Personality types and student performance in an introductory physics course

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

[1] *Department of Physics, University of Toronto, Toronto, Ontario M5S 1A6, Canada*
[2] *Faculty of Engineering, McMaster University, Hamilton, Ontario L8S 0A3, Canada*

(Received 23 March 2017; published 27 October 2017)

We measured the personality type of the students in a large introductory physics course of mostly life science students using the True Colors instrument. We found large correlations of personality type with performance on the precourse Force Concept Inventory (FCI), both term tests, the postcourse FCI, and the final examination. We also saw correlations with the normalized gain on the FCI. The personality profile of the students in this course is very different from the profile of the physics faculty and graduate students, and also very different from the profile of students taking the introductory physics course intended for physics majors and specialists.

DOI: 10.1103/PhysRevPhysEducRes.13.020124

I. INTRODUCTION

Classification of people into different personality types goes back to at least Hippocrates (460–370 BC). A more modern theory of personality types comes from C. G. Jung. Jung published the results of his almost 20 years of research in 1921 in *Psychological Types* [1]. Twenty years later Isabel Briggs Myers and her mother Katharine Cook Briggs slightly changed Jung's theory [2], and the resulting Myers-Briggs Type Indicator (MBTI) is now the most widely used psychological typing tool.

The MBTI is based on four "dichotomies": orientation (extraversion–introversion), cognitive perceiving function (sensing–intuition), cognitive judging function (thinking–feeling), and attitude of the functions (judgment–perceiving). Thus, there are $2^4 = 16$ different personality types in the MBTI taxonomy. Keirsey simplified the Myers-Briggs classification into four "temperament" groupings based on the dichotomies he considered to be most important [3]. He named the temperaments guardians, artisans, idealists, and rationals. Keirsey has a lot of followers in the area of learning and teaching styles [4]. In 1979 Lowry introduced a metaphor for Keirsey's temperaments, using four colors in a system called True Colors [5]. There are many variations of the assessment instruments based on the Kiersey and the True Colors taxonomy. One version of the True Colors test instrument has correlated well with the MBTI [6]. Table I shows the Keirsey temperaments, their corresponding color metaphor, the Myers-Briggs classification, and a very brief summary of each type. Although many people have a dominant personality type, some have two or more types equally, and virtually nobody has a dominant type with no aspects of other types. In addition, some people with a dominant color have almost equal scores for one or more of the others. Below we will refer to the different types by their colors.

As discussed by Shen *et al.*, the MBTI, Keirsey, or True Colors test instruments have been given to students in engineering, psychology, economics, pharmacy, dentistry, and more. For example, data on 3784 students in 8 engineering schools in the U.S. shows that 40.16% of the students' dominant color was gold, 33.79% were green, 19.12% were orange, and 6.94% were blue [7].

The True Colors instrument shown in the Appendix is given every year to 2nd year civil engineering students at McMaster University, Hamilton Ontario. Initially we assigned a dominant color type based on the one with the largest score. For the 86 students in 2016 the results showed that the dominant colors were 31% green, 28% orange, 25% gold, and 16% blue.

The MBTI was given to 20 applied physics students at Central Queensland University, Queensland Australia and collapsing the 16 types into True Colors showed that 63% of the students' dominant color was green, 27% were gold, 10% were blue, and none were Orange [8].

However, the concept of personality type and its measurement can be overused and/or misused. There are troubling questions about the test instruments' statistical structure, reliability, robustness, and validity [9]. In addition, almost all people have a mixture of personality types, and focusing on just one dominant type can be highly misleading. Also, some misguided career counselors use the results to recommend that a person should choose a particular profession: any such attempt to put an individual into a particular "box" is oversimplified to the point of

[*] Corresponding author.
ameyerth@physics.utoronto.ca

TABLE I. The four personality types.

Temperament	Guardian	Artisan	Idealist	Rational
True Color	Gold	Orange	Blue	Green
Myers-Briggs classification	Sensing, Judging	Sensing, Perceiving	Intuition, Feeling	Intuition, Thinking
Characteristic	Love to plan, detail oriented, trustworthy	Playful, energetic, risk-taker	Mediators, optimistic, passionate	Intellectual, idea person, philosophical

being wrong. In this study we attempt to restrict ourselves to using the personality type data to investigate if there are trends and correlations between personality type and performance by the students that can guide us to adjust the structure of our course and the types of pedagogy that we use in order to make it more effective for a larger number of our students.

There are some preliminary studies attempting to correlate brain activity with personality. For example, Koelsch, Skours, and Jentschke used functional magnetic resonance imaging and other techniques, and found some small correlations of the data with the results of a questionnaire on personality type [10]. The questionnaire was the NEO instrument, based on a 5 factor model of personality [11]. They conclude by calling for improved questionnaires based on neurological data.

Here we have measured the personality type of the students in a 1000-student introductory physics course intended primarily for life science students, and correlated the personality type with performance on the Force Concept Inventory [12], both precourse and postcourse, and with the two term tests and the final examination. The course features interactive engagement forms of pedagogy throughout, and is described more fully elsewhere [13]. The fall 2016 session studied here uses the same structure and pedagogy as in the 2014 session described in Ref. [13], except that the textbook is Wolfson [14] instead of Knight [15].

We have also measured the personality type of the physics faculty, first-year physics majors, and physics graduate students. These results were compared to the results of the life science students.

II. METHODS

The MBTI test instrument consists of 93 questions, Keirsey has 71 questions, and NEO has two common versions having 90 and 240 questions. The length and resulting time necessary to answer all the questions makes them unsuitable for use in our pre-course assessment. Further, the MBTI and Keirsey instruments are both used by many major corporations for assessing their employees, and are available only for a fee. This also precludes them from use in our study. There are different versions of the True Colors test instrument, and some of them also require a fee for use.

For this study, we chose a True Colors instrument that is free and fairly short [16]. In its original format, it can be self-scored. The Appendix is a slightly modified version of that instrument. This version of the test instrument is referred to as the "Word Cluster" version.

In our precourse assessment, given during the first week of the term, we converted this instrument into multiple-choice format. The major changes from the original are that we have omitted 1 page of introductory material and another page at the end describing the characteristics of each color. In the table, we also changed "confrontational" to "confrontive." For example, referring to set A of Group I in the Appendix, one of the 20 questions in this format is

Question 15: For set A in Group I:
A. Set A is *most* like me.
B. Set A is *a lot* like me.
C. Set A is *somewhat* like me.
D. Set A is *least* like me.

Then there are 3 more similar questions about Set B, C, and D of Group I. There are then four more sets of 4 questions each for the other four Groups of words. We scored "most" as 4 points, "a lot" 3 points, "somewhat" 2 points, and "least" 1 point, the total number of points for all five groups should be $5 \times (1 + 2 + 3 + 4) = 50$. The result of the True Colors assessment is a numerical value for each of the 4 colors, ranging from 5, the lowest, to 20, the highest. As with all precourse assessments, the students are not given their results and are therefore not explicitly told about their measured personality type.

Each of the four colors of the True Colors assessment is measured five times, once for each of the five groups of word clusters. One measure of the reliability of the assessment instrument is to calculate Cronbach's α for the five measurements of each color [17]. Table II summarizes the results of the calculations. Values between 0.70 and 0.90 are heuristically described as "acceptable" and values much larger than 0.90 are often considered to be too good to be true [18]. This analysis supports the reliability of our modified-multiple-choice version of the True Colors instrument.

TABLE II. Cronbach's α of the color score for each of the five groups of word clusters.

Color	α
Blue	0.79
Gold	0.76
Green	0.75
Orange	0.77

Also included in the precourse assessment were the 14 questions from the first of two half-length FCI tests [19]. These two tests each take much less time to administer and are shown to be valid alternatives to the full FCI test instrument [20]. The postcourse assessment consisted only of the second half-length FCI test and was given to the students during the last week of the term.

Initially we assigned a color type for each student by choosing the color with the highest value. 978 students wrote the precourse assessment, which included the True Colors questions. This was almost all the students who were enrolled at that time. However, 12 students did not answer all the questions on personality types and are excluded from our analysis. For the 120 students who had 2 or more color types equally dominant all combinations were represented. The two most prevalent ties were blue-gold (32 students of $120 = 27\%$) and gold-green (32 students of $120 = 27\%$); 2 students had all four types equally, and 11 students had three types equally dominant. Various weighting schemes were attempted to account for the ties and also for the values of the nondominant color values. Finally, we settled on an analysis based on the *centroid* of the colors scores.

For example, if a student has scores of (blue, gold, green, orange) $= (7, 15, 15, 13)$, we can plot the blue score in the first quadrant $(x, y) = (7, 7)$, the gold score in the second quadrant $(x, y) = (15, -15)$, the green score in the third quadrant $(x, y) = (-15, -15)$, and the orange score in the fourth quadrant $(x, y) = (13, -13)$. Then we calculate the following centroid of the scores:

$$x_C = \frac{\text{blue} + \text{orange-gold-green}}{4},$$
$$y_C = \frac{\text{blue} + \text{gold-green-orange}}{4}. \quad (1)$$

For the example student, this gives $(x_C, y_C) = (-2.5, -1.5)$. Figure 1 illustrates for this student. The central red dot is the value of the centroid. As shown in Table I, there is some justification for the color assignments to these specific quadrants. In the original Meyers-Briggs classification, green and blue are both intuitive, but they are opposite in regards to thinking or feeling. Similarly, gold and orange are both sensing, but are opposite in regards to judging or perceiving. The arrangement of the quadrants proposed here allows us to examine the MBTI so-called FT dimension (feeling vs thinking) as well as the JP dimension (judging vs perceiving). These two dimensions lie on diagonal lines with slopes of 1 and -1, respectively, and pass through the origin. In Fig. 1, the student has a gold-green tie for highest value. What places the centroid into the green is not that the student had a high orange (13) but that the student had a low blue (7). Thus thinking is “beating” feeling (higher green-blue difference of −8) by more than the judging is beating perceiving (lower gold-orange difference of −2).

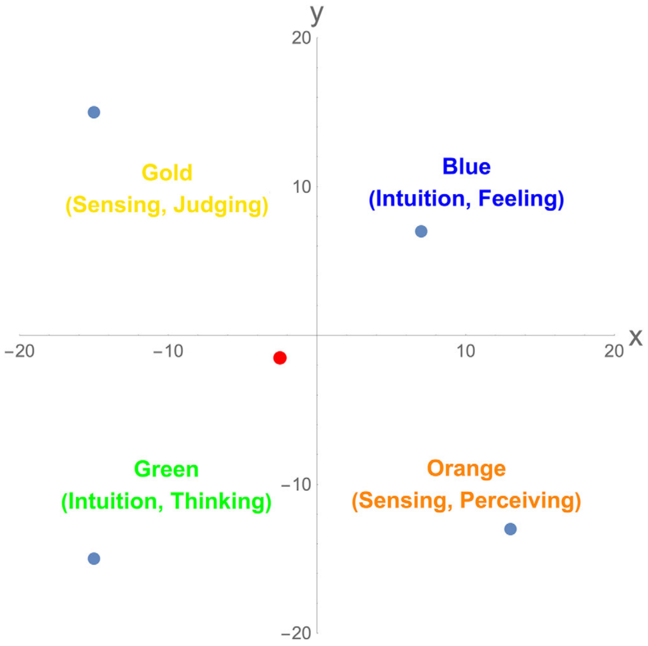

FIG. 1. Illustrating the calculation of the centroid.

We then assign a color type for each student based on their centroid values. Defining a cutoff value c, then a blue student is $(x_C, y_C) = (>c, >c)$, a gold student is $(<c, >c)$, a green student is $(<c, <c)$, and an orange student is $(>c, <c)$. Results such as grades on the first term test are surprisingly insensitive to the value of the cutoff c. For example, below we will show that blue students consistently exhibit the weakest performance on test and examination grades, and on FCI scores. Table III shows the mean value of the first term test for blue students as defined by various by values of c.

Therefore, we chose $c = 0.0$ to assign a color type for each student.

There were a total of 108 students from our sample who had x_C and/or $y_C = 0$, so did not have a single color type. Of these, 9 students had centroids at the origin. There were 37 students whose centroids were equally blue and gold, i.e., $(x_C = 0, y_C > 0)$. Similarly there were 29 students with equal gold-green centroids, 12 students with equal green-orange centroids, and 12 students with equal blue-orange centroids. It is fairly easy to show that if the colors

TABLE III. Mean test 1 grades for blue students for different cutoff values.

Cutoff c	Mean test grades (%)
0.0	42.4 ± 1.2
0.2	42.5 ± 1.2
0.4	42.1 ± 1.3
0.6	42.6 ± 1.5
0.8	43.7 ± 1.7
1.0	41.6 ± 2.0

scores for 3 colors are equal with the fourth score different, the centroid will not lie on either axis, and similarly if 2 colors for opposite quadrants are equal with the other 2 colors different from each other the centroid will also not lie on either axis.

We did some analysis of results using three methods: the simple-minded assigning of color types by choosing the color score with the highest value, a one-dimensional weighting procedure to try to account for all 4 color scores [21], and the two-dimensional centroid method discussed here. All three showed qualitatively the same trends, but somewhat different quantitative values. We believe the centroid method is the most accurate of the three in assessing the impact of colors on student performance, and that is what is used in the remainder of this study.

The methodology described above was also applied to the results obtained by administering the True Colors instrument to the physics faculty, first-year physics majors, and physics graduate students.

III. RESULTS

A. Centroid calculations and color assignments

We gave the True Colors assessment to the physics faculty at the University of Toronto in 2016; 26 of 63 faculty (41%) responded to the anonymous survey. Just using the highest score, not the centroid, to assign a color, the dominant colors were 19 green, 1 gold, 1 blue, 1 orange, 3 green-gold ties, and 1 orange-gold tie. There is a separate first year course for our physics majors and specialists, and we also gave the True Colors assessment to those students. 29 of 208 students, 14%, responded to the anonymous survey. The dominant colors were 18 green, 6 gold, 2 orange, and 0 blue. There was 1 green-blue tie, 1 orange-blue tie, and 1 orange-green tie. We also gave the True Colors assessment to Toronto physics graduate students; 30 of about 200 students (15%) responded to the anonymous survey. The colors were 16 green, 3 gold, 4 blue, 3 orange, 1 green-gold tie, 1 orange-green tie, and 2 green-blue ties.

For the physics faculty, Fig. 2(a) shows the values of the centroids, and the open circle shows the mean value of the centroids. Figure 2(b) shows the centroids for the students in our 1st year course for physics majors and specialists, and Fig. 2(c) shows the centroids for our physics graduate students.

Figure 3 shows the centroids for all students in the 1000-student introductory physics course for life science students that we are studying here, and the mean value of centroids as the open circle. Also shown are histograms of the values of centroids.

Perhaps not surprisingly, the physics majors and specialists are much closer to the physics faculty than the mostly life science students in the course of this study. The physics graduate students are similar to the physics faculty in that the mean centroid is also located in the Green quadrant, however, it is shifted somewhat towards the orange quadrant.

Table IV shows the mean value of the centroids for the physics faculty, students in the course for physics majors and specialists, the physics graduate students, and for the students in the course being studied here. The stated uncertainties are the standard "error" of the mean $\sigma_m = \sigma/\sqrt{N}$ [22].

Table V shows the distribution of color types for the 1000-student course as determined by the value of the centroid. Students whose centroid fell on one of the axes did not have a single color type and were, therefore, not included.

B. Statistical tests and student performance

There were five assessments in the course: the precourse half FCI, the first term test, the second term test, the postcourse half FCI, and the final examination. We will examine these in order. Then the results of an ANOVA regression for all five assessments are discussed.

1. Precourse FCI scores

The precourse half FCI was given in the first week of the term. 978 students wrote the assessment. As discussed in, for example, Ref. [13], the distribution of FCI scores is not Gaussian, so the median is more appropriate than the mean to characterize the results. We report FCI scores in percent.

Table VI shows the median precourse FCI scores for all students and for students with defined color types. The uncertainties are $1.58 \times \mathrm{IQR}/\sqrt{N}$, where IQR is the interquartile range and N is the number of students in the sample [23]. This uncertainty is roughly taken to indicate a 95% confidence interval, i.e., it is equivalent to $2 \times \sigma_m$ for a normal distribution.

The highest median value is for green students and the lowest median value is for the blue students. This is a pattern that we will see for all the other assessments discussed below. For the precourse FCI scores, the difference between the orange and green students is not significant: $(57.1 \pm 4.8) - (50.0 \pm 5.4) = 7.1 \pm 7.2$. The gold-green difference is nonzero within uncertainties: $(57.1 \pm 4.8) - (50.0 \pm 2.8) = 7.1 \pm 5.6$. Further analysis of the various pairs of colors is in Sec. III. B. 6 below.

Remembering that the claimed uncertainty is roughly equivalent to $2 \times \sigma_m$ for a normal distribution, the combined uncertainty of the difference between the green and blue students is about 6 "standard deviations"; i.e., since $(57.1 \pm 4.8) - (35.7 \pm 4.4) = 21.4 \pm 6.5$, then calling Δm the difference in the median values, and u the uncertainty in Δm,

$$\Delta m/u = 21.4/6.5 \simeq 3 - 6 \, \text{standard deviations.}$$

We used Cliff's δ to examine the effect size of the difference between the green and blue students. The Cliff δ

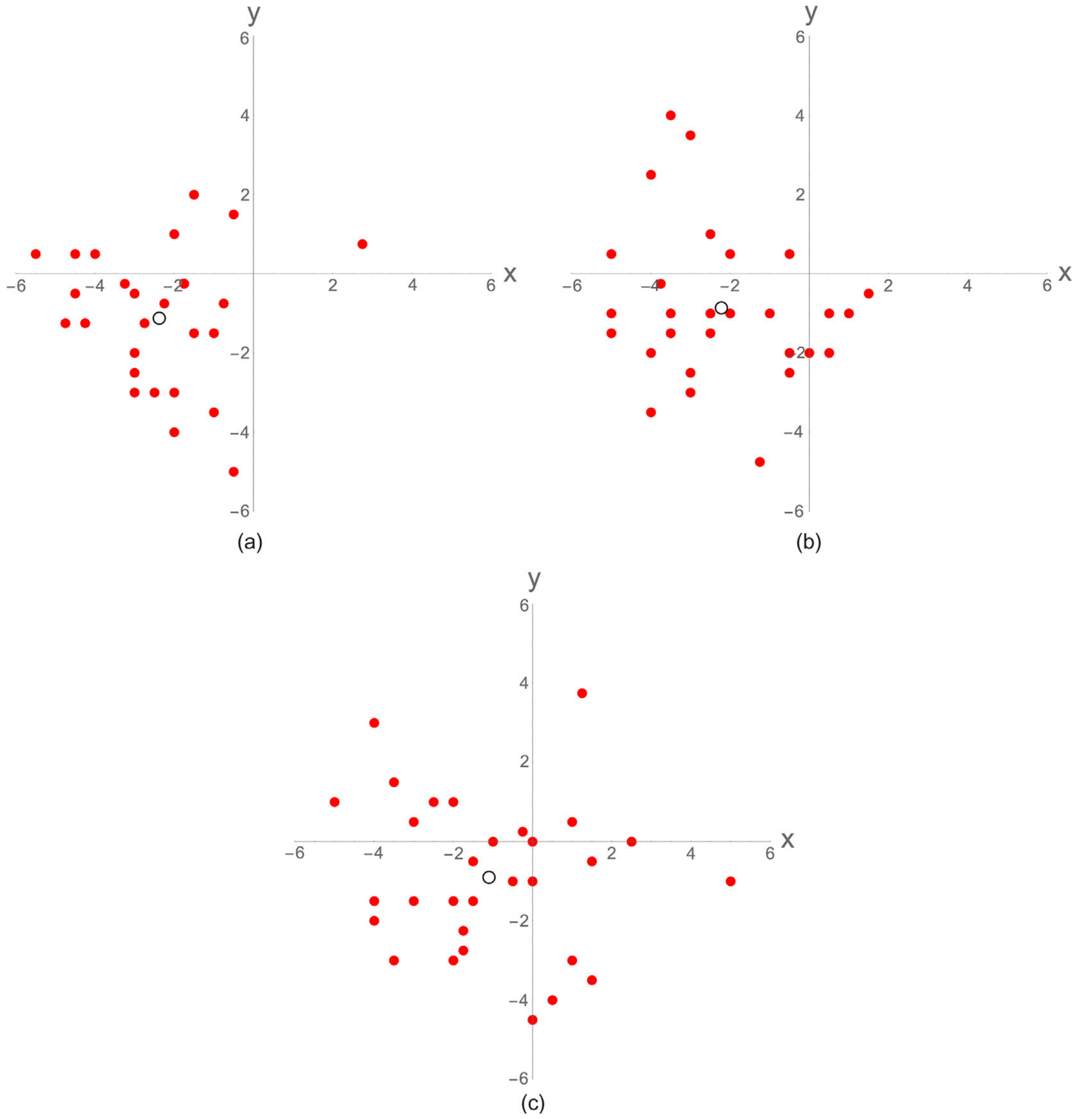

FIG. 2. The centroids (the red dots) and the mean of the centroids (the open circle). (a) Physics faculty. (b) 1st year physics majors and specialists. (c) Physics grad students.

for 2 samples is the probability that a value randomly selected from the first group is greater than a randomly selected value from the second group minus the probability that a randomly selected value from the first group is less than a randomly selected value from the second group. It is calculated as

$$\delta = \frac{\#(x_1 > x_2) - \#(x_1 < x_2)}{N_1 N_2}, \qquad (2)$$

where indicates counting. The values of δ can range from -1, when all the values of the first sample are less than the values of the second, to $+1$, where all the values of the first sample are greater than the values of the second. A value of 0 indicates samples whose distributions completely overlap.

Calculating the value of Cliff's δ for green and blue precourse FCI scores gave $\delta = 0.37$, which is heuristically characterized as "medium." The 95% confidence interval range is 0.26–0.47; since this range does not include zero, the difference is statistically significant.

The box plot is a particularly nice way of visually comparing distributions such as FCI scores and test grades. Figure 4 shows the box plot of the precourse FCI scores for the different personality types of the students. The "waist" on the box plot is the median, the "shoulder" is the upper quartile, and the "hip" is the lower quartile. The vertical lines extend to the largest (smallest) data point value less (greater) than a heuristically defined outlier cutoff [24]. The "notch" around the median value represents the statistical uncertainty in the value of the median.

2. First term test

The first term test was given early in the term, after 3 weeks of classes. 927 students wrote the test, which was

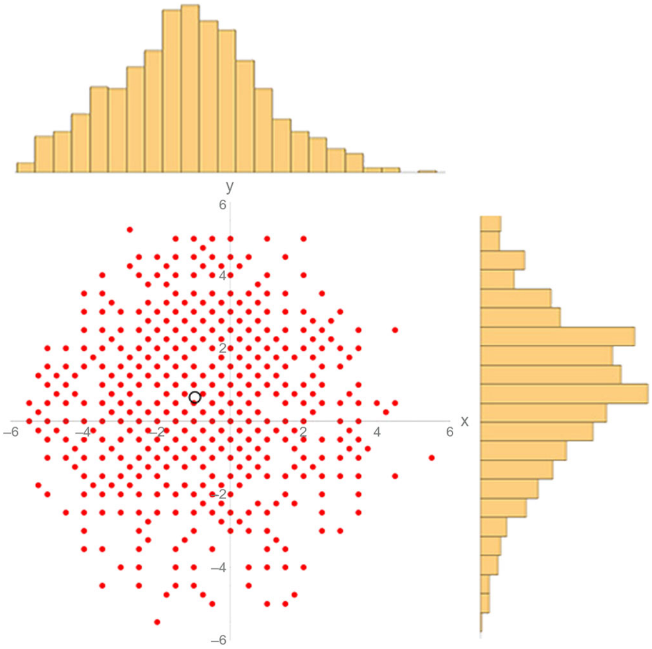

FIG. 3. The centroids of the students, the mean of the centroids (the open circle), and histograms of the centroid values.

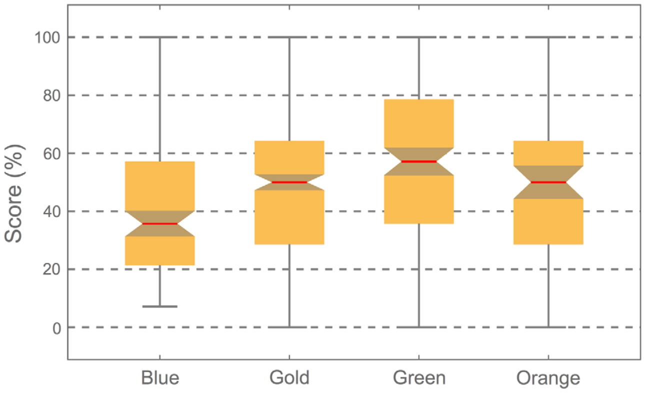

FIG. 4. Precourse FCI scores for different personality types.

80 min long. The mean grade, in percent, was 47.23 ± 0.46 which was lower than we intended, and was rescaled in calculating a final grade in the course. However, since the mean is close to 50%, it is almost perfect for discriminating between students [25]. The stated uncertainty is the standard "error" of the mean, $\sigma_m = \sigma/\sqrt{N}$. Table VII shows the mean scores for students with defined color types. We see that green students outperformed blue ones by over $5 \times \sigma_m$: $(51.1 \pm 1.1) - (42.5 \pm 1.2) = 8.6 \pm 1.6$.

TABLE IV. Mean centroid values.

	$(\bar{x}_C, \bar{y}_C)$
Physics faculty	$(-2.4 \pm 0.3, -1.1 \pm 0.3)$
Students in the first year course for physics majors and specialists	$(-2.2 \pm 0.3, -0.9 \pm 0.4)$
Physics graduate students	$(-1.1 \pm 0.4, -0.9 \pm 0.4)$
Students in the course being studied	$(-0.97 \pm 0.06, 0.65 \pm 0.06)$

TABLE V. Dominant color types.

Color Type	N	Percentage
Blue	162	18
Gold	419	48
Green	200	23
Orange	98	11

TABLE VI. Median precourse FCI scores.

	N	Median precourse FCI score (%)
All students	978	50.0 ± 1.8
Blue	162	35.7 ± 4.4
Gold	419	50.0 ± 2.8
Green	200	57.1 ± 4.8
Orange	98	50.0 ± 5.4

Cliff's δ for the blue and green students is 0.32, which is heuristically characterized as "small." The 95% confidence interval is 0.19–0.43; since this does not include zero, the difference is statistically significant.

For distributions which are Gaussian, such as test grades, an alternative to Cliff's δ is Cohen's d [26]. It is defined as

$$d \equiv \frac{|\mathrm{mean}_1 - \mathrm{mean}_2|}{\sigma_{\mathrm{pooled}}}, \tag{3}$$

where

$$\sigma_{\mathrm{pooled}} = \sqrt{({\sigma_1}^2 + {\sigma_2}^2)/2}. \tag{4}$$

Cohen's d is somewhat easier to interpret than Cliff's δ. Note that it uses the standard deviation, not the standard error of the mean.

Comparing the blue and green student grades on the test gives $d = 0.60$, so the difference in the means is over one-half of the pooled standard deviation. This value is heuristically defined as a medium difference. The 95% confidence interval for d is 0.36–0.83: since this range does not include zero, the difference is statistically significant.

Figure 5 shows the boxplot of the test grades for the different personality types. The dots are data points that lie outside of the cutoffs, and are considered to be outliers.

TABLE VII. Mean test 1 grade for students with a dominant personality type.

	N	Mean test 1 grade (%)
Blue	134	42.5 ± 1.2
Gold	373	47.2 ± 0.7
Green	172	51.1 ± 1.1
Orange	89	49.6 ± 1.4

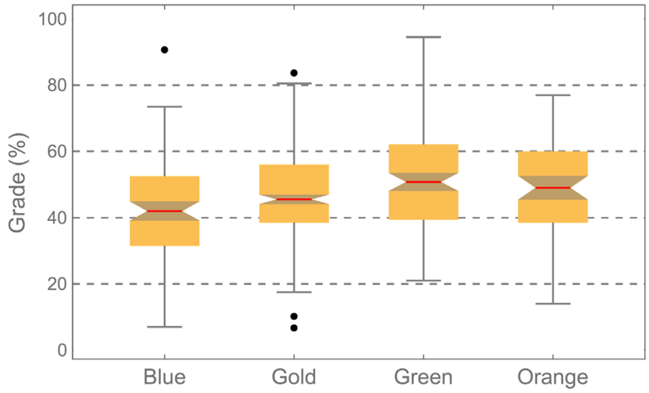

FIG. 5. Box plots of grades on test 1 for different personality types.

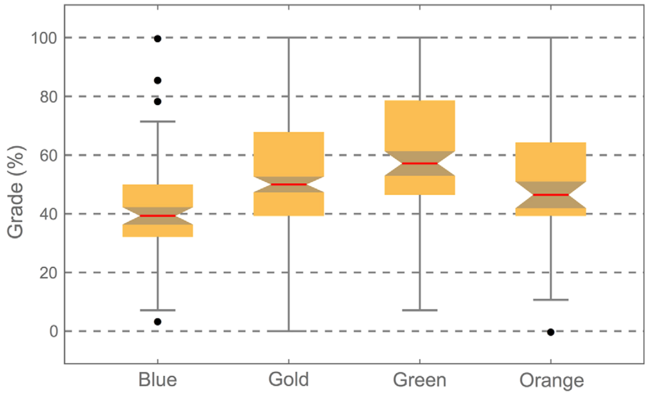

FIG. 6. Box plots of grades on test 2 for different personality types.

3. *Second term test*

The second term test was given during the 9th week of classes. 716 students wrote the test, which was 80 min long. Once again, the overall mean on the test was lower than we intended: it was 51.77 ± 0.81. Table VIII shows the test grades for different personality types. Students lacking a clearly defined color were excluded from the table.

The same pattern we have seen for the precourse FCI and the first term test is true here: the green students outperformed the blue students. The difference between the green and blue students is about $6.5 \times \sigma_m$: $(60.4 \pm 1.8) - (42.9 \pm 2.0) = 17.5 \pm 2.7$. This difference is the largest of the three assessments we have examined so far by a small amount.

The Cliff δ is also the largest. It is 0.47 (medium) with a 95% confidence interval of 0.33–0.59.

Cohen's d is also larger for this test than for the first one. It is 0.85 ("large") with a 95% confidence interval of 0.57–1.13.

The difference in test grades is confirmed by the box plot in Fig. 6.

4. *Postcourse FCI and FCI gains*

The postcourse half FCI was given during the last week of the term. 671 students wrote the assessment. Table IX summarizes the results.

Once again, the green students outperformed the blue students. The difference between the green and blue scores is $(85.7 \pm 5.3) - (57.1 \pm 7.1) = 28.6 \pm 8.9$. Calling Δm the difference in the median values, and u the uncertainty in Δm, the difference is about the same as observed for the precourse scores:

$$\Delta m / u = 28.6/8.9 \simeq 3 - 6 \,\text{standard deviations.}$$

Cliff's δ for the blue and green scores is 0.37 (medium) with a 95% confidence interval of 0.22–0.50. These values are also comparable to the ones for the precourse.

One hopes that the students' performance on the FCI is higher at the end of the course than at the beginning. Comparing Table VIII to Table V, the postcourse scores are higher than the precourse ones for all categories of students. The box plot of postcourse scores, which is not shown, looks similar to that of the precourse ones, Fig. 4, except for the upward shift in values.

As in Ref. [13], we characterize the gains from the precourse to the postcourse by the median normalized gain:

$$\langle g \rangle_{\text{median}} = \frac{\langle \text{postcourse\%} \rangle - \langle \text{precourse\%} \rangle}{100 - \langle \text{precourse\%} \rangle}, \tag{5}$$

where the angle brackets on the right-hand side indicate medians. We examined the gains for the 628 "matched" students who wrote both the precourse and the postcourse FCI. Table X summarizes.

Once again, Green students outperformed blue ones, with gold and orange students in the middle. The difference between the green and blue students is

TABLE VIII. Mean test 2 grade for students with a dominant personality type.

	N	Mean test 1 grade (%)
Blue	90	42.9 ± 2.0
Gold	282	53.3 ± 1.2
Green	145	60.4 ± 1.8
Orange	75	50.0 ± 2.3

TABLE IX. Median postcourse FCI scores.

	N	Median postcourse FCI score (%)
All students	671	71.4 ± 2.6
Blue	90	57.1 ± 7.1
Gold	202	71.4 ± 2.8
Green	139	85.7 ± 5.3
Orange	73	71.4 ± 7.9

TABLE X. Median normalized FCI gains for matched students.

	$\langle g \rangle_{\text{median}}$
All students	0.43 ± 0.06
Blue	0.25 ± 0.13
Gold	0.33 ± 0.08
Green	0.60 ± 0.16
Orange	0.43 ± 0.17

$(0.60 \pm 0.16) - (0.25 \pm 0.13) = 0.35 \pm 0.21$. This difference is roughly 3 standard deviations.

Another way of examining gains is to calculate a normalized gain G for each individual student, defined as

$$G = \frac{\text{postcourse\%} - \text{precourse\%}}{100 - \text{precourse\%}}. \qquad (6)$$

Figure 7 shows the box plot of G for different personality types. The vertical scale is chosen to not display the 34 students who either had a precourse score of 100 or a value of $G < -0.5$.

Cliff's δ for the values of G for blue and green students is 0.23, which is heuristically characterized as small. The 95% confidence interval is 0.09–0.38, so the difference is statistically significant.

5. Final examination

The final examination was 2 h long, and was written by 696 students. The overall mean grade was 64.4 ± 0.7; at the University of Toronto, this is a letter grade of C. Table XI shows the mean grades for different color types.

The same pattern is evident that has been shown for all the other assessment instruments: with green students outperforming blue students. In this case the difference between the green and blue performance is $(70.0 \pm 1.6) - (58.4 \pm 1.8) = 11.6 \pm 2.4$, which is almost a 5 standard deviation difference.

Cliff's δ and Cohen's d for green and blue students are somewhat smaller than for the second term test, 0.36 and 0.64, respectively. Both of these are heuristically characterized as medium. The 95% confidence intervals are 0.21–0.48 and 0.36–0.91, respectively, so both statistics indicate a statistically significant difference.

The box plot, which is not shown, also shows no surprises.

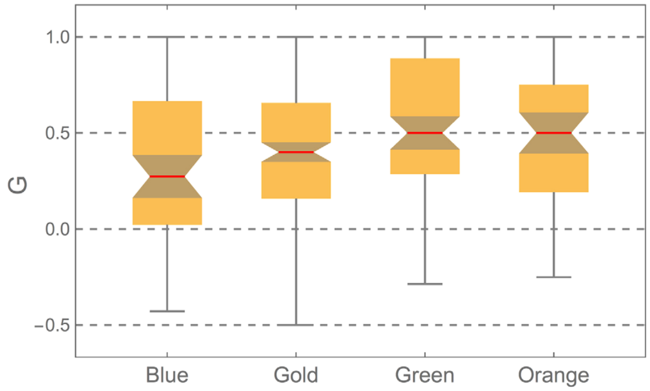

FIG. 7. The normalized gain for different personality types.

TABLE XI. Mean final exam grade for students with a dominant personality type.

	N	Mean final exam grade (%)
Blue	89	58.4 ± 1.8
Gold	273	65.6 ± 1.1
Green	143	70.0 ± 1.6
Orange	73	63.5 ± 2.2

6. ANOVA results

Above we used Cliff's δ and Cohen's d to compare performance for two of the four colors, blue and green. In addition, we did a one-way analysis of variance (ANOVA) of the means of all course assessments for all four colors. The results are summarized in Table XII. We found that the means of all course assessments had statistically significant differences when broken into groups of color types. Note that ANOVA assumes the values are normally distributed, which is not correct for FCI scores, so those values should be treated with particular caution.

From the results of the ANOVA, we used Tukey's honest significance test for a 95% confidence level to examine where those differences lie [27]. The results are summarized in Table XIII.

Because all assessments have a wide spread of values and the small number of blue and orange students in our sample, it is difficult to interpret some of these values. Nonetheless there are some trends. With the exception of the FCI gain, the green-blue differences are all much less than the accepted statistically significant p value of <0.05. The orange-gold differences are not significant for any assessment.

Trying to draw further conclusions from the data is probably not appropriate without better statistics and a deeper analysis of the assessment instruments.

TABLE XII. One-way analysis of variance (ANOVA).

Assessment	F test	p value	F critical
FCI Pre	14.026	6.2×10^{-9}	2.615
Test 1	10.788	6.0×10^{-7}	2.617
Test 2	14.195	6.1×10^{-9}	2.620
FCI post	9.243	5.6×10^{-6}	2.621
FCI gain	2.876	0.036	2.621
Final exam	7.768	4.3×10^{-5}	2.620

TABLE XIII. p values for Tukey's honest significance test for pairs of colors.

Assessment	Gold-Blue	Green-Blue	Orange-Blue	Green-Gold	Orange-Gold	Orange-Green
FCI Pre	0.0010	1.0×10^{-9}	0.011	6.0×10^{-4}	0.97	0.094
Test 1	0.0034	3.0×10^{-7}	8.3×10^{-4}	0.011	0.44	0.84
Test 2	1.9×10^{-4}	2.8×10^{-9}	0.12	0.0039	0.60	0.0021
FCI Post	0.042	3.1×10^{-6}	0.37	0.0031	0.95	0.015
FCI Gain	0.93	0.063	0.48	0.065	0.68	0.85
Final Exam	0.0066	1.6×10^{-5}	0.29	0.088	0.81	0.061

IV. DISCUSSION

Earlier, we provided some justification for how we assigned the colors to the four quadrants of an x-y plot. The data showing significant differences in student performance between blue and green students with the other colors in the middle provides another justification: surely these blue and green students should be in opposite quadrants of the plot. The fact that the assignments that we made contain a mnemonic (the color names are assigned to the quadrants in alphabetical order) is a coincidence. Other assignments, such as green-orange-blue-gold, should be equally valid so long as the green and blue scores are in opposite quadrants.

It should be made clear at the outset that just because we see correlations between color type and student performance does not mean we are suggesting a simple causal relationship. We are also not advocating for using color type to assess the suitability of a student for our course: a glance at, for example, the distribution of grades on the first term test in Fig. 5 makes it clear that there are high performing and low performing students for all color types. However, thinking about personality types provides a new perspective on our students and some of the difficulties they may have in doing well in our course.

For example, it is clear that there is an "impedance mismatch" between the strongly green physics faculty and graduate student TAs, and our students in the introductory physics for the life sciences course, who are mostly gold but with significant numbers of blue and orange color types. In order to accommodate the detail-oriented gold students, faculty should be sure to make expectations, deadlines, etc., extremely clear.

Similarly, to accommodate the blue students, we should emphasize the benefits of physics for the public in general and for health care in particular. It could also be useful for these students, who value intuition, to point out, as Livio wrote, "More than 20 percent of Einstein's original papers contain mistakes of some sort. In several cases, even though he made mistakes along the way, the final result is still correct. This is often a hallmark of great theorists: They are guided more by intuition than by formalism" [28].

To make the course more relevant to the orange students, it could be worthwhile to devote some time in making the risks of scientific inquiry clear by emphasizing that good scientists need the courage to be wrong. It could also be useful for these students to point out, as Gopnik *et al.* wrote, that "Science is a kind of institutionalized childhood" [29].

None of these recommendations are particularly revolutionary. However, putting these issues in the context of personality types may make them particularly compelling.

We need to beware of thinking statements such as "I/you/he/she am/are/is/is measured to have an orange personality type" are the same as "I/you/he/she am/are/is/is an orange personality type." As with all such psychological assessments, the result can be faked to one degree or another. For example, a person who is inherently a playful risk taker (orange) can consciously choose to answer the True Colors assessment questions to come out as a detail-oriented person (gold). Even without such a conscious decision, we all have a self-image, which perhaps we acquired from what we have been told by our parents, peers, or former teachers. In such a case we will unconsciously choose answers that conforms to that self-image. And, of course, such self-images can be self-fulfilling prophecies. So a student who believes he or she is an intellectual idea person (green) will have the confidence necessary to do well in a physics course. This is one reason why we cautioned against interpreting the correlations we see between personality type and performance as indicating a simple causal relationship.

Steele and Aronson introduced the phrase "stereotype threat" in 1995 in the context of test performance of African-American students [30]. Since then the phrase has been applied to the gender issue in physics courses [31,32]. We are proposing that it is also appropriate in the case of a mismatch between a measured personality type and the ability to do well in a physics course.

A related perspective on the issue of color type is that physics is generally perceived by the public to be difficult and requires considerable raw intellectual talent. In terms of personality type, this is most similar to green. In 2015 Leslie, Cimpian, Meyer, and Freeland published a study of U.S. postsecondary institutions [33]. They looked at disciplines that are perceived as requiring different levels of intellectual talent. Those perceptions are negatively correlated with the percentage of female Ph.D. students in the disciplines: the greater the perception of required raw talent, the fewer females in the discipline. Evidence strongly suggests this is true not only in the STEM fields of science, technology, engineering, and mathematics, but also in the social sciences and humanities. A similar correlation was found in the percentage of African-American Ph.D. students, but not Asian-American Ph.D. students. Although there are no data on whether or not the perception that some disciplines require more raw talent than others is actually correct, they argue that in either case stereotype threat is a factor in participation rates.

We think it is likely that physics faculty and graduate students in general are strongly green, as are the faculty and graduate students at the University of Toronto. It would be interesting to examine the color type of faculty and graduate students in other disciplines to see if the fields that are believed to require raw talent are also green compared to fields that are generally considered to be "easier".

In physics education research, the normalized gain on the FCI, $\langle g \rangle$, has played a crucial role for 25 years. It is widely taken to be a measure of the quality of instruction. Its value has been shown repeatedly to depend strongly on the type of pedagogy used, and therefore has played a leading role in the adoption of *interactive engagement* types of teaching. Although precourse and postcourse FCI scores have been shown to depend on a number of factors, the value of $\langle g \rangle$ turns out to be surprisingly insensitive to these factors. For example, previous results suggest the value of $\langle g \rangle$ does not depend on factors such as whether or not the student took a senior-level high school physics course or the student's motivation for taking our course [34]. Furthermore, a previous study suggests that values of $\langle g \rangle$ are consistent when comparing the normal 12-week term of the course studied here to the compressed 6-week version given in the summer [35]. In Ref. [13] we presented evidence that the value of $\langle g \rangle$ is statistically the same for teams of students with roughly equal strength compared to teams with a mixture of student strengths. Hoellwarth and Moelter showed that in a particular implementation of Studio Physics, $\langle g \rangle$ was independent of the instructor [36]. Wood, Galloway, and Hardy showed $\langle g \rangle$ was largely independent of whether or not the student is capable of suppressing an intuitive and spontaneous wrong answer in favor of a reflective and deliberative right one [37], a result that we have replicated [38]. Therefore, the fact that our results suggest a correlation between $\langle g \rangle$ and color type is particularly dramatic and troubling. Evidently our research-based pedagogy does not serve our blue students as well as it should.

When a performance gap is discovered for some factor, such as gender, socioeconomic background, Piagetian cognitive level or, here, personality type, one hopes to find ways to reduce it. Here we have used Cliff's δ and Cohen's d as one way to quantify the performance gap between blue and green students. Figure 8 illustrates this for the precourse FCI, the first term test, the second term test, the postcourse FCI, and the final examination in the course. These are the order in which the students did them. The solid black is for Cliff's δ and the red dashed is for Cohen's d. Note that d is not calculated for FCI scores, since d assumes a normal distribution. It is clear that our current course does not reduce the gap.

The uncertainties in Fig. 8 need some explanation. Earlier, for each of the assessment instruments, we presented a value D for Cliff's δ or Cohen's d, and then the lower value L of the 95% confidence interval range and upper value U of the 95% confidence interval range. We can write this as $D \pm (D - L) = D \pm (U - D)$. However, the uncertainties from the 95% confidence interval correspond to $2 \times \sigma$. In plots the displayed uncertainties are usually the standard deviation, not twice the standard deviation. Therefore, the displayed uncertainties in Fig. 8 are $(D - L)/2 = (U - D)/2$.

Probably because the first term test was much too hard, the dropout rate for this session of the course, about 25%, was higher than usual. We attempted to correlate the color type with the dropouts, and did not see a large correlation. We also attempted to compare student learning teams comprised of students with the same personality type to teams with a mix of personality types, but for a number of reasons this attempt failed.

There is a somewhat troubling issue with our data on personality types. For each of the five groups of four sets of words, the students are asked to choose which set is most like them, a lot like them, somewhat like them, or least like them; the example question shown in Sec. II is the question for set A of group I. Then there are 3 more similar questions about set B, C, and D of group I. There are then four more sets of 4 questions each for the other four groups of words. Since "most" is scored 4 points, "a lot" 3 points, "somewhat" 2 points, and "least" 1 point, the total number of points for all five groups should be $5 \times (1+2+3+4) = 50$. However, this was only true for just under one-half of the students who did the precourse assessment (419 of 978), please see Table XIV [39].

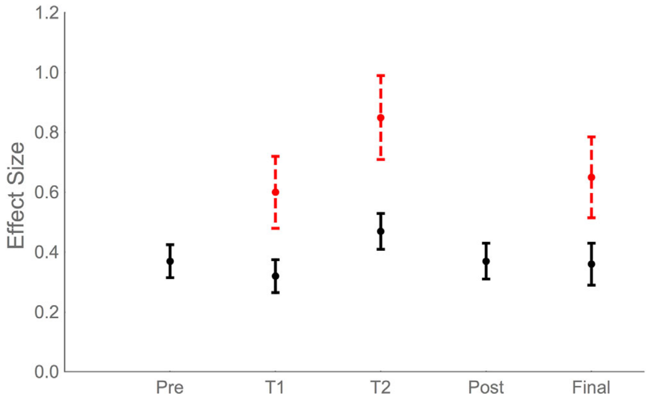

FIG. 8. Comparing blue and green students. Cliff's δ (solid black) and Cohen's d (dashed red) for the precourse FCI (Pre), the first term test (T1), the second term test (T2), the postcourse FCI (Post), and the final examination (Final).

TABLE XIV. Student total color points Fall 2016 PHY 131.

Total number of true color points	Percentage of class (%)
Less than 50 points	17
50 points	43
Greater than 50 and less than or equal 55	19
Greater than 55 and less than or equal 60	11
Greater than 60	10

For any assessment instrument, like this one, where students are given credit for answering all the questions regardless of what they answered, a disturbing issue is that some students will not take their answers seriously and will, for example, answer randomly or just choose A, B, C, or D in order or something similar. In Ref. [8] we showed some data for the postcourse FCI indicating that particularly the good students were not trying to give their best answers. Inserting a question in the middle of the instrument to check that the students are at least reading all the questions can check this, and it turns out that most students seem to take giving accurate answers fairly seriously. So, for the personality type questions, perhaps some students were somewhat confused or sloppy, or perhaps they decided that two word sets were equal in ranking. We have assumed that, despite these issues, the personality scores we measured reflect to some degree the personality types of the students. This assumption is supported by the results shown in Table II.

V. CONCLUSIONS AND FUTURE WORK

In the early days of the Royal Society of London in the 17th century, members regularly performed and reported on experimental measurements [40]. Many of these experiments were crucial in the development of the sciences of mechanics, the gas laws, optics, and more. However, some of those experiments in retrospect look silly. For example, Boyle investigated the difference in behavior of a butterfly, a bee, a hen-sparrow, and a mouse when placed in a partially evacuated chamber [41]. However, it is only in retrospect that those experiments seem silly: at the time people did not understand the issues and in this case oxygen had not even been discovered.

We are not making such grandiose claims for the experiments on personality type described here. However, like those early experimentalists, we are not sure just what we are measuring, or exactly how it relates to student learning and performance. Nonetheless, the correlations that we see between measured personality type and student performance makes it obvious to us that assessments of personality type, however flawed, are measuring something relevant to physics education.

At this stage of our research, it is important to view our results primarily as observations. Some of us were initially skeptical about whether we would see any significant correlation between measured personality types and student performance, and have been very surprised by the size of the correlations that we have observed. An investigation into possible operational strategies for reducing the green-blue performance gap will form the second phase of our research into personality types. A few of our specific intentions are described below. Hopefully the strong correlations observed in this study will entice other researchers to investigate strategies for improved pedagogy based on an understanding of personality type.

We have shown that student performance on the precourse FCI, two term tests, and postcourse FCI, normalized gain on the FCI, and the final examination correlate to the color type, with green students consistently outperforming the blue. For all but the normalized gain on the FCI, the difference in blue-green performance was 5σ or better; for the normalized gain it was somewhat less, at about 3σ.

We believe our observed correlations of personality with student performance are probably true in a much broader context than just students at the University of Toronto. There has been a study of Singapore university students that is similar to ours for two of their courses, one for a first year mechanics course with 110 students and the other for a second year quantum mechanics course with 80 students. Although their statistics are limited because of the small number of students, the results are consistent with ours [42]. A similar study performed on first year chemistry students at the University of Sydney used the Myers-Briggs Type Indicator and found a correlation between student performance and the Myers-Briggs FT dimension (feeling-thinking, i.e., blue-green). With students scoring high in “thinking” outperforming students who scored high in “feeling” [43].

Important questions that we have not addressed here involve what characteristics of these personality types are contributing to the performance gap we have observed, and how can we modify our pedagogy to address these differences. We intend to address these issues in at least three ways.

First, we will be forming two focus groups of students. One will be all blue students and the other all green students. We wish to probe the differences in the ways that these students interact with each other and the material of the course. An individual from outside the department will facilitate these focus groups.

Second, we intend to form the learning teams of 4 students two ways: one will be homogeneous in terms of measured personality type, and the other will be a mixture of different measured personality types. This is similar to our study of effective teams of Ref. [13], except there the teams were formed on the basis of the results of the precourse FCI, not the measured personality type.

Third, it may be that the Investigative Science Learning Environment (ISLE) provides a perspective on pedagogy that addresses the observed gap between the performance of blue and green students [44]. We will be explicitly modifying the activities we use for collaborative learning to incorporate the rubrics developed by ISLE. Our hope is that this will not only benefit all students, but will also reduce the blue-green performance gap.

ACKNOWLEDGMENTS

We thank April Seeley and Xing Xingxing of the Department of Physics, University of Toronto, for assistance in the data collection. We benefited by discussion of personality types with Aephraim Steinberg, the Department of Physics, University of Toronto. We also thank Ho Shen Yong of Nanyang Technical University, Singapore, for sharing his data. The anonymous referees of this journal also made many helpful comments.

APPENDIX: THE TRUE COLORS TEST INSTRUMENT

The original form of the True Colors assessment instrument is given here.

Instructions: Compare all 4 boxes in each row. Do not analyze each word; just get a sense of each box. Score each of the four boxes in EACH row from most to least as it describes you: 4 = most, 3 = alot, 2 = somewhat, 1 = least.

Group I	**Set A**	**Set B**	**Set C**	**Set D**
	Active	Organized	Warm	Learning
	Variety	Planned	Helpful	Science
	Sports	Neat	Friends	Quiet
	Opportunities	Parental	Authentic	Versatile
	Spontaneous	Traditional	Harmonious	Inventive
	Flexible	Responsible	Compassionate	Competent
	Score	Score	Score	Score
Group II	**Set E**	**Set F**	**Set G**	**Set H**
	Curious	Caring	Orderly	Action
	Ideas	People Oriented	On-time	Challenges
	Questions	Feelings	Honest	Competitive
	Conceptual	Unique	Stable	Impetuous
	Knowledge	Empathetic	Sensible	Impactful
	Problem Solver	Communicative	Dependable	
	Score	Score	Score	Score
Group III	**Set I**	**Set J**	**Set K**	**Set L**
	Helpful	Kind	Playful	Independent
	Trustworthy	Understanding	Quick	Exploring
	Dependable	Giving	Adventurous	Competent
	Loyal	Devoted	Confrontational	Theoretical
	Conservative	Warm	Open Minded	Why questions
	Organized	Poetic	Independent	Ingenious
	Score	Score	Score	Score
Group IV	**Set M**	**Set N**	**Set O**	**Set P**
	Follow rules	Active	Sharing	Thinking
	Useful	Free	Getting along	Solving problems
	Save money	Winning	Feelings	Perfectionistic
	Concerned	Daring	Tender	Determined
	Procedural	Impulsive	Inspirational	Complex
	Cooperative	Risk Taker	Dramatic	Composed
	Score	Score	Score	Score
Group V	**Set Q**	**Set R**	**Set S**	**Set T**
	Puzzles	Social causes	Exciting	Pride
	Seeking info	Easy going	Lively hands	Tradition
	Making sense	Happy endings	On	Do things right
	Philosophical	Approachable	Courageous	Orderly
	Principled	Affectionate	Skillful on	Conventional
	Rational	Sympathetic	Stage	Careful
	Score	Score	Score	Score

Total orange score: Sum of A, H, K, N, S __________
Total green score: Sum of D, E, L.P, Q ____________
Total blue score: Sum of C, F, J, O, R _____________
Total gold score: Sum of B, G, I, M, T ____________
If any of the scores are less than 5 or greater than 20 you have made an error. Please go back and read the instructions.

[1] C. G. Jung, Psychological Types, in *The Collected Works of C. G. Jung*, edited by R. F. C. Hull, translated by H. G. Baynes, Bollingen Series XX (Princeton University Press, Princeton, NJ, 1976), Vol. 6.

[2] See, for example, I. B. Myers and P. B. Myers, *Gifts Differing: Understanding Personality Type* (Davies-Black Publishing, Mountain View, CA, 1980).

[3] D. Keirsey, *Please understand me II* (Prometheus Nemesis Book Company, Carlsbad, CA, 1998).

[4] See, for example, R. M. Felder and R. Brent, Understanding student differences, J. Eng. Educ. **94**, 57 (2005).

[5] https://truecolorsintl.com/about-us/what-is-true-colors/ (Retrieved May 17, 2017).

[6] J. A. Whichard, Reliability and Validity of True Colors, True Colors International (2006), http://truecolorsintl.com/wp-content/uploads/2013/05/Research-Validity-and-Reliability-I.pdf.

[7] S.-T. Shen, S. D. Prior, A. S. Whitel, and M. Karamannoglu, Using personality type differences to form engineering design teams, Eng. Educ. **2**, 54 (2007).

[8] B. W. Boreham and J. D. Watts, Personality Type in Undergraduate Education and Physics Students, J. Psychol. Types **44**, 26 (1998).

[9] See, for example, D. J. Pittenger, Measuring the MBTI …, and Coming Up Short, J. Career Planning Employment **54**, 48 (1993).

[10] S. Koelsch, S. Skouras, and S. Jentschke, Neural correlates of emotional personality: A structural and functional magnetic resonance imaging study, PLoS One **8**, e77196 (2013).

[11] R. McCrae, P. T. Costa, Jr., and T. A. Martin, The NEO-PI-E: A more readable revised NEO personality inventory, Jour. of Personality Assessment **84**, 261 (2005).

[12] D. Hestenes, M. Wells, and G. Swackhammer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[13] J. J. B. Harlow, D. M. Harrison, and A. Meyertholen, Effective student teams for collaborative learning in an introductory university physics course, Phys Rev. Phys. Educ. Res. **12**, 010138 (2016).

[14] R. Wolfson, *Essential University Physics* (Pearson, London, 2016).

[15] R. Knight, *Physics for Scientists and Engineers: A Strategic Approach*, 3rd ed. (Pearson, Toronto, 2013).

[16] https://media.wix.com/ugd/23a234_46bbbb5084a74759aa5e053eb16ad4e3.pdf (Retrieved May 17, 2017).

[17] L. J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika **16**, 297 (1951).

[18] M. Tavakol and R. Dennick, Making sense of Cronbach's alpha, Int. J. Med Educ. **2**, 53 (2011).

[19] J. Han, L. Bao, L. Chen, T. Cai, Y. Pi, S. Zhou, Y. Tu, and K. Koenig, Dividing the Force Concept Inventory into two equivalent half-length tests, Phys. Rev. ST Phys. Educ. Res. **11**, 010112 (2015).

[20] J. Han, K. Koenig, J. Fritchman, D. Li, W. Sun, Z. Fu, and L. Bao, Experimental validation of the half-length Force Concept Inventory, Phys Rev. ST Phys. Educ. Res. **12**, 020122 (2016).

[21] For each student, we calculated a weight which was the color score with the highest value minus the mean of the other three scores. This weight was multiplied times test grades or FCI scores to form a weighted value. Then the weighted values were compared for different students color types.

[22] Particularly for students the word "error" is misleading since it implies that some mistake has been made. We strongly prefer the word "uncertainty" although the phrase standard error of the mean is standard and we here place *error* in quote marks to indicate that it is a poor choice of words.

[23] R. McGill, J. W. Tukey, and W. A. Larsen, Variations of box plots, Am. Statistician **32**, 12 (1978) (retrieved November 15, 2014). Note that in this article the multiplier is 1.57, not 1.58: since the uncertainty itself is largely heuristic, the difference in these values is trivial. Also, in Refs. and we incorrectly omitted the factor of 1.58 entirely.

[24] There are various conventions for the cutoff definition. We use 1.5 times the inter-quartile range extending from the upper, and lower quartiles, which was proposed in J. D. Emerson and J. Strenio, Boxplots and Batch Comparison, in *Understanding Robust and Exploratory Data Analysis*, edited by D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Wiley-Interscience, Toronto, 1983), p. 58. This cutoff definition is the usual one.

[25] See, for example, D. Harrison, Designing a Good Test (1999), http://www.upscale.utoronto.ca/PVB/Harrison/TestDesign/TestDesign.html.

[26] J. Cohen, A power primer, Psychol. Bull. **112**, 155 (1992).

[27] J. W. Tukey, Comparing individual means in the analysis of variance, Biometrics **5**, 99 (1949).

[28] M. Livio, *Brilliant Blunders* (Simon and Schuster, New York, 2013).

[29] A. Gopnik, A. N. Meltzoff, and P. K. Kuhl, *The Scientist in the Crib* (Morrow, New York, 1997).

[30] C. M. Steel and J. Aronson, Stereotype threat and the intellectual test performance of African Americans, J. Personality Social Psychol. **69**, 797 (1995).

[31] D. M. Harrison, Factors correlated with students' scientific reasoning in an introductory university physics course, in *The Physics Educator: Tacit Praxes and Untold Stories*,

edited by K. A. MacLeod and T. G. Ryan (Common Ground Publishing, Champaign IL, 2016), Chap. 11, pp. 186–212.

[32] L. E. Kost-Smith, S. J. Pollock, N. D. Finkelstein, G. L. Cohen, T. A. Ito, and A. Miyake, Gender differences in physics I, the impact of a self-affirmation intervention, AIP Conf. Proc. **1289**, 197 (2010)

[33] S. J. Leslie, A. Cimpian, M. Meyer, and E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines, Science **347**, 262 (2015).

[34] J. J. B. Harlow, D. M. Harrison, and A. Meyertholen, Correlating student interest and high school preparation with learning and performance in an introductory university physics course, Phys. Rev. ST Phys. Educ. Res. **10**, 010112 (2014).

[35] J. J. B. Harlow, D. M. Harrison, and E. Honig, Compressed-format compared to regular-format in a first-year university physics course, Am. J. Phys. **83**, 272 (2015).

[36] C. Hoellwarth and M. J. Moelter, The implications of a robust curriculum in introductory mechanics, Am. J. Phys. **79**, 540 (2011).

[37] A. K. Wood, R. K. Galloway, and J. Hardy, Can dual processing theory explain physics students' performance on the Force Concept Inventory?, Phys. Rev. Phys. Educ. Res. **12**, 023103 (2016).

[38] D. M. Harrison, Cognitive reflection and physics student performance (2016), http://www.upscale.utoronto.ca/PVB/Harrison/CognitiveReflection/CognitiveRelflectionPhysicsTests.pdf.

[39] The physics faculty, graduate students, and 1st year physics majors and specialists used the form of the True Colors test in Appendix A. For the faculty only 15 out of 26 respondents (58%) got total scores of 50. For our graduate students, 19 out of 30 (63%) got total scores of 50. The students in our 1st year course for physics majors and specialists did better: 25 of 29 (86%) got total scores of 50.

[40] The Royal Society of London, founded in 1660, was not the first such group. The Academia dei Lincei was founded in 1603 in Rome and included Galileo among its members.

[41] Experiment 40 in his New Experiments Physico-Mechanicall, Touching the Spring of the Air, and its Effects (1660), described in J. B. West, Robert Boyle's landmark book of 1660 with the first experiments on rarified air, J. Appl. Physiol. **98**, 21 (2005).

[42] Ho Shen Yong (private communication).

[43] A. Yeung, J. Read, and S. Schmid, Students' learning styles and academic performance in first year chemistry, *Uniserve Science Blended Learning Symposium Proceedings* (Uniserve Science, Sydney, NSW 2005), pp. 137–142. The phrase learning-styles used in the title of this paper is now largely obsolete; however, the instrument used in this paper was actually the MBTI.

[44] The website is http://www.islephysics.net/. See also, for example, D. Brookes and E. Etkina, Physical phenomena in real time, Science **330**, 605 (2010); E. Etkina and G. Planinisic, Thinking like a scientist, Phys. World **27**, 48 (2014).