Adv Educ Res Eval, 2021, 2(1): 138-145
DOI: 10.25082/AERE.2021.01.005

RESEARCH ARTICLE

Holding teachers accountable: An old-fashioned, dry, and boring perspective

David Trafimow

Psychology Department, New Mexico State University, Las Cruces, NM 88003-8001, USA

Correspondence to: David Trafimow, Psychology Department, New Mexico State University, PO Box 30001, Las Cruces, NM 88003-8001, USA;
E-mail: dtrafimo@nmsu.edu

Received: March 16, 2021;
Accepted: April 23, 2021;
Published: April 25, 2021.

Citation: Trafimow D. Holding teachers accountable: An old-fashioned, dry, and boring perspective. *Adv Educ Res Eval*, 2021, **2**(1): 138-145. https://doi.org/10.25082/AERE.2021.01.005

Abstract: Few would disagree with the desirability to hold teachers accountable, but student evaluations of teaching and department head evaluations of teaching fail to do the job validly. Although this may be due, in part, to difficulties conceptualizing teaching effectiveness and student learning, it also is due to insufficient attention to measurement reliability. Measurement reliability sets an upper bound on measurement validity, thereby guaranteeing that unreliable measures of teaching effectiveness are invalid too. In turn, for measures of teaching effectiveness to be reliable, the items in the measure must correlate well with each other, there must be many items, or both. Unfortunately, at most universities, those who are tasked with teaching assessment do not understand the basics of psychometrics, thereby rendering their assessments of teachers invalid. To ameliorate unsatisfactory assessment procedures, the present article addresses the relationship between reliability and validity, some requirements of reliable and valid measures, and the psychometric implications for current teaching assessment practices.

Keywords: teachers accountable, teaching effectiveness, teaching assessment, reliability, validity

1 Introduction

We live in a society where there is a trend in the direction of increasing accountability. Like other areas of human activity that affect people, the trend has influenced teaching too. The hope is that holding teachers accountable will increase teaching effectiveness, increase student learning, and eventually translate to bettering the human condition. To the extent that holding teachers accountable really does increase teaching effectiveness, and all the rest, accountability is a good thing. However, Stroebe (2020) [1] came to the opposite conclusion in his comprehensive review of the empirical literature, which is that our current practices for evaluating teaching, particularly depending on student evaluations of teaching, decreases teaching effectiveness and student learning. How can it be that holding teachers accountable is good, in theory; but bad, or at least not good, in practice? (Hanushek and Rivkin (2010) [2] provided an excellent review showing that "observed teacher characteristics do not represent teacher quality" (p. 267). This review dovetails nicely with the more recent review by Stroebe (2020) [1].

To answer the question, consider a fanciful scenario where teachers are evaluated based on how much they express a liking for eating chocolate—the more the liking for eating chocolate, the better the evaluation. Because this evaluation procedure is blatantly invalid, it renders salient that no matter how good a practice is, in theory; the evaluation procedure must be valid to render it similarly good, in practice. However, validity is a very technical subject, often considered old-fashioned, dry, and boring; though necessary to get a handle on the disconnect between the theoretical desirability of holding teachers accountable and that it does not work properly when attempted [3, 4]. And there is no way to reasonably discuss validity, without discussing reliability first as a prerequisite for validity [5].

2 Reliability

Although many alleged experts at evaluation consider reliability a technical topic that has little to do with evaluating teaching effectiveness (Citations omitted to protect the guilty), it is easy to disprove this point of view by simple mathematics that harken back to Charles Spearman's (1904) [6] ground breaking work [7, 8]. Skipping the technical details of the mathematical proof, Spearman showed that the reliability of the measures of two constructs is

mathematically related to the extent to which the measures could be expected to correlate (the validity coefficient). Equation (1) provides the equation, using modern symbols:

$$\rho_{XY} = \rho_{T_X T_Y} \sqrt{\rho_{XX'} \rho_{YY'}} \tag{1}$$

where

(1) ρ_{XY} is the observed correlation between X and Y (i.e., the validity coefficient);

(2) $\rho_{T_X T_Y}$ is the true correlation between X and Y (the correlation that would be observed if the two variables were measured with perfect reliability);

(3) $\rho_{XX'}$ is the reliability of X;

(4) $\rho_{YY'}$ is the reliability of Y.

To dramatize the importance of reliability, imagine that the reliability of one of the measures is zero. Instantiating 0 into Equation (1) for $\rho_{XX'}$, we see that the validity coefficient also must be zero, no matter the values of the other variables: $\rho_{XY} = \rho_{T_X T_Y} \sqrt{\rho_{XX'} \rho_{YY'}} = \rho_{T_X T_Y} \sqrt{0 \times \rho_{YY'}} = 0$. Going to the other extreme, imagine that both variables are measured with perfect reliability. Instantiating 1 into Equation (1) for $\rho_{XX'}$ and $\rho_{YY'}$ reduces Equation (1) to the following: $\rho_{XY} = \rho_{T_X T_Y}$. In other words, the observed correlation equals the true correlation, which is the ideal situation to have as far as reliability is concerned.

Of course, it rarely happens that reliabilities equal 0 or equal 1; they usually are somewhere between these extreme values. To understand what happens in the in-between cases, let us make a simplifying assumption, that we can agree on what we mean by student learning and that we can measure it with perfect reliability and validity. This is not true, which will be explored later, but the ideal assumption is nevertheless clarifying. Using student learning as the criterion variable, and instantiating 1 for $\rho_{YY'}$ in Equation (1), implies Equation (2) below:

$$\rho_{XY} = \rho_{T_X T_Y} \sqrt{\rho_{XX'}} \tag{2}$$

Based on Equation (2), Figure 1 relates the validity coefficient predicting student learning from teaching effectiveness (ρ_{XY}) to the true correlation between both variables ($\rho_{T_X T_Y}$), letting the reliability of the measure of teaching effectiveness ($\rho_{XX'}$) vary between 0 and 1. In Figure 1, the top curve refers to when the true correlation equals 1 and the successively lower curves represent when the true correlation equals 0.7, 0.4, and 0.1, respectively. Figure 1 shows three important psychometric facts of life. First, and most obvious, the true correlation matters. Second, the greater the reliability of the measure of teaching effectiveness, the greater the validity coefficient. Third, there is an interaction between the two foregoing effects. Specifically, the greater the true correlation, the more that reliability matters for influencing the validity coefficient. Note that the top curve, when the true correlation is maximized at 1, has a full range going from 0 to 1, depending on the reliability of teaching effectiveness, whereas the other curves are not only lower, but also have decreased ranges.

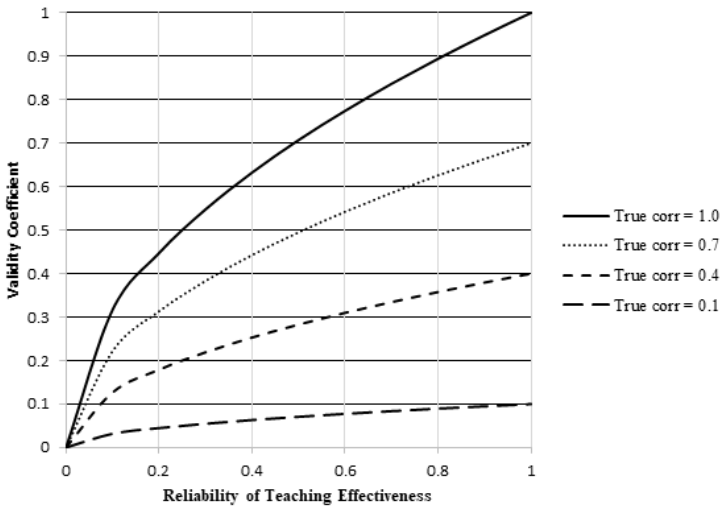

Figure 1 The validity coefficient predicting student learning from teaching effectiveness ranges along the vertical axis as a function of the reliability of teaching effectiveness along the horizontal axis when the true correlation equals 1.0 (top curve), 0.7 (second curve), 0.4 (third curve), or 0.1 (bottom curve).

The implications of Figure 1 and the equations strongly indicate the importance of reliability for validity, and so it is surprising how often alleged experts in assessing teaching effectiveness ignore it. For example, Stroebe's comprehensive review, though replete with various types of validity coefficients, ignored the issue of reliability. This is not a criticism of Stroebe, who performed an admirable job of conveying the relevant literature, with implications, but rather an indictment upon the field more generally for insufficient attention to reliability.

Moreover, the problem might be worse than insufficient attention. To relate a personal experience, I once attempted to create a reliable and valid measure of teaching effectiveness in psychology. In the preliminary stages, a higher-up was highly critical of the strong focus on reliability, despite the explanation I had provided that reliability is a prerequisite for validity (Citation omitted to protect the guilty). It was clear from the tenor of the message that, from the higher-up's point of view, my goal was wrong. The higher-up's goal was not to have a reliable and valid assessment device, but rather to have an assessment device that appeared valid and provided support for outcomes the higher-up desired to have. Of course, the higher-up had no patience for slogging through old-fashioned, dry, and boring reliability issues!

And lest we forget, Figure 1 could be characterized as optimistic because the assumption was that we had a perfect measure of student learning. Under a more realistic assumption about the measure of student learning, the values in Figure 1 would drop, thereby rendering the reliability lesson even more poignant.

It is possible to argue that reliability does not matter in the context of an experiment (An anonymous reviewer made this suggestion). For example, suppose that teaching evaluations are performed under two different experimental conditions, and a statistically significant result is obtained. In that case, the experiment worked, with one condition resulting in better teaching evaluations than in the other condition, so why worry about reliability? But such reasoning is extremely flawed. Statistical significance only indicates that the data deviate from the statistical model; but statistical significance has little to say about whether the problem in the statistical model is the null hypothesis, the assumption of random selection from the population, the assumption of random assignment of participants to conditions, or countless other possibilities. Even more important, statistical significance ignores the size of the effect. For example, suppose the effect size is 0.10, a very small effect, though statistically significant. In that case, even though the experiment 'worked' according to the significance test, the effect size is so small that it is meaningless in a practical sense. Once it is admitted that the size of the effect matters, reliability matters too because reliability sets an upper limit on experimental effect sizes just as it does for correlational effect sizes. Specifically, there is the issue of the reliability of the dependent variable, but there also is the issue of the reliability of the independent variable (e.g., does it 'take' equally for everyone?). Just as Equation (1) shows that the true correlation is multiplied by the square root of the product of the reliabilities to obtain the observed correlation; the true effect size in an experimental paradigm is multiplied by the square root of the product of the reliabilities to obtain the observed effect size. Therefore, lack of reliability is just as much of a problem in an experimental context as it is in a correlational context.

3 Validity

As Strobe (2020) [1] showed, recent literature has reversed older literature by showing that student evaluations fail to predict student learning, when the studies are properly conducted [9–12]. Moreover, we can broaden the perspective to include specific sources of invalidity, and even to consider issues outside student evaluations of teaching. Regarding the former, it is possible to argue that if student evaluations of teaching correlate with variables with which they should not correlate, that provides evidence of invalidity.

For example, there is ample evidence that student evaluations of teaching correlate with teacher attractiveness [13–21]. Under the assumption that attractiveness has nothing to do with teaching effectiveness or student learning, this correlation could be argued to be evidence of invalidity. A counterargument might be that teacher attractiveness might increase student learning. For example, attractive teachers might garner increased student attention, thereby increasing student learning. Perhaps alarmingly, however, Wolbring and Riordan (2016) [21] performed research that differed from the usual correlational paradigms and showed experimentally that the same lecture received differential ratings of quality from participants, depending on whether they had been shown pictures of an attractive or unattractive person as the ostensible lecturer. Hence, although it is possible to invent scenarios justifying the effects of attractiveness on ratings, such potential justifications have received little support from the literature.

Likeability might be considered a generalization of attractiveness. For example, attractive teachers might also be more likeable, and perhaps likeability increases student evaluations of

teaching [22]. There is ample evidence that likeability is correlated with student evaluations of teaching [23–25] and this correlation leads to the question of whether teacher likeability has anything to do with student learning. If one believes that students learn more under likeable teachers, then it is possible argue that the correlation of student evaluations of teaching with likeability is both justified and supports the validity of student evaluations of teaching. In contrast, however, given the lack of any evidence that teacher likeability influences student learning, a more plausible argument is that the correlation between likeability and student evaluations of teaching constitutes more evidence against the validity of student evaluations of teaching.

And then we come to two politically charged issues. Specifically, student evaluations of teaching are associated both with minority status [26–28] and gender [13, 17, 20, 29–31]. Although it is possible to invent scenarios where teachers with minority member status are less effective than white teachers, or that female teachers are less effective than male teachers, these scenarios seem implausible in the absence of positive evidence. That implausibility is suggestive that the correlations of student evaluations of teaching with minority status and gender provide further evidence of invalidity.

And moving beyond student evaluations of teaching, let us consider that although department heads evaluate their faculty members with the aid of student evaluations of teaching, it is department heads that make evaluations in most departments in most universities. In fact, some universities have dispensed with numerical student evaluations of teaching and the department heads have to depend on qualitative student comments about the teacher and their own intuitions. And yet, it seems unlikely that removing numerical evaluations really solves the invalidity problem. That is, given that quantitative student evaluations of teaching are influenced by attractiveness, likeability, minority status, gender, and others, is there any reason to believe that qualitative student comments would be immune from such influences? Worse yet, whether student evaluations of teaching are quantitative or qualitative, it seems implausible that department heads would reliably—much less validly—assess their faculty. Imagine two different acting department heads, with different relationships with their faculty, different moods at the time of evaluation, and countless other differences. It simply is not plausible that these two people's assessments would correlate very well across a wide swath of faculty, thereby indicating low interrater reliability. And as Equation (1) and (2) and Figure 1 demonstrated, with insufficient reliability, validity is compromised. Worse yet, remember again that Figure 1 was constructed based on an unrealistically optimistic assumption of a perfect measure of student learning. A more realistic assumption would result in a more pessimistic conclusion. For example, suppose that the true correlation between department head evaluations of teaching effectiveness and student learning was an optimistic value of 0.50 (remember that student variance likely would be much more important than teacher variance, so a more realistic true correlation would be around 0.2 or 0.3). And suppose that department head reliability is at the 0.30 level (based on informal data, I consider this optimistic) and that the reliability of the student learning measure is at the conventional level of acceptability of 0.70 (this is probably quite optimistic given faculty disagreements about what constitutes student learning or differing degrees of student learning). Although 0.70 is considered by most authorities to be the lower bound of "acceptability" for a reliability coefficient, consider that this translates into a coefficient of determination of 0.49. In that case, based on Equation (1), the validity coefficient would be 0.23. If we use the more realistic value of 0.30 as the true correlation, the validity coefficient would reduce to 0.14. And further realism would reduce the validity coefficient yet further. Finally, if anyone believes that department head evaluations are more reliable than I give them credit for being, I can only ask you to consider the different department heads you know, their relations with various faculty, and reevaluate your belief.

4 Discussion

Validity of assessing teaching effectiveness is crucial, but we have seen that the technical issue of reliability is intimately connected with that validity. Put simply, reliability is a prerequisite for validity. In terms of Aristotle's necessary and sufficient causes, reliability is a necessary, but not sufficient, cause of validity. Although Figure 1 illustrates nicely that reliability is a necessary cause of validity, that reliability is not a sufficient cause of validity might require more explanation. To see this quickly, imagine a measure of teaching effectiveness with the following item: "How many inches tall are you?" This item doubtless would achieve extremely good test-retest reliability as people tend not to change their height from day to day, or even week to week. But despite the excellent reliability, teacher height would nevertheless be a poor predictor of student learning.

Given that reliability is necessary for validity, but not sufficient, let us address the twin issues of obtaining reliability and validity. With respect to reliability, although there are many ways to index it, the most widely used reliability index is Cronbach's (1951) [32] alpha, expressed below as Equation (3) (This is not to say that Cronbach's alpha is the best reliability index, because it is not. However, better methods are not widely used, possibly because they are more complex, and it is sufficient for present purposes to remain with Cronbach's alpha):

$$\text{reliability as indexed by } \alpha_{\text{standardized}} = \frac{K\bar{r}}{1+(K-1)\bar{r}} \tag{3}$$

where
(1) $\bar{r}$ = the average inter unit correlation;
(2) K is the number of units.

Equation (3) clarifies that there are two ways to improve reliability. Increase the extent to which the items making up one's assessment device correlate with each other and increase the number of units (items). Figure 2 shows how increasing the average interitem correlation can influence reliability assuming 2 items, 4 items, or 8 items. Figure 2 illustrates that more items imply more reliability as do better interitem correlations. Finally, the effect of interitem correlations interacts with the number of items, as the different curves show. Thus, a researcher who wishes to have even a chance at acceptable validity, needs to have good reliability first, as Figure 1 indicates. In turn, to have good reliability, the items need to correlate well, there needs to be many items, or both. Ignoring these technical details is tantamount to guaranteeing to tank the validity of one's assessment of teacher effectiveness, regardless of whether it is based on student evaluations of teaching, subjective opinions of department heads, evaluations of people who observe teaching, an exposition of teaching techniques used in the course, and so on. The first question to ask is, "Why should I believe that this method is reliable?" Or, if one is attempting to construct an assessment device, the question is, "How can I make my assessment device more reliable?" The answer to the latter question is, (a) increase the similarity of the items to each other so interitem correlations increase or (b) have more items.

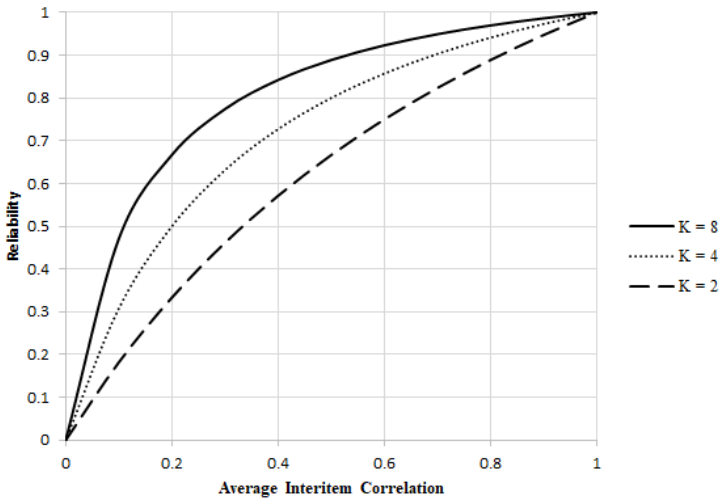

Figure 2 Reliability ranges along the vertical axis as a function of the average interitem correlation along the horizontal axis, when there are 8 items (top curve), 4 items (middle curve), or 2 items (bottom curve).

A potential counterargument to the notion of increasing the similarity of the items is that the assessment device might not capture all that is meant by teaching effectiveness. But the counterargument is flawed because if very disparate items are needed to cover "all of the construct," then that implies one has multiple constructs that ought to be measured separately. For example, if one believes that good lectures constitute a component of teaching effectiveness and that effective use of technology does too, combining them into a single measure of teaching effectiveness will almost guarantee poor reliability due to the poor interitem correlations that can be expected. In turn, the poor reliability will guarantee even poorer validity. Instead, it would be better to construct separate, and reliable, measures of lecturing and use of technology. In addition to the reliability issue, another potential gain of separating the variables is that their relative contributions to student learning could be investigated using multiple regression or structural equation modelling procedures.

Moving to validity, the first validity question to ask is: "Valid for what?" If the goal is validity at predicting student learning, then we have to agree on that which constitutes student learning. This might include student regurgitation of material, ability to apply class material to solve problems, ability to connect class material with material from previous classes, ability to think outside-the-box, or many other possibilities. And there is no reason to believe that these different types of student learning will correlate well with each other. Thus, in a previous example, where we assumed that the reliability of the hypothetical measure of student learning is 0.70, this likely was wildly optimistic. We might agree that many different kinds of learning are important, and that is fine, but amalgamating them into a single score is guaranteed to result in low reliability by Equation (3). It would be better to keep them separate and have reliable measures of each type of learning.

Well, then, if we have separate measures of different kinds of teaching effectiveness, and separate measures of different kinds of student learning, there would be many different validity coefficients. For instance, imagine we identified three types of teaching effectiveness (*A*, *B*, and *C*) and three types of student learning (*L*, *M*, and *N*), in which case there would be nine validity coefficients to estimate: ρ_{AL}, ρ_{AM}, ρ_{AN}, ρ_{BL}, ρ_{BM}, ρ_{BN}, ρ_{CL}, ρ_{CM}, and ρ_{CN}. Likely, some types of teaching effectiveness would correlate more with some types of student learning whereas different types of teaching effectiveness would correlate more with other types of student learning. And with reliable measures of different types of teaching effectiveness and reliable measures of different types of student learning, those interested in the scholarship of teaching would be able to determine convincingly where the validity coefficients are impressive and where they are not.

This thinking suggests difficult issues. For example, suppose that ρ_{AL} is an impressive number but ρ_{AM} is not, whereas ρ_{BL} is not impressive but ρ_{BM} is. The implication would be that teaching effectiveness with respect to *A* matters more than teaching effectiveness with respect to *B*, as long as we are concerned with student learning of type *L*. But if we switch to a concern with student learning of type *M*, then the reverse is true. Thinking in this way opens up a potential debate not only about which types of student learning matter most, but also about how to measure those types of student learning reliably and validly. More than that, there would have to be a recognition that particular types of teaching effectiveness might be differentially effective for different types of learning. This last point should be commonsensical, but I have yet to see it recognized in my own university, or when I have been asked by other universities to evaluate their candidates for tenure, promotion, or both.

It is interesting to compare universities to sports teams. A basketball coach, for example, would not settle for an overall evaluation of basketball ability, but would categorize each of his players with respect to shooting, passing, dribbling, rebounding, team defense, individual defense, and others. And further, the coach would differentially apply the importance of categories depending on player and position. A power forward might be expected to rebound more effectively whereas a point guard might be expected to be more effective at dribbling and passing. And practice sessions would reflect that there would be different priorities for different players, different positions, and so on. In contrast, the tendency is for teachers to be assessed with a single number. Given the diversity of classes, goals for student learning, and other differences, this is silly on the face of it; just as it would be silly for a basketball coach to use a single number to assess each player on the team. The silliness increases when technical issues pertaining to (a) the relationship between reliability and validity and (b) what is required for reliability are considered. The psychometric facts of life dictate the importance of reliability for valid assessment of teaching. In turn, the psychometric facts of life also dictate that amalgamation across disparate types of items is tantamount to guaranteeing a lack of reliability, and consequently a lack of validity too.

In conclusion, we commenced with a contradiction between the theoretical desirability of holding teachers accountable and Strobe's (2020) [1] comprehensive review showing that practically, it does not work. The present essay uncovers an underlying cause of the contradiction. There are crucial psychometric facts that pertain to the assessment of teaching effectiveness, and unless these are considered, assessment of teaching will remain blatantly invalid. This is not to say that nobody cares about psychometric facts. On the contrary, there are journals, such as *Educational and Psychological Measurement*, that are replete with technical articles pertaining to reliability, validity, statistical, and mathematical issues. In addition, the reason many consider these issues old-fashioned is because they are old-fashioned, dating back at least to Spearman's seminal work in 1904. Unfortunately, those who handle teaching assessment issues at universities have little clue about the foundational psychometric prerequisites for validity that they deem old-fashioned, dry, and boring, if there is any awareness whatsoever. I know many who have talked a very impressive teaching assessment game, but none of them have a satisfactory knowledge of the technical issues that are a psychometric necessity for

valid teaching assessment. Until those who typically assess teaching are force-fed the old-fashioned, dry, and boring psychometric facts of life, teaching assessment will continue to be invalid. Perforce, the contradiction between the desirability to hold teachers accountable and the practical problem of doing so validly, will continue.

References

[1] Stroebe W . Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis. Basic and Applied Social Psychology, 2020, **42**(4): 276-294.
https://doi.org/10.1080/01973533.2020.1756817

[2] Hanushek EA and Steven GR. Generalzatons about usng value-added measures of teacher qualty. American Economic Review, 2010, **100**(2): 267-271.
https://doi.org/10.1257/aer.100.2.267

[3] Crocker L and Algina J. Introduction to classical & modern test theory. United States: Wadsworth, 1986.

[4] Rosenthal R and Rosnow RL. Essentials of behavioral research: Methods and data analysis. Boston, M. A.: McGraw-Hill, 2008.

[5] McLeod SA. What is reliability?. Simply Psychology, 2007.
https://www.simplypsychology.org/reliability.html

[6] Spearman C. The proof and measurement of association between two things. International Journal of Epidemiology, 2010, **39**(5): 1137-1150.
https://doi.org/10.1093/ije/dyq191

[7] Gulliksen H. Theory of mental tests. Hillsdale, NJ: Lawrence Erlbaum Associates, 1987.

[8] Lord FM and Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

[9] Abrami PC, d'Apollonia S and Cohen PA. Validity of student ratings of instruction: What we know and what we do not. Journal of Educational Psychology, 1990, **82**(2): 219-231.
https://doi.org/10.1037/0022-0663.82.2.219

[10] Clayson DE. Student evaluations of teaching: Are they related to what students learn?. Journal of Marketing Education, 2009, **31**(1): 16-30.
https://doi.org/10.1177/0273475308324086

[11] Clayson DE, Frost TF and Sheffet MJ. Grades and the student evaluation of instruction: A test of the reciprocity effect. Academy of Management Learning & Education, 2006, **5**(1): 52-65.
https://doi.org/10.5465/amle.2006.20388384

[12] Uttl B, White CA and Gonzalez DW. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching rating and student learning are not related. Studies in Educational Evaluation, 2017, **54**: 22-42.
https://doi.org/10.1016/j.stueduc.2016.08.007

[13] Boehmer DM and Wood WC. Student vs. faculty perspectives on quality instruction: Gender bias, "hotness", and "easiness" in evaluating teaching. Journal of Education for Business, 2017, **92**(4): 173-178.
https://doi.org/10.1080/08832323.2017.1313189

[14] Felton J, Koper PT, Mitchell J, *et al*. Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofesssors.com. Assessment & Evaluation in Higher Education, 2008, **33**(1): 45-61.
https://doi.org/10.1080/02602930601122803

[15] Fisher AN, Stinson DA and Kalajdzic A. Unpacking backlash: Individual and contextual moderators of bias against female professors. Basic and Applied Social Psychology, 2019, **41**(5): 305-325.
https://doi.org/10.1080/01973533.2019.1652178

[16] Freng S and Webber D. Turning up the heat on online teaching evaluations: Does "hotness" matter. Teaching of Psychology, 2009, **36**(3): 189-193.
https://doi.org/10.1080/00986280902959739

[17] Hamermesh DS and Parker A. Beauty in the classroom: instructors pulchritude and putative pedagogical productivity. Economics of Education Review, 2005, **24**(4): 369-376.
https://doi.org/10.1016/j.econedurev.2004.07.013

[18] Johnson RR and Crews AD. My professor is hot! Correlates of RateMyProfessors.com ratings for criminal justice and criminology faculty members. American Journal of Criminal Justice, 2013, **38**(4): 639-656.
https://doi.org/10.1007/s12103-012-9186-y

[19] Riniolo TK, Johnson k, Sherman T, *et al*. Hot or not: Do professors perceived as physically attractive receive higher student evaluations. The Journal of General Psychology, 2006, **133**(1): 19-35.
https://doi.org/10.3200/GENP.133.1.19-35

[20] Rosen AS. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. Assessment & Evaluation in Higher Education, 2018, **43**(1): 31-44.
https://doi.org/10.1080/02602938.2016.1276155

[21] Wolbring T and Riordan P. How beauty works. Theoretical mechanism and two empirical applications on students evaluation of teaching. Social Science Research, 2016, **57**: 253-273. https://doi.org/10.1016/j.ssresearch.2015.12.009

[22] Gurung RAR and Vespia KM. Looking good, teaching well? Linking liking, looks, and learning. Teaching of Psychology, 2007, **34**(1): 5-10. https://doi.org/10.1177/009862830703400102

[23] Delucchi M. Don't worry. Be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. Teaching Sociology, 2000, **28**(3): 220-231. https://doi.org/10.2307/1318991

[24] Feistauer D and Richter T. Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. Studies in Educational Evaluation, 2018, **59**: 168-178. https://doi.org/10.1016/j.stueduc.2018.07.009

[25] Reysen S. Construction of a new scale: The Reysen likability scale. Social Behavior and Personality: An International Journal, 2005, **23**: 2001-2208. https://doi.org/10.1037/t66417-000

[26] McPherson MA and Jewell R. Leveling the playing field: Should student evaluation scores be adjusted? Social Science Quarterly, 2007, **88**(3): 868-881. https://doi.org/10.1111/j.1540-6237.2007.00487.x

[27] Reid LD. The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors. Journal of Diversity in Higher Education, 2010, **3**(3): 137-153. https://doi.org/10.1037/a0019865

[28] Smith BP. Student ratings of teacher effectiveness: An analysis of end-of-curse faculty evaluation. College Student Journal, 2007, **41**: 788-800.

[29] Arceo-Gomez EO and Campos-Vazquez RM. Gender stereotypes: The case of MisProfesores.com in Mexico. Economics of Education Review, 2019, **72**: 55-65. https://doi.org/10.1016/j.econedurev.2019.05.007

[30] Boring A. Gender biases in student evaluations of teaching. Journal of Public Economics, 2017, **145**: 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

[31] Mengel F, Sauermann J and Zlitz U. Gender bias in teaching evaluation. Journal of the European Economic Association, 2019, **17**(2): 535-566. https://doi.org/10.1093/jeea/jvx057

[32] Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, **16**(3): 297-334. https://doi.org/10.1007/BF02310555