Impact of equity models and statistical measures on interpretations of educational reform

Idaykis Rodriguez,[1] Eric Brewe,[2,1] Vashti Sawtelle,[1] and Laird H. Kramer[1]
[1]*Department of Physics, Florida International University, Miami, Florida 33199, USA*
[2]*Department of Teaching and Learning, Florida International University, Miami, Florida 33199, USA*
(Received 22 November 2011; published 26 July 2012)

We present three models of equity and show how these, along with the statistical measures used to evaluate results, impact interpretation of equity in education reform. Equity can be defined and interpreted in many ways. Most equity education reform research strives to achieve equity by closing achievement gaps between groups. An example is given by the study by Lorenzo *et al.* that shows that interactive engagement methods lead to increased gender equity. In this paper, we reexamine the results of Lorenzo *et al.* through three models of equity. We find that interpretation of the results strongly depends on the model of equity chosen. Further, we argue that researchers must explicitly state their model of equity as well as use effect size measurements to promote clarity in education reform.

DOI: 10.1103/PhysRevSTPER.8.020103 PACS numbers: 01.40.Fk, 01.40.G−, 01.40.gf

I. UNDERSTANDING EQUITY IN EDUCATION REFORM

Equity claims are becoming more prevalent as physics education research investigates how different learning environments impact diverse learners. In this paper we argue that interpretations of equity claims are understood through the underlying equity model chosen by researchers. Furthermore, we argue that the choice of equity model necessitates that researchers carefully provide an operational definition of equity, describe the measures used, and interpret the results within the model of equity. We present three standard models of equity (equity of individuality, of parity, and of fairness) and show how effect sizes and confidence intervals on the effect size provide nuanced, but crucial, information for evaluating outcomes within a model of equity. We also demonstrate how effect size statistics can be used to make equity claims when comparing groups.

Equity in physics education has been a goal of science reform in response to the need of ''science for all'' [1]. The science education literature on equity has focused primarily on differences in performance, opportunity, or access of certain underrepresented groups. Historically underrepresented groups in physics include students of diverse cultures, students with low socioeconomic status, and females. Researchers working on questions of equity usually attend to issues such as social justice, gender, race, ethnicity, socioeconomic status, equality of education, quality of education, or fairness [2–4]. One prominent tradition of equity research in science education is the focus on ''gaps'' [5–7]. Gaps most often refer to differences in performance between groups of students on a quantitative measure such as test scores and grades, but may also include differences in opportunity or access. For the purposes of this paper, we acknowledge that there are many forms of equity research, but we will focus on the interpretation of gaps in performance on quantitative measures. We argue that interpretations of group result comparisons often include a tacit model of equity, and that making these models explicit may change or elucidate the interpretations.

II. THREE MODELS OF EQUITY

An explicit definition of equity is necessary to guide educational reform and education research. Reforms striving for equity vary according to the ideology and perspective of the reformers, which can be seen in the nature of the research question and in the interpretation of the results [3–6]. One place the ideology is communicated lies in decision of whether to compare the measures of grades or gains in standardized tests. The model of equity also impacts decisions such as what groups to compare and how these comparisons are carried out, for example, comparing across treatments or within treatment, or comparing Hispanic students at a large research-intensive school with Hispanic students in a similar class at a community college. The researcher's model of equity, along with measures of effectiveness employed, contributes to making these decisions. In this section we present three models of equity; *equity of parity, equity of fairness*, and *equity of individuality.*

A. Equity of parity

One approach employed to increase the participation of historically underrepresented groups in science has been to improve their achievement compared to that of the majority group. This social justice perspective describes the equity of parity model. In this paper, parity will follow

Secada's [4] definition, indicating that parity is achieved when the distributions of achievement scores in two or more groups have the same postinstruction average despite differences in preinstruction averages. Equity of parity strives to close achievement gaps, requiring students in different groups to achieve different gains in order to close gaps between groups. Thus, one must acknowledge that the instruction benefits the ''less prepared'' students more than the ''well prepared'' students. Lynch [3] describes the equity of parity model as equal outputs, in which all students achieve the same outcome on exams or conceptual assessments, regardless of incoming scores. However, she also mentions that sources of the initial gaps of incoming scores may result from unequal inputs such as opportunities, resources, and conditions for learning.

B. Equity of fairness

Equity and equality are often used interchangeably [3,4]. The equity of fairness model distinguishes equity from equality by defining equity as that which promotes justice and freedom from bias or favoritism [2]. Students come into classrooms with differing conceptions of physics [8] as well as with ethnic, gender, and socioeconomic differences that may influence their incoming performance [3]. The equity of fairness perspective implies an impartial treatment that would be reflected in an equal gain (posttest score − pretest score) in conceptual understanding for all students groups, regardless of initial understanding or group.

This view of equity ignores the social factors that created the initial differences in the first place. Maintaining equal gain only perpetuates the initial differences; thus, an underrepresented group will always exist. Although the equity of fairness model of equity does not favor one group over another, it includes inherent disadvantages in not attending to students' identities and/or cultures [2].

C. Equity of individuality

The third model of equity moves beyond the group comparisons and examines individual excellence. Research focusing on achievement gaps relies on an underlying assumption that there is a difference between groups. Questioning the origin of the lower achieving group leads to racial, ethnic, and socioeconomic comparisons. Gutierrez argues that relying on a comparison group perpetuates the idea that marginalized students are not worth studying in their own right and therefore sustains upper-middle class whites as the normative group [5,6].

The equity of individuality model promotes individual excellence of students within groups. Taking this perspective does not limit research to examining achievement scores; it also informs researchers about group participation within a discipline or practice. An equity of individuality perspective may lead to investigating how a specific group's participation increases from year to year, or inquiring whether a group is represented within a field as compared to the total population. However, neglecting the comparison group potentially perpetuates the current differential status between groups. Further, even if achievement, participation, and representation of a group increase over time, the increases carry little value for stakeholders without comparisons to another group [7].

III. EFFECT SIZE AS A BETTER MEASURE OF EQUITY

The use of effect sizes and confidence intervals on effect sizes provides a consistent and reliable mechanism to analyze results statistically and evaluate the equity model outcomes. In previous research in physics education, equity claims have focused on significant differences in normalized gains (that take into account differences on pretest scores) [9–13], possibly muddying the interpretation and making comparisons across research studies difficult. However, comparison of the effect sizes and confidence intervals from pretest to posttest or across treatments provides a transparent way of evaluating whether the desired equitable outcome has been achieved. Further, they simplify cross-study comparison and meta-analysis.

A. Null hypothesis significant testing

A standard approach to examining the impact of a course reform on students' test scores has been null hypothesis significance testing in which statistically significant differences in the achievement results of student group 1 and student group 2 are reported [10–13]. Statistical significance is inferred when the means of the two student groups are statistically different; i.e., the null hypothesis is rejected. The p value indicates the strength of evidence for a difference between the two groups. Statistical significance does not indicate what led to the difference between the two groups, but only that the two are different. It also does not indicate the magnitude of the differences between groups. Further, as the number of participants grows large even very small differences in means are likely to become ''significant'' in null hypothesis significance testing [14].

B. Effect size and confidence intervals

Effect size (Cohen's d, Glass's delta, η^2, or adjusted R^2) provides a measure of the effect the experimental condition or the grouping has had on the outcome, and thus indicates the strength of the conclusion, as well as characterizing the extent to which the null hypothesis is not valid [15]. The confidence interval on an effect size estimates the uncertainty associated with the effect in the population based on the effect measured in the sample. In the case of equity research, effect sizes convey how equitable a treatment is through comparisons between groups within a treatment. For example, comparing the effect of treatment for men and women on pretest scores, posttest scores, and/or gain

indicates how equitable the treatment was for men and women.

Statisticians have criticized the rampant use of null hypothesis significance testing [14,16] and suggested that researchers turn to reporting statistics that focus attention on the size of the effect. The American Psychological Association (APA) also encourages authors to report effect sizes in statistical analysis. The 6th edition of the APA Publication Manual [17] stresses that null hypothesis significance testing "is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results" (p. 33). Because of recommendations from the APA, editorial boards of 23 journals require effect size reporting [18]. Details regarding the use of effect size and the substance of the statistics can be found in [15].

C. Effect size provides for meta-analysis

Use of effect sizes and confidence interval on effect sizes paves the way for meta-analyses, enhancing our ability to interpret results across studies. Thompson [19] states that a focus on providing confidence intervals around effect size encourages meta-analytic thinking; that is, researchers are better able to compare results of different studies when effect sizes and confidence intervals on these effect sizes are reported. Thompson [19] contends that the most informative use of confidence intervals lies in the ability to compare intervals across studies, and thus compare the size of the obtained effect to the size of effects in similar studies. Smithson [20] provides an excellent description of the need for confidence intervals and details on how confidence intervals can be calculated for various statistical tests.

IV. APPLYING EFFECT SIZE AND INTERPRETING EQUITY

In order to demonstrate the insight achieved through the use of effect size, we reanalyze the study of Lorenzo *et al.* [21] of interactive engagement by calculating effect sizes and confidence intervals on the effect sizes to provide further interpretation of the results.

The use of effect size and confidence intervals facilitates cross-study comparisons for equity research. Lorenzo *et al.* claim to have decreased the gender gap in introductory physics by using interactive engagement methods and promoting in-class interaction and collaboration. They studied three treatments, traditional, interactive engagement 1 (IE1), and interactive engagement 2 (IE2), and compared each treatment in terms of effectiveness. IE1 primarily implemented Peer Instruction [22], which involves mini lectures and conceptual questions discussed by students in small groups in class. In IE2 they added a two-hour workshop using *Tutorials in Introductory Physics* [23] and cooperative quantitative problem-solving activities [24] to the implementation of peer instruction. *Tutorials in Introductory Physics* [23] emphasizes conceptual reasoning and hands-on activities while cooperative problem-solving activities reinforce the students' problem-solving skills. Lorenzo *et al.* measured their students' achievement using standard pretesting and posttesting with Force Concept Inventory (FCI) to their classes. The FCI assesses a student's conceptual understanding of Newtonian mechanics [25].

In their paper, Lorenzo *et al.* claim the gender gap is reduced through the use of higher levels of interactive engagement. They use null hypothesis significance testing between women and men on FCI pretest scores and FCI posttest scores as evidence to support these claims. They argue that there are significant differences between the FCI pretest scores of men and women in both IE1 and IE2. They then claim that in IE1 the differences in the FCI posttest mean scores of men and women decrease, but remain significant. In IE2, they claim that the posttest scores of men and women in IE2 are not significantly different and thus the gender gap has closed. These data are shown in the aggregate averages in Table I.

Though Lorenzo *et al.* do not explicitly reference equity in their study, we can interpret their investigation of interactive engagement narrowing the gender gap as an *equity of parity* model where the goal is for all students to achieve the same outcomes. We will now illustrate how using effect size to analyze their data enhances the understanding of their gender data and the treatment's impact on equity. In the following sections we calculate the effect of gender on the scores and interpret the result for each treatment. We will evaluate the results for IE1 and interpret them from the perspectives of the equity models. We will then do the same for IE2 treatment. Finally, we evaluate the effect of treatment on the scores and consider new interpretations for equity.

TABLE I. Summary of data from Lorenzo *et al.* where the normalized gain is $\langle g \rangle = (\text{posttest} - \text{pretest})/(100 - \text{pretest})$.

	Men ($N_{IE1} = 432$, $N_{IE2} = 161$)				Women ($N_{IE1} = 251$, $N_{IE2} = 99$)			
	Pretest % (SD)	Posttest % (SD)	Raw gain %	$\langle g \rangle$	Pretest % (SD)	Posttest % (SD)	Raw gain %	$\langle g \rangle$
IE1	73 (16)	88 (8.5)	15	0.56	61 (16)	81 (12)	20	0.51
IE2	71 (19)	91 (11)	20	0.69	61 (19)	89 (9.4)	28	0.72

TABLE II. Effect sizes and confidence intervals (C.I.) comparing gender for students in IE1 and IE2, using the FCI pretest and FCI posttest from Lorenzo *et al.*

	Cohen's d of gender on FCI pretest (95% C.I.)	Cohen's d of gender on FCI posttest (95% C.I.)
IE1	0.75 (0.59–0.91)	0.70 (0.54–0.86)
IE2	0.53 (0.27–0.78)	0.19 (0.06–0.44)

A. Effect of gender on scores for each treatment

We begin by calculating effect size and confidence interval for the Lorenzo *et al.* data [21]. We utilize Cohen's d, as it is the most prevalent effect size used to compare two groups, given by

$$d = \frac{\mu_1 - \mu_2}{\sigma_{\text{pooled}}}, \tag{1}$$

where μ_i is the sample mean and σ_{pooled} is the pooled standard deviation of the two samples. We calculated Cohen's d for the effect of gender on FCI pretest and FCI posttest scores, within each treatment (IE1 and IE2), as well as the confidence intervals on this effect size, from the aggregated averages and standard deviations. These are provided in Table II.

1. Evaluating the claims of Lorenzo et al. of decreasing the gender gap in IE1

We begin by reexamining the claim of Lorenzo *et al.* of decreased gender gap in IE1 from pretest to posttest instruction. First, the effect sizes and confidence intervals on these effect sizes provide a subtly different interpretation of statistical significance from that of Lorenzo *et al.* [21]. In IE1, the effect of gender remains the same on both the FCI pretest and the FCI posttest, as indicated by nearly identical effect sizes that are both within the medium rage of effect according to Cohen's d guidelines [26]. The effect sizes also have mostly overlapping confidence intervals (Table II). The effect size remaining the same pretest to posttest provides no evidence that the gender gap, or the effect of gender on the FCI scores, decreased in IE1, which contrasts the findings of Lorenzo *et al.* of a deceased gender gap.

2. Interpreting results of IE1 from differing models of equity

We now analyze the IE1 results on the basis of the equity models. Examining the results through a lens of gap closing in an equity of parity model, we disagree that IE1 has decreased the effect of gender. As measured, the effect of gender on student scores shows the gap maintained for the IE1 treatment group. No effect of gender on students' scores would be indicated by an effect size close to zero and confidence intervals that cross over zero.

From an equity of fairness perspective in which the scores of each group should have gained the same, we see from Table I that a comparison of raw gain for men (15%) and women (20%) does not clearly determine if IE is free from bias. Examining effect sizes from Table II, we observe the effect of gender on FCI pretest ($d = 0.75$) and posttest scores ($d = 0.70$) to be nearly identical and within the same medium range effect, with mostly overlapping confidence intervals. We conclude that IE1 has the same effect on the scores of men and women, maintaining a continuous gap. Maintaining the gap gives students a fair and just treatment, thereby achieving equity of fairness.

Looking at IE1 from an equity of individuality model, we must recall that this model does not require a comparison group. Using the data in Tables I and II, we find that gender has an effect on pretest scores and therefore the groups should be treated separately and not compared. We will look further at interpretations of the equity of individuality model when we discuss the effect of treatment on individual group scores in Sec. IV B 2

3. Evaluating the claims of Lorenzo et al. of closing the gender gap in IE2

Lorenzo *et al.* [21] show no significant differences between the scores of men and women on the FCI posttest scores in IE2, leading to their conclusion that they effectively eliminated the gender gap. Use of effect sizes allows us to reevaluate the result. First, as seen in Table II, the effects of gender on FCI pretest ($d = 0.53$) and FCI posttest ($d = 0.19$) are different, since 0.53 is a medium effect and 0.19 is a small effect, indicating the gender gap decreased because the FCI posttest effect is smaller. However, the confidence intervals allow us to determine whether the gap has closed. In IE2, the confidence intervals on the effect size overlap (pretest 0.27–0.78 and posttest 0.06–0.44). Using Cohen's guidelines on magnitudes of effects [26], this would indicate that the effect of gender on the pretest score is small to medium, and the effect on the posttest score is no effect to small, indicating the gender gap decreased. As the 95% confidence interval on the effect does not include zero and does include small effects, we conclude that the effect of gender is not entirely eliminated, contrary to the claim by Lorenzo *et al.*

4. Interpreting results of IE2 from differing models of equity

An analysis of effect sizes and confidence intervals for the Lorenzo *et al.* [21] data shows that effect sizes and confidence intervals provide a more thorough and complete understanding of the results. Gender effects on FCI pretest and FCI posttest scores in IE2 decreased from pretest to posttest, but the confidence intervals indicate a small gender effect still remains in the posttest scores. From an equity of parity perspective, the results indicate that IE2 is moving toward achieving the goal of closing

gaps, since the posttest effect size is small and the confidence interval almost overlaps zero.

For IE2, equity of fairness goals were not achieved, as the effect of gender decreased from pretest to posttest. Since women had a lower pretest score than men, the difference in effect size indicates that women achieved higher gains than men.

Within the equity of individuality model, we come to a similar conclusion as in IE1 for IE2: because there is an initial gender effect on student pretest scores, the student scores should be treated individually and not compared across gender groups.

B. Effect of treatment on scores for each gender

Equity researchers share a goal of developing treatments that are effective for all groups of students. Effect sizes on the gain [see Eq. (2)] allow us to confirm the effect of a treatment on a group of students:

$$d = \frac{\mu_{\text{post}} - \mu_{\text{pre}}}{\sigma_{\text{pooled}}} \tag{2}$$

In their original study, Lorenzo *et al.* [21] used normalized gains of men's and women's scores to support their claim that the IE2 treatment had eliminated the gender gap but not the IE1 treatment. We calculate the effect of each of the two treatments on the raw gain FCI scores for men and women using Eq. (2). Table III includes the Cohen's d value and the confidence interval of the effect sizes.

1. Extending analyses beyond inferences

Interpreting the effect sizes of each of the treatments for men and women leads to differing conclusions from those originally deduced from an analysis of normalized gain. In each treatment, as seen in Table I, the normalized gain is roughly the same for men and women (0.56 versus 0.51 in IE1 and 0.69 versus 0.72 in IE2). Considering these normalized gains alone led Lorenzo *et al.* to conclude that "equity" has been achieved since the normalized gain for each gender is not different for both IE1 and IE2. Effect sizes, however, tell a more complete story. An effect size analysis indicates that both IE1 and IE2 had a greater positive effect[1] on women's scores than on men's scores. The confidence intervals for men and women in IE1 do overlap, and it cannot be determined whether it clearly benefits women more than men. In IE2 the confidence intervals do not overlap, and we can confidently conclude that treatment IE2 had a larger effect on women than on men. We now turn to understanding how the three underlying equity models impact the interpretations of these results.

TABLE III. Effect sizes and confidence intervals of treatment for men and women in IE1 ($N_{\text{men}} = 432$, $N_{\text{women}} = 251$) and IE2 ($N_{\text{men}} = 161$, $N_{\text{women}} = 99$) derived from the data of Lorenzo *et al.*

	Cohen's d of IE1 on FCI gain (95% C.I.)	Cohen's d of IE2 on FCI gain (95% C.I.)
Men	1.17 (1.03–1.31)	1.29 (1.05–1.53)
Women	1.41 (1.22–1.61)	1.87 (1.53–2.19)

2. Expanding interpretations of the effect of treatment using models of equity

In an equity of parity model the goal is to close achievement gaps. In order to evaluate if this goal was achieved, we require a focus on final scores. Comparing the effect of treatments for men and women does not allow us to analyze the equity of parity, as this model requires comparison of final scores.

The equity of individuality model encourages development of individuals; thus, we analyze groups without comparisons. Researchers using an equity of individuality model could argue that having large effects on women's gain in both IE1 and IE2 simply focuses attention on responding to a specific group of individuals in the classroom. As seen in Table III, we see that both treatments had large effects for both men and women; thus, equity of individuality was achieved for both men and women in treatments IE1 and IE2. We do see that both treatments favor the female students, as their effect size is larger than men.

V. DISCUSSION

A. Two sides of equity models

No model of equity is ideal, as each of the three models varies in their goals. The most common model discussed in literature is equity of parity, where groups of students finish with the same test scores, reducing the achievement gap. Such a model is supported by the national science research recommendations that are a part of Science for All [1]. This equity perspective focuses the attention on minorities, low socioeconomic status students, and marginalized students such as females in science, in order to create reforms and curricula that enhance their learning. To achieve an equity of parity goal is admirable, as it greatly influences our ability to produce top students from a variety of diverse backgrounds and with varying levels of preparation who have the opportunity to contribute new ideas and diversity to our science community [2]. However, even though there is a need and an obligation to create such opportunities, an

[1]All of the effect sizes of treatment on the FCI gain are "large" according to Cohen's benchmarks [26], yet Cohen advised users that these are guidelines and not absolute indicators. Instead, effects should be considered within the context of the study. For example, discovering a small effect in a cancer drug trial may be important if no other drug has achieved similar results. Similarly, in education research a semester or year-long treatment should have considerable effects on scores. Therefore, especially with pretest or posttest designs, we may need to recalibrate what constitutes a large effect.

equity of parity goal can be achieved only if the underrepresented group gains more than the majority, where the majority is often white and middle- to upper-class students. Comparing such groups can set groups in opposition to one another and fuel insecurities and prejudice between groups [5,6]. An alternative pathway to achieving equity of parity could be to address Lynch's [3] equal inputs of access and opportunity and target the gaps before they can appear. In other words, we can address these differences early and throughout the educational process to prevent the differences from arising.

The equity of fairness model provides equitable treatment with all groups demonstrating equal gains in achievement, not favoring one group over another. However, being free from bias does come with caveats. For example, the data analysis of Lorenzo *et al.* [21] using effect sizes suggests that their IE1 classroom achieved an equity of fairness outcome where all students gained the same amount. However, a common criticism of achieving equal gains between men and women lies in the perpetuation of initial gender differences.

To avoid placing groups in opposition, one could study specific groups of students in their own right [5,6,27]. Females in interactive engagement classes of Lorenzo *et al.* are at an advantage (indicated by the effect size analysis), as the course had a greater impact on their FCI scores than it did on the scores of men. Showing that interactive engagement differently impacts females in one treatment over another is an important finding on its own without having to compare females to males. Researching a specific group (e.g., female, black, or Hispanic) for its own advancement and excellence embraces that group's individuality and is responsive to diversity [27]. This is the perspective the equity of individuality model takes, which focuses on a group's identity and culture. Although looking at a group's excellence over time satisfies the equity of individuality goals, the value of the increase in excellence is not immediately meaningful to stakeholders without comparing a minority group to a majority group [7].

B. Equity measurements

We advocate that researchers interested in comparing groups should use effect sizes and confidence intervals on those effect sizes to interpret their results regardless of their model of equity. Focusing analysis on effect sizes and confidence intervals provides a comprehensive analysis and interpretation of any data. We demonstrated that, in contrast to the claims of Lorenzo *et al.* that they had closed the gender gap, an analysis of effect sizes showed gender still impacted FCI scores and that the effect was never completely eliminated. The nuanced information contained in the effect sizes and confidence intervals on the effect size analysis enhanced the interpretation by providing estimates of the *size* of the differences between groups.

Effect size and confidence intervals also provide an avenue for opening discourse about community norms that will support meta-analytic thinking. As a community, we should be moving toward measures and norms that allow for straightforward comparison across treatments and groups; effect size analysis satisfies that goal. The shift away from null hypothesis significance test analysis is motivated by increasing expectations for better analysis and making more specific equity claims in physics education research. Having an explicit equity model that guides the research leads to more nuanced interpretations.

VI. MAKE EQUITY EXPLICIT

Equity in science education research does not carry an inherent definition; it carries many. In this paper we proposed three different models of equity that incorporate the common trends in equity literature. First, *equity of parity* perceives equity as equal outcomes in achievement for all groups. Second, *equity of fairness* follows the literal definition of equity, not choosing any one group over another, which may be interpreted as equal achievement gains between groups. Third, *equity of individuality* considers the uniqueness of a specific group for the purpose of advancing that group. The three models each have their advantages and disadvantages. As researchers, we make explicit decisions and must interpret our data within our model of equity.

The predominant trend in research is to think of equity as equality, or closing gaps, and, therefore, the trend is to compare students and check for equal outcomes. However, it is important to realize that equity goes beyond quantitative differences in achievement and can be concerned with the individual, with excellence, and with fairness and justice [4]. We have seen how different models of equity underlie particular decisions and questions, and we have demonstrated how each of the models supports different interpretations of the data. Equity has many social and cultural connotations that often are inconsistent. Therefore, as a research community physics education research should strive to define what we mean by equity explicitly and recognize how it guides our research, how it affects our choice of measures, and, ultimately, how it influences the claims we make.

[1] F. J. Rutherford and A. Ahigren, *Science for All Americans* (Oxford University Press, New York, 1989).

[2] O. Lee, Equity implications based on the conceptions of science achievement in major reform documents, Rev. Educ. Res. **69**, 83 (1999).

[3] S. Lynch, *Equity and Science Education Reform* (Lawrence Erlbaum Associates, Mahwah, NJ, 2000).

[4] W. G. Secada, *Equity in Education* (Falmer Press, New York, 1989).

[5] R. Gutierrez, A ''gap-gazing'' fetish in mathematics education? Problematizing research on the achievement gap, J. Res. Math. Educ. **39**, 357 (2008) [http://www.jstor.org/stable/40539302].

[6] R. Gutierrez and E. Dixon-Roman, *Mapping Equity and Quality in Mathematics Education* (Springer Science, New York, 2011)

[7] S. T. Lubienski, On ''gap gazing'' in mathematics education: The need for gaps analyses, J. Res. Math. Educ. **39**, 350 (2008) [http://www.jstor.org/stable/40539301].

[8] E. F. Redish, New models of physics instruction based on physics education research, in *Proceedings of the Deustchen Physikalischen Gesellschaft Didaktik der Physik, Jena, Germany, 1996*, edited by K. H. Lotze, http://eric.ed.gov/PDFS/ED438184.pdf, p. 51.

[9] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[10] L. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[11] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and Introductory Physics Performance, in *Proceedings of the 2008 Physics Education Research Conference*, edited by C. Henderson, M. Sabella, and L. Hsu (AIP, New York, 2008), p. 15.

[12] L. McCullough, Gender, context, and physics assessment, J. Int. Women's Stud. **5**, 20 (2004) [http://www.bridgew.edu/SOAS/jiws/May04_Special/Gender.pdf].

[13] V. P. Coletta and J. A. Phillips, Interpreting FCI scores: Normalized gain, pre-instruction scores, and scientific reasoning ability, Am. J. Phys. **73**, 1172 (2005).

[14] J. Cohen, The earth is round ($p < .05$), Am. Psychol. **49**, 997 (1994).

[15] R. J. Grimsson and J. J. Lim, *Effect Sizes for Research: A Broad Practical Approach* (Lawrence Erlbaum Associates, Mahwah, NJ, 2005).

[16] R. Carver, The case against statistical significance testing, revisited, J. Exp. Educ. **61**, 287 (1993) [http://www.jstor.org/stable/20152382].

[17] *American Psychological Association Publication Manual* (American Psychological Association, Washington, DC, 2010), 6th ed.

[18] M. Capraro and R. Capraro, Exploring the APA fifth edition Publication Manual's impact on the analytic preferences of journal editorial board members, Educ. Psychol. Meas. **63**, 554 (2003).

[19] B. Thompson, What future quantitative social science research could look like: Confidence intervals for effect sizes, Educ. Researcher **31**, 25 (2002).

[20] M. Smithson, *Confidence Intervals*, Quantitative Applications in the Social Sciences Series Vol. 104 (Sage Publications, Thousand Oaks, CA, 2002).

[21] M. Lorenzo, C. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[22] E. Mazur, *Peer Instruction: A Users Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997).

[23] L. C. McDermott and P. S. Schaffer, *Tutorials in Introductory Physics* (Prentice-Hall, New Jersey, 2001).

[24] P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. 1. Group versus individual problem solving, Am. J. Phys. **60**, 627 (1992); P. Heller and M. Hollabaugh, Teaching problem solving through cooperative grouping. 2. Designing problems and structuring groups, *ibid.* **60**, 637 (1992).

[25] S. E. Lewis and J. E. Lewis, Seeking effectiveness and equity in a large college chemistry course: An HLM investigation of peer-led guided inquiry, J. Res. Sci. Teach. **45**, 794 (2008).

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

[27] C. P. Benbow and J. C. Stanley, Inequity in equity: How ''equity'' can lead to inequity for high-potential students, Psychol. Publ. Pol. Law **2**, 249 (1996).