] M.M. Mukaka, "A guide to appropriate use of Correlation coefficient in medical research," Malawi Med. Jour. **24**(3), 69 (2012).
[9] J.J.B. Harlow, D.M. Harrison, and A. Meyertholen, "Effective students teams for collaborative learning in an introductory physics course," Phys. Rev. PER **12**, 010138, (2016).
[10] The values of the centroids are all evenly divisible by 0.5. In this plot and Fig. 3 of Ref. 1 the points have been "jittered" somewhat.
[11] D.M. Harrison, "Analyzing class performance on tests: gray swans," (2017): https://faraday.physics.utoronto.ca/PVB/Harrison/GraySwan/GraySwan.pdf
[12] These students had a wide variety of reasons for choosing this answer. For example, a few had already graduated and were taking this course because it is required for admission to a professional faculty such as medicine. A few were in architecture. There were a number of other reasons.
[13] J.J.B. Harlow, D.M. Harrison, and A. Meyertholen, "Correlating student interest and high school preparation with learning and performance in an introductory physics course," Phys. Rev. ST PER **10**, 010112 (2014).
[14] D.M. Harrison, "Factors correlated with students' scientific reasoning ability in an introductory physics course," in K.A. MacLeod and T.G. Ryan, eds., *The Physics Educator: Tacit Praxes and Untold Stories* (Common Ground Publishing, Champaign IL, 2016), Chapter 11, pg. 186 - 212.
[15] W.K. Adams, K.K. Perkins, N.S. Podolefsky, M. Dubson, N.D. Finkelstein, and C.E. Wieman, "New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey," Phys. Rev. ST PER **2**, 010101 (2006).

[16] A density plot is a histogram which has used kernel smoothing. Here each value is the weighted average such that closer points are given a larger weight. In this case a Gaussian weighting function is used.
[17] S. Pociask, D. Gross, and H.-Y. Shih, "Does Team Formation Impact Student Performance, Effort, and Attitudes in a College Course Employing Collaborative Learning?" Jour. of the Scholarship of Teaching and Learning **17**(3), 19 (2017).
[18] J.-A. Brown, "Using Psychology in the Physics Classroom: Five Steps to Improving Classroom Effectiveness," The Physics Teachers **56**(1), 32 (2018). https://doi.org/10.1119/1.5018687
[19] L.J. Lee, M.E. Connolly, C.R. Henderson, and W.M.Christensen, "A comparison of student evaluations of instruction vs. students' conceptual learning gains," Am. J. Phys. **86**(7), 531 (2018).
[20] R. Thornton and D. Sokoloff, "Assessing students learning of Newton's laws," Am. J. Phys. **66**(4), 228 (1998).
[21] D.R. Riso and R. Hudson, *Discovering Your Personality Type: The Essential Introduction to the Enneagram, Revised and Expanded*, (Houghton Mifflin Harcourt, 2003).
[22] W.M. Marston, *Emotions of Normal People*, (Harcourt, Brace and Co., 1928).
[23] http://www.herrmannsolutions.com/ (Retrieved Sept. 23, 2017).
[24] http://www.hoganassessments.com/assessment/hogan-personality-inventory/ (Retrived Sept. 23, 2017).
[25] http://www.kilmanndiagnostics.com/overview-thomas-kilmann-conflict-mode-instrument-tki (Retrieved Sept. 23, 2017).
[26] For example B. Dattner, "Most Work Conflicts Aren't Due to Personality, Harvard Business Review (2014). https://hbr.org/2014/05/most-work-conflicts-arent-due-to-personality (Retrieved Sept. 24, 2017).
[27] S. Baron-Cohen, "The extreme male brain theory of autism," Trends in Cognitive Science **6**(6), 248 (2002).
[28] F. Focquaert, M.S. Steven, G.L. Wolford, A. Colden, and M.S. Gazzaniga, "Empathizing and systemizing cognitive traits in the sciences and humanities," Personality and Individual Differences **43**, 619 (2007).
[29] A. Wakabayahi, S. Baron-Cohen, S. Wheelwright, N. Goldenfeld, J. Delaney, D. Fine, R. Smith, and L. Weil, "Development of short forms of the Empathy Quotient (EQ-Short) EQand the Systemizing Quotient (SQ-Short)," Personality and Individual Differences **41**(7), 929 (2006).
[30] D.M. Harrison and A. Meyertholen, "Personality type and physics student performance: comparing two assessments of personality," (2018). Available from david.harrison@utoronto.ca.

Last revision: Thursday, August 9, 2018