2020

Do Differences in Teaching Evaluations Really Matter? An Investigation into What Constitutes a Meaningful Difference in Evaluations of Professors

Catherine Bain
Butler University

Do Differences in Teaching Evaluations Really Matter?

An Investigation into What Constitutes a Meaningful Difference in Evaluations of Professors

A Thesis

Presented to the Department of Psychology

College of Liberal Arts and Sciences

of

Butler University

In Partial Fulfillment

of the Requirements for Graduation Honors

Catherine Marie Bain

May 6th 2020

Abstract: This study sought to determine what constitutes a minimally meaningful difference in student evaluations of their professors, when students are asked to rate their professors on the traditional 5-point teaching effectiveness item commonly used in higher education. A minimally meaningful difference is the smallest difference between two ratings that: 1) exceeds chance variation and 2) corresponds to a difference deemed meaningful using some external anchor or standard. Data was obtained through a series of surveys given to students at Butler University and to an online nationwide sample. Analysis occurred through both an anchor-based approach, using data obtained from a single survey, and a distribution-based method, using data obtained from two surveys administered two weeks apart. Both methods were used to find the minimal meaningful difference in student evaluations of professors. A meaningful difference of .84 was found when participants were asked to distinguish between two professors of higher quality. A meaningful difference of .75 was found when participants were asked to distinguish between two professors of lower quality. Both differences exceeded chance variation.

Do Differences in Teaching Evaluations Really Matter?

An Investigation into What Constitutes a Meaningful Difference in Evaluations of Professors

Introduction

At the end of each semester, students anxiously await their teacher evaluation email asking them to give feedback on their professors from that semester. Some students love this time of year when they are given the chance to praise their favorite professor and give constructive feedback to the professors they may not like quite so much. Yet, some students loathe this process, for it often seems as though those evaluations have little to no impact on which professors are brought back the following year or what classes a professor may teach. This leaves students asking themselves if their opinion even matters, or rather, how different their ratings need to be about one professor as compared to the next for them to have an effect. This study seeks to answer that question, or more specifically, how different ratings should need to be to warrant change. Universities use those student evaluations when considering which instructors they will bring back for the following academic year, who they will recommend for a tenure track position, and other critical decisions affecting faculty. Should a professor who is given a 5.00/5.00 rating be given a raise over someone who received at 4.75/5.00? Is that difference truly meaningful? My thesis intends to answer this question: In the context of student evaluations, how large does the difference in ratings need to be to constitute a meaningful difference?

The medical field faces many of the same issues when evaluating different drug or therapy treatments. Researchers often look at whether treatment effects are statistically significant, meaning the difference seen with the treatment is not simply due to chance. Yet, just because a new drug or therapy produces a statistically significant change or difference does not mean it is necessarily successful. If the change is not large enough to be meaningful (or even noticable) to patients, the

therapy is arguably of little value. So, how much of a change does the drug or therapy need to make in the patients life to truly be considered successful? In an attempt to answer this question, researchers have developed what are known as minimally meaningful differences. This refers to the amount of change that is meaningful to the patient. For example, if an arthritis drug is prescribed, patients receiving the drug may experience a change that is statistically significant, but the change may be so small that patients do not feel as though they have better range of motion without pain, and thus the change is not meaningful to them. Therefore, this drug would not produce a meaningful difference. For example, Motl et al. (2014) performed a study containing patients with Multiple Sclerosis. They found that although a statistically significant difference occurred at a 6-point change in scores on the Multiple Sclerosis Walking Scale (MSWS)-12, patients did not find that there was any meaningful change on their quality of life (Motl et al., 2014). In response to this dilemma, medical researchers have developed techniques to determine what constitutes a meaningfully significant (as opposed to statitistically significant) difference. Over the years, as research addressing this type of change has grown in popularity, various terms have developed to refer to the same concept, including the following: meaningful difference, clinically significant difference, clinically important changes, and just noticable difference.

This study is meant to apply these principles, originated primarily in medical research, to the higher education system in an attempt to provide a better understanding of meaningful differences in student evaluations of teachers. I will be investigating the minimally meaningful differences in teacher evaluations at Butler University with hopes to expand the study to other universities across Indiana and perhaps America.

Review of the Literature

Identifying a minimally meaningful difference is a relatively new concept, and as such, there is no true preferred method for doing so. Previous research, however, seems to be converging on two standardized methods: the anchor based approach and the distribution based approach. The first, the anchor based approach, relies on an external, previously established, significant criterion (the anchor) against which changes in the new data can be compared. This anchor may range from a patient's self-reports of change to a clinical outcome or anything in between. There are also various ways in which an anchor can be obtained. For example, a transition assessment item could be used where a subject is asked to complete a series self-reports where they determine how much, if any, change they have experienced (e.g., have you experienced no change, a small amount, a moderate amount, a large amount, etc.).

Review of Anchor Based Studies

In 2003, Zisapel and Nir performed a study to determine the minimal clinically significant difference of patient sleep quality on the Leeds Sleep Evaluation Questionnaire (LSEQ) through an anchor-based approach using the Visual Analog Scale (VAS) as the anchor. They had a sample of 428 patients, all age 55 or above, who were all diagnosed with insomnia. They gave all participants a placebo for the first two weeks of the study followed by three weeks of a prolonged release of 2mg of melatonin and a placebo for two more weeks. They made sure to assess sleep quality using the LESQ after each time period, which used a five-point severity rating scale (Zisapel & Taliner, 2003). They found that a one-point change on the LESQ was associated with an average change of 10.3 mm on the VAS. The VAS is often used by those involved in sleep research to measure aspects of sleep and daytime functioning and the effects therapeutic interventions have on them. Past research suggests that the minimally meaningful change as seen by patients occurs at 10 mm on the VAS, suggesting that a one-point change on the LESQ is a meaningful difference.

Salaffi et al. (2003) also chose to examine minimal clinically important differences (MCIDs) using a sample of 825 patients with chronic musculoskeletal pain. They used the patient's global impression of change (PGIC) questionnaire as an anchor to determine the MCID for a numerical rating scale (NRS). Salaffi et al. performed a prospective cohort study assessing patient's pain intensity using the NRS twice, the first being a baseline and the second a follow up three months later. In addition, they used the PGIC questionnaire to assess which patients believed they had experienced a noticeable improvement (defined as being "much better"). The results indicated that patients reporting a noticable improvement moved about 2 points on the NRS, suggesting a 2-point difference on the NRS should be considered "much better" (i.e., meaningful). These results are consistent with previously published findings that the use of a "much better" improvement rating is a useful tool to find MCID in medical cases.

Miller and Manuel (2008) also conducted research addressing meaningful differences; however, they looked at whether or not the difference was meaningful to practitioners. To do this, they used an estimation method for which they surveyed fifty substance abuse treatment providers that were all involved in the National Institute on Drug Abuse (NIDA) clinical trials. Miller and Manuel had practitioners identify thresholds for clinically meaningful differences by indicating the size of the effect that would justify their learning a new treatment method (2008). In this case, the anchor used comes from the previously established outcomes percentages, and the meaningful difference is determined by the practitioners' self-reports. Their research found that the practitioners felt there was a meaningful difference between the new treatment and the current treatment if the outcomes percentages improved at least 10 points (Miller & Manuel, 2008). Although all of these differences were found to be non-statistically significant in a least some prior research, they appear to be clear and meaningful differences from the perspective of practitioners.

Review of Distribution Based Studies

The second approach used to determine minimally meaningful differences is that of the distribution-based approach, which relies on the statistical properties of patient reported outcome data. The most valuable characteristic of this approach is that it allows for the identification of differences that are likely too large to have occurred by chance or from random measurement error. There are a variety of ways by which to execute a distribution-based approach. One of the most commonly used is one that relies on the effect size associated with a given difference or change. A common effect size used is the standard mean difference (SMD). To calculate the SMD one divides the difference between two sample means by the pooled standard deviation. This is commonly then interpreted through Cohen's guidelines, which deem that the effect sizes of .20 to .49, .50 to .79, and .80 and above are considered small, medium, and large, respectively (Cohen, 1988). Thus, according to this interpretation, any difference associated with an effect size of .20 or greater would possess one of the necessary properties of being meaningful. However, this approach does assume that relatively small differences cannot be meaningful, which is a potential limitation.

Raman et al. (2016) examined clinically important differences as they pertain to quality of life of bone metastases patients who are undergoing palliative radiotherapy. This study used both an anchor-based and a distribution-based method to determine what Raman termed the minimally clinically important difference (MCID), which is conceptually equivalent to a minimally meaningful difference, the term we employ in the current thesis. The quality of life data were collected from a sample of patients who were involved in a randomized phase III trial (Raman et al., 2016). Patients were classified as improved, stable, or deteriorated, which was used as the anchor to characterize changes on a multi-item quality of life measure after a 42 day follow up. The distribution-based approach was also used to determine the smallest amount the MCID could be. To do this, the

investigators calculated the standard error of measurement (SEM). The SEM is the standard deviation of an individual's scores on a specific measure and can be calculated by multiplying the sample standard deviation by the square root of one minus the reliability of the measure. The SEM captures how much a reported score is likely to differ on average from the corresponding true score due to random factors, such as measurement error (Harvill, 1991; Horn, 1971). Thus, the SEM can be used to provide an estimate of how large the difference between two scores must be to be relatively confident that the difference is 'real' and not the result of chance or error (Wyrwich, et al., 2005; Giesler, 2011). In other words, for a difference to be deemed meaningful, a necessary but not sufficient criterion should be that the difference exceeds the SEM. How much greater than the SEM a meaningful difference should be is somewhat controversial, although at least some authors suggest one SEM (see Giesler, 2011). After calculating the SEM, the investigators found that the MCID's determined by the anchor and distribution approaches 'agreed' for patients who were improving (i.e., the amount of change deemed meaningful according to the anchor approach exceeded what could be expected from chance variation). However, for declining patients, the two approaches did not always 'agree', with the anchor-based approach producing some MCID's that did not exceed the SEM, specifically the MCID's for emotional pain, and constipation. Several of the MCID's found in this study were not large enough to be statistically significant, but were still meaningful to the patients.

Most research has indicated that using a combination of the anchor-based and distribution-based approaches will provide the best picture of a meaningful difference. As previously mentioned, using a distribution-based approach ensures that the difference will exceed that which could simply be expected by chance while the anchor-based approach ensures that the difference is truly meaningful. By combining the two methods, researchers can ensure that the difference they find will be both meaningful to the patient and larger than that which could be the result of chance.

Methodology

Both a distribution-based approach and an anchor-based approach were used to investigate minimal meaningful differences in the context of teacher evaluations using multiple surveys.

Distribution-Based Approach

Participants.

Participants (n=65) were recruited and surveyed through the SONA system, which is the subject pool management software utilized by the Department of Psychology at Butler University. Participants received extra credit for participation. The sample was 80.5% female.

Procedure.

Two online surveys were given to the same participants near the middle of the semester, separated by a gap of two weeks. Participants answered the survey only after giving informed consent. Each survey asked the participant to evaluate one of their current professors on a set of items using five point, Likert-type response scales. The items were taken from the teaching evaluation survey originally used by Butler University's College of Liberal Arts and Sciences during the prior decade. The critical item, which was included in the set of standard evaluation items, asked respondents to simply rate their instructor's overall teaching effectiveness on a 5-point scale anchored by poor and excellent. This item was chosen because of its near universal presence on college instructor teaching evaluations. Participants were then asked to re-rate their professor using decimals in their ratings, which was encouraged by the use of a "sliding indicator". The item used to obtain these ratings was: "Now, if you could rerate this professor on overall teaching effectiveness using decimals, which will allow an even more precise rating, how would you rate them?" The same participants were then asked to complete the same set of items again two weeks later. This gap in time was used to minimize potential carryover effects, although this procedure cannot fully eliminate

the possibility that some participants may have relied on recall of time one ratings when completing time two ratings. Administering the full set of items at both time points was also used to reduce carryover effects and to better reflect the conditions under which college students typically provide teaching evaluations (i.e., students usually answer multiple questions about their instructor). The data from the critical item was used to calculate the item's test-retest reliability, which in turn can be used to compute the standardized error of measurement (SEM). The SEM provides an estimate of how great a difference in the ratings would have to occur in order for the difference to exceed chance variation. The SEM can be calculated by multiplying the sample standard deviation by the square root of one minus the reliability of the measure to help determine the size of the difference needed to be confident that the difference exceeds chance variation (Giesler, 2011; Harvill, 1991). We describe the results of the Distribution-Based Approach with the results from the Anchor-Based Approach in the general Results section below, because the findings of each study so closely inform the other.

Anchor-Based Approach

Participants.

In order to increase the generalizability of the results, a nationwide sample of 258 participants was obtained through a survey on CloudResearch, which was possible due to a Butler University thesis grant. CloudResearch is a service provided by Amazon that social scientists use to recruit participants for online studies. CloudResearch participants are able to view available tasks and the associated pay rate and then decide which tasks they want to do. The samples provided by CloudResearch, previously known as MTurk, have been studied extensively to determine their characteristics. These studies have generally found that MTurk/CloudResearch samples produce results similar in quality to those obtained in a more traditional manner, but that appropriate methods

should be utilized to screen out respondents who provide low quality data (Buhrmester et al., 2011; Chmielewski & Kucker, 2020).

Of the 258 participants, 135 were male, 120 were female, and 3 reported their gender as "other". Ages ranged from 20-70 years old with a mean age of 36.1 (SD=10.326). About 73% of participants were white, 6.2% were Black or African American, 3.1% were American Indian or Alaskan Natives, 10.1% were Asian, 6.6% were Hispanic/LatinX, and .8% were of another race. About 30% of respondents did not have a college degree but had taken or were taking college coursework, 54% had a bachelor's degree, 14% had a master's degree, and 2.3% had a professional or doctorate degree.

Procedure

An online survey advertised on CloudResearch was used to collect participants' responses to a series of questions about participants' previous or current professors. Participants answered the survey only after giving informed consent. Throughout the entire process, the professors were never named, thus keeping their identity anonymous. This survey took no more than 15 minutes to complete and contained questions asking participants to rate professors in several ways, including using the five-point scale anchored by poor and excellent. These data were then used to identify, and partially validate, a meaningful difference on the critical five-point poor to excellent scale.

The primary approach used to identify a meaningful difference consisted of asking respondents to think of the best professor they have had and rate that professor on the five-point scale, then to think of a professor who was also good, but noticeably less good than their best professor. That second professor was then rated on the same scale (Question 1, Appendix A)

Participants were allowed to use decimals in their ratings, which was encouraged by the "sliding indicator" they were asked to use to indicate their ratings. The difference between the best

professor's score and second-best professor's score was computed for each participant and then averaged across participants. This approach allowed for the discovery of a meaningful difference when evaluating student ratings of 'good' instructors (i.e., instructors whose ratings fall on the upper end of the scale). Because individuals often rate positive and negative stimuli differently (Baumeister et al., 2001), a similar approach was used to determine if the same meaningful difference would emerge on the negative end of the scale. Specifically, respondents were asked to think of the worst professor they have had and rate that professor on the five-point scale, then to think of a professor who was also of lower quality, but noticeably better than their worst professor. That second professor was then rated on the same scale, and the difference between worst and second worst was calculated (Question 2, Appendix A).

Then, in an attempt to acquire evidence for convergent validity, scores from two other "positive" questions and two more "negative" questions were examined. Positive questions asked respondents to rate professors of higher quality, whose ratings generally fell on the positive end of the 5-point scale; negative questions asked respondents to rate professors of lower quality, whose ratings generally fell on the negative end of the 5-point scale. One of the positive items asked students to think of three of the best professors they have had. We asked them to rank them in order of quality (i.e. best, second best, third best). We then asked them if they thought the difference in quality between each pair of professors was meaningful. Finally, we had them rate each professor (Question 3, Appendix A). The second positive item aimed at obtaining evidence of converging validity asked participants to think of two of the better professors they have had, rank them, and rate each professor on a 5-point scale using decimals. They were then told to imagine that the better professor would be getting a $1000 raise and were asked to indicate if the difference between professors was large enough to justify giving the better one a raise. If participants did not find the

difference meaningful, they were asked to keep thinking of new lesser quality professors until the difference was meaningful and then rate that lesser quality professor (Question 5, Appendix A). However, due to time constraints, we did not examine the data from the second part of this question. We only examined the difference between scores provided by participants who said the initial difference was meaningful, as detailed in the results.

One of the negative items used to provide convergent validity evidence asked students to think of the three worst professors they have had and rank them, indicate whether the difference between each pair was meaningful and then rate each professor on a 5-point scale using decimals (Question 4, Appendix A). The second negative item asked participants to think of two of the worst professors they have had and rank and rate them on a 5-points scale using decimals. Then participants were asked to imagine that they were the dean of a college who had to fire one of the two professors as the result of budget cuts. We asked participants if they felt the difference between professors was meaningful enough to warrant firing the worst professor. If they said no, we asked them to keep thinking of new, but still low-quality, professors until they found one that was sufficiently better than the worst professor such that the firing of the worst professor could now be justified. As with the parallel positive item, due to time constraints, we only focused on the responses of participants who initially said the difference was large enough to be meaningful. (Question 6, Appendix A).

Results

Distribution-Based Approach

Before computing the SEM based on the test-retest reliability information obtained from the Distribution data set, outliers were first removed from the data set. Participants' time one ratings of their professor on the critical item were subtracted from their time two ratings. Theoretically, this

difference should be near zero as the same professor is being evaluated only two weeks apart. Outliers identified by difference values exceeding two standard deviations were removed from the data set. This left us with a sample of 61 participants. From this cleaned data set, the standardized error of measurement (SEM) was calculated. The SEM value was calculated by multiplying the average of the time one and time two standard deviations by the square root of one minus the test-retest correlation, a measure of reliability. This data produced an SEM value of .3349[1]. In keeping with prior work (see Giesler, 2011), any differences found using the anchor-based approach were deemed meaningful if the difference found was larger than the SEM value of .3349.

Anchor-Based Approach

Finding a meaningful difference was approached in two different ways, as described earlier. For the "positive" approach, respondents were asked to compare the best professors they have had to identify the size of meaningful differences occurring on the upper or positive end of the scale; for the "negative" approach, participants were asked to compare the worst professors they have had. Before analyzing the data, we screened for participants whose responses suggested they were not paying attention to or properly comprehending the questions, a practice commonly recommended when using samples recruited through online methods. We removed any participant who provided ratings of professors that were not consistent with their rankings of professors from questions one, two, three, and four, which most directly solicited rating and ranking responses. For example, if the participant was asked to compare their best ranked professor to their second-best ranked professor, their responses would not be included if they rated the second-best professor higher than the best

[1] This value was calculated using the decimal ratings. One could argue that the test-retest correlation used to compute the SEM should be determined from the ratings typically produced by students, which are usually collected on discrete 1 to 5 scales. However, the test-retest correlation when discrete ratings were used was 1.00, resulting in an SEM of 0. Because of this, we chose to compute the SEM using ratings that allowed decimals to ensure that the meaningful difference determined by the anchor-based approach truly exceeded chance variation. Thus, the approach we chose is arguably overly conservative.

professor. This produced a sample size of n=163, resulting in the exclusion of 95 participants. Although a sizeable proportion of the sample was lost through these procedures, recent investigations into the quality of online samples suggest our sample was not atypical in this respect (Chmielewski & Kucker, 2020).

Positive Approach.

As a reminder, questions one, three and five approached the difference from a positive standpoint. Descriptive statistics for each of these questions can be found in Table 1.

Question one asked participants to simply rate their best and noticeably second-best professors on a 5-point scale and was used to establish the meaningful difference. This produced a mean difference in question one of .8402 with a standard deviation of .4667. This mean difference is larger than the SEM of .3349 found through the distribution-based approach. This suggests that a mean difference of .8402 is large enough to exceed chance variation, which in turn suggests that one can be relatively confident that a difference of approximately .84 on a five-point scale is meaningful. The standard deviation is larger than we had hoped, however, and does call into question the precision of the meaningful difference.

To establish convergent validity, question three asked participants to rank then rate three of their best professors on the same scale and to indicate which pairs of professors differed meaningfully in terms of overall quality. We only examined ratings from participants who reported that the difference between the best and second-best professor was meaningful, giving us a sample size of 112. Question three produced a mean difference of .6891 between best and second best with a standard deviation of .3224. We correlated the difference identified in question one with the difference identified in question three, which as just described, only included participants who indicated that the difference between the best professor and second-best professor was meaningful.

This left us with 112 participants. When the differences from question one were correlated with the differences from question three, a significant moderate correlation was found ($r(110)=.499$, $p<.001$).

As a further check on the validity of the difference identified in question one, we analyzed the ratings elicited by question five. Question five asked participants to rate the two best professors they have had, and then asked if the difference in their rating was meaningful enough to warrant giving a one-thousand dollar raise to the professor with the higher rating. We only considered scores from those who said it was. This left us with 120 participants. The mean difference produced by this approach was .5414 with a standard deviation of .3399. A low, marginally significant correlation ($r(118)=.170$, $p=.064$) was found between the differences identified by question one and question five. When examining the scatterplot, it appeared that the lower than expected correlation was the result of two outliers. We retained the outliers, however, as we had not made a priori plans to screen for outliers for this particular analysis.

Negative Approach.

Questions two, four, and six approached the difference from a negative standpoint. We used the same analytic strategy for these questions as described above. Descriptive statistics for each of these questions can be found in Table 2.

Question two, which asked participants to simply rate their worst professor then another lower quality professor who was noticeably better than their worst professor, was used to establish the meaningful difference. This produced a mean difference of .7505 with a standard deviation of .4409 (n=163). This mean difference is larger than the SEM from the distribution-based approach (i.e., .3349), indicating the difference exceeds chance variation. Information from the other questions was then used to assess convergent validity. To determine whether the negative difference differed statistically from the positive difference, a within samples t-test was run comparing the difference

score from question one to the difference score from question two. This test indicated that there was a significant difference between the positive difference and negative difference, t(162)=29.383, p<.001.

Question four asked participants to rank then rate three of their worst professors on the same scale and to indicate which pairs of professors differed meaningfully in terms of overall quality. We were primarily interested in the ratings from those who said the difference between their worst and second-worst professor was meaningful, which theoretically should be similar to the difference identified in question two. Eliminating those who said it was not meaningful gave us a sample size of 93. The difference between the worst and second worst professor from question four produced a mean difference of .7319 with a standard deviation of .4201. When the differences from question two were correlated with the differences from question four, a significant, moderate to strong correlation was found ($r(91)=.658$, $p<.001$). Question six asked participants to rate the two worst professors they have had, and then asked if the difference was meaningful enough to fire the worst professor of the two. We only considered scores from those who said it was (n=110). The mean difference produced by this method was .6763 with a standard deviation of .4579. A significant moderate to large correlation was found ($r(110)=.693$, $p<.001$) between the question six differences and the question two differences.

Discussion

The goal of this study was to establish a minimally meaningful difference for students' evaluations of professors on the traditional 5-point scale. Because the minimally meaningful difference is something typically used in a medical setting, it has yet to receive much attention from other fields. We first employed a distribution-based approach to determine the standard error of measurement (SEM) associated with students' ratings, which can be used as an indicator of how

large a difference would have to be in order for it to exceed chance variation using data from a sample of Butler students. We then used an anchor-based approach to determine how large of a difference would be considered meaningful by students. As indicated earlier, a thesis grant provided the opportunity to acquire a more diverse sample through CloudResearch for this anchor-based approach. The anchor-based values were calculated using this data sample and were compared to the SEM found through the distribution-based approach to determine if the difference found exceeded what could be expected from chance variation.

The minimally meaningful difference for professors on the positive end of the scale was .84; the minimally meaningful difference for professors on the negative end of the scale was .75. These differences were then correlated with differences elicited by conceptually similar but alternative methods to assess for convergent validity. The correlations were mostly significant and relatively large, supporting the meaningful differences' validity.

The findings from our two different anchor-based approaches (i.e. positive and negative) were surprisingly close, but produced two different values. Although the difference between the values was found to be statistically significant, the fact that they differed by less than .10 of a point on a five-point scale suggests that the meaningful difference between two professors is relatively stable across the scale. Additional replications with more diverse and larger samples will be needed to confirm the stability of the difference, as well as studies that focus on differences occurring near the midpoint of the scale, but our findings suggest that, roughly speaking, a .80 difference can generally be considered meaningful. Our confidence in this value would be heightened if the associated standard deviation had been somewhat lower, but replications with larger sample sizes will likely result in smaller standard deviations.

However, with respect to our two minimally meaningful difference values, it is interesting to note why they differed. Despite their similarity, a larger meaningful difference was found when asking participants about good professors, while a smaller meaningful difference was found when asking participants about bad professors. This difference between good and bad was seen again from the differences generated by questions five and six, which imposed the conditions of giving a raise to a good professor or firing a bad professor.

One possible reason for this is that we struggle more to determine a difference between things we feel positively about compared to things we feel negatively about. That is to say, perhaps it is harder for us to distinguish between people we like than it is for us to distinguish between people we dislike. If this is true, it could explain why the minimally meaningful difference must be larger for professors who students rate positively than it is for those they rate negatively, because students probably like professors they rate positively more than those they rate negatively. In addition, it is possible that it is harder for students to justify rewarding a professor than it is to justify firing a professor (i.e., they find it easier to fire a professor than to give a professor a raise because bad teaching is easier to recognize than good teaching). This would be an interesting point for future research to address.

Social styles, which refers to how a person generally interacts with others in interpersonal contexts, of students and professors may also influence the student's opinion of the professor. Previous research has indicated that a student's social style has significant effects on evaluations of professors (Schlee, 2005). Specifically, students tend to rate a professor more highly if the student and professor share the same social style. It is possible, then, that students struggle more to determine differences in evaluations of professors they perceive as more effective because they

perceive those professors to be similar to themselves whereas for professors they perceive as dissimilar to themselves, they may be able to more easily make distinctions.

Another factor that likely contributed to the minimally meaningful differences identified in our study is the distribution of good vs. bad professors experienced by our participants. Our results were highly dependent on the various types of professors our respondents happened to have experienced. The size of the difference between best and meaningfully second-best professor, for example, will differ across respondents depending upon just how good and/or how less good their best and meaningfully second-best professor happened to have been. Students who have had primarily good professors will have provided different responses to our survey compared to students who may have frequently had poor or mediocre professors. Ideally, all of our respondents would have had a large number of professors spanning the full poor to excellent continuum, but that is unlikely. Again, replications with larger samples will be needed to ensure sufficiently generalizable minimally meaningful difference values.

It is also important to recognize that a minimally meaningful difference of .80 may not be universally applicable. Previous research has shown that students give female professors significantly poorer evaluations than male professors even, even when controlling for the gender of the student, course division of the class, the professor's years of teaching, and tenure status (Basow & Silberg, 1987). Although the number of female professors in the United States has risen in recent years, there is evidence that students demand more from female professors than they do from male professors; specifically, female professors are expected to grant more special favor requests than male professors (El-Alayi, 2018). This gives female professors more opportunities to disappoint the student and could produce lower ratings that have nothing to do with the quality of instruction. Thus,

when comparing faculty, caution is warranted, whether or not minimally meaningful differences are being used.

In addition, there is evidence that race contributes to the ratings of professors. Students on average evaluate African American professors to be significantly less competent than white or Asian professors (Bavishi et al., 2010). Given these findings, it is possible that race or gender could have impacted our results. Perhaps a respondent was thinking of two white men when asked to evaluate the two best professors, whereas they were comparing a white man to an African American woman when asked to compare the two worst professors. Because race and gender are often associated with, and can contribute to, perceptions of teaching effectiveness, it is possible that the differences in the positive and negative minimally meaningful differences could have been arisen from these types of factors.

A potentially important limitation of the current investigation is that a number of assumptions were made when we imported methods used to identify meaningful health differences experienced by patients to the identification of meaningful differences in teaching evaluations. For example, one assumption was that a two-week period was the proper interval to assess the test-retest reliability of the traditional 5-point teaching effectiveness item. One could argue for a shorter or longer period of time, or that the assessment should have been conducted closer to the end of a semester. We also made assumptions about the types and sizes of samples that would be needed for the study. Generally, larger sample sizes are better for these types of analyses, but our sample sizes were primarily samples of convenience. Also, because we calculated the SEM using a sample of Butler University students, it could be argued that we should have used a Butler sample for the anchor-based approach, even though the CloudResearch sample provides findings that are probably more generalizable. However, we in fact did administer an anchor-based survey to a sample of

Butler students as part of the current thesis. Time constraints prevented us from fully analyzing the data, but at first opportunity, we intend to compare ratings from the Butler and CloudResearch samples.

Finally, in the anchor-based part of the investigation, we assumed that the most direct approach to establishing a meaningful difference was the best approach. For example, question one directly asks participants to identify the best professor and then a professor who is still good, but meaningfully less so, and uses the resulting difference to establish the minimally meaningful difference. Contrast that approach to the approach utilized by question three. Question three asks participants to make judgments about three different good professors, who may or may not differ from each other in terms of perceived quality. The ratings from participants who found the best of the three professors to differ meaningfully from the second best was used to compute a 'meaningful difference' that was used to validate the meaningful difference from question one. Theoretically, question three's meaningful difference could be used as an estimate of the minimally meaningful difference. Because it was elicited through a more indirect approach, however, we assumed the values it produced would be less valid. Thus, we opted to use the most direct method to identify the minimally meaningful difference.

An issue related to validity is whether and how minimally meaningful differences should be used when evaluating professors. The methods we employed to establish minimally meaningful differences are promising, and the data suggest the minimally meaningful differences possess acceptable levels of validity. However, as noted earlier, the relatively large standard deviation associated with the minimally meaningful differences suggests caution is warranted when using minimally meaningful differences in an evaluative context. However, even with the somewhat limited samples utilized in the current investigation, confidence intervals could be computed around

minimally meaningful difference values to provide insight into the smallest and largest values the minimally meaningful difference could be. Ultimately, of course, when evaluating teaching effectiveness, multiple pieces of evidence should be used, not just ratings from a single teaching effectiveness item.

In terms of future research, studies should examine minimally meaningful differences while controlling for gender, race, and social style of professor to see if the differences between positive and negative minimally meaningful differences persist. In addition, future research should see if results change across institutions or institution type (e.g., is the minimally meaningful difference different at a small, private college compared to a large, public university). It would also be interesting to control for the number of times a student has the professor they rate to see if an individual's rating changes depending on how well they know that professor. Lastly, once minimally meaningful differences are more firmly established, it would be interesting to obtain qualitative information from students about their professors to determine which aspects of the professors and their teaching contribute most to perceptions of quality differences across instructors.

This study sought to establish a minimally meaningful difference for student's evaluations of professors on the traditional 5-point teaching effectiveness item. Two minimally meaningful differences were found. One from the positive perspective (i.e., when students evaluate good professors) and one from the negative perspective (i.e., when students evaluate bad professors). The positive approach produced a minimally meaningful difference of .84; the negative approach produced a minimally meaningful difference of .75. Although the difference between minimally meaningful differences was statistically significant, the difference was small in absolute terms, suggesting the minimally meaningful difference is relatively stable across the 5-point scale typically employed when assessing teaching effectiveness.

References

Basow, S.A., and N.T. Silberg. (1987). Student evaluations of college professors: Are female and male professors rated differently. *Journal of Educational Psychology 79*(3):308-314. Doi:10.1037/0022-0663.79.3.308.

Baumeister, R.; Finkenauer, C.; Vohs, K. (2001). Bad is stronger than good. Review of General Psychology, 5, 323–370.

Bavishi, A., Madera, J.M., and R. Michelle. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education 3*(4):245-256. Doi: 10.1037/a0020763.

Buhrmester, M., Kwang, T., and SD Gosling. (2011) Amazon's mechanical turk: A new source of inexpensive, yet reliable data? *Perspect Psychol Sci. 6*(1):3-5. Doi:10.1177/1745691610393980

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. Social Psychological and Personality Science, 11(4), 464–473. https://doi.org/10.1177/1948550619875149

Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.

El-Alayi, A., Hasen-Brown, A.A., and Ceynar, M. (2018). Dancing backwards in high heels: Female professors experience more work demands and special favor requests, particularly from academically entitled students. *Sex Roles: A Journal of Research 79*(3-4):136-150. Doi: 10.1007/s11199-017-0872-6.

Falissard, B., Sapin, C., Loze, J., Landsberg, W., & Hansen, K. (2016). Defining the minimal

clinically important difference (MCID) of the Heinrichs-carpenter quality of life scale (QLS). *International Journal of Methods in Psychiatric Research 25*(2):101-111. DOI:10.1002/mpr.1283.

Feldman, K. A. (1976) Grades and college students' evaluations of their courses and teachers. *Research in Higher Education 4*(1):69-111. Doi: https://doi.org/10.1007/BF00991462

Giesler, R. B. (2010). Detecting clinically significant differences when assessing quality of life. In Kattan, M. W., ed. *Encyclopedia of Medical Decision Making.* Thousand Oaks, CA: Sage Publications.

Horn, J. L. (1971). Integration of Concepts of Reliability and Standard Error of Measurement. Educational and Psychological Measurement, 31(1), 57–74. https://doi.org/10.1177/001316447103100104

Miller, W. R. & Manuel, J.K. (2008) How large must a treatment effect be before it matters to practitioners? An estimation method and demonstration. *Drug and Alcohol Review 27*(5):524-528. DOI: 10.1080/09595230801956165

Motl, R.W., Learmonth, Y.C., Pilutti, L.A., Dlugonski , D., & Klaren, R. (2014). Validity of Minimal Clinically Important Difference Values for the Multiple Sclerosis Walking Scale-12. *European Neurology 71*(3-4):196-202. DOI: 10.1159/000356116

Raman, S., Ding, K., Chow, E., and Meyer R.M., Nabid, A., Chabot, P., … Brundage, M. (2016). Minimal clinically important differences in the EORTC QLQ-BM22 and EORTC QLQ-C15-PAL modules in patients with bone metastases undergoing palliative radiotherapy. *Qual Life Res 25*(10):2535-2541. Doi: 10.1007/s11136-016-1308-4.

Ringash, J., O'Sullivan, B., Bezjak, A., & Redelmeier, D.A. (2007) Interpreting Clinically Significant Changes in Patient-Reported Outcomes. *Cancer, 110*(1):196-202.

DOI: 10.1002/cncr.22799.

Salaffi, F., Stancati, A., Silvestri, C.A., Ciapetti, A., & Grassi, W. (2003) Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *European Journal of Pain 8*(4):283-291. DOI:10.1016/j.ejpain.2003.09.004

Schlee, R.P. (2005). Social styles of students and professors: Do student's social styles influence their preference for professors? *Journal of Marketing Education 27*(2):130-142. Doi: 10.1177/0273475305276624.

Wolpert, M., Görzig, A., Deighton, J., Fugard, A.J.B., Newman, R., & Ford, T. (2015). Comparison of indices of clinically meaningful change in child and adolescent mental health services: difference scores, reliable change, crossing clinical thresholds, and 'added value' – an exploration using parent rated scores on the SDQ. *Child and Adolescent Mental Health 20*(2):94-101. DOI:10.1111/camh.12080

Wyrwich, K. W., Bullinger, M., Aaronson, N., Hays, R. D., Patrick, D. L., Symonds, T., and the Clinical Significance Consensus Meeting Group (2005). Estimating clinically significant differences in quality of life outcomes. Quality of Life Research, 14, 285- 295.

Zisapel, N. & Nir, T. (2003) Determination of the minimal clinically significant difference on a patient visual analog sleep quality scale. *J. Sleep Res., 12*(4):291-298. PMID: 14633240.

Appendix A

Question 1. We'd like you to think about two different professors.

First think of the best professor you've had. Now think of a professor who is also good, but a noticeable/meaningful step lower in quality than the best professor you've had.

Next, we ask you to rate each professor using this scale.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Question 2. Now we'd like you to think of two new professors.

Think of the worst professor you've had. Now think of a professor who is also bad, but a noticeable/meaningful step better in quality than the worst professor you've had.

Next, we ask you to rate each professor using this scale.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Question 3. Think of three of the better professors you've had. Imagine you had to rank them (no ties allowed) in terms of general quality.

Take a minute and think to yourself who of the three is the best, who is second best, and who is third best.

Do you agree that the difference in quality between the best and the second best is big enough to be a meaningful difference?

- Yes
- No
- Not Sure

Do you agree that the difference in quality between the second best and the third best is big enough to be meaningful?

- Yes
- No
- Not Sure

Do you agree that the difference in quality between the best and the third best is big enough to be a meaningful difference?

- Yes
- No
- Not Sure

Now please rate each professor on a 1 to 5 scale where 1 is poor and 5 is excellent. You can use any kind of decimal (e.g. 3.80) you need to make your rating accurate.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Question 4. Now think of the three worst professors you have ever had. Imagine you had to rank them. Please take a moment to rank them in your head.

Do you agree that the difference in quality between the worst and the second worst is big enough to be a meaningful difference?

- Yes
- No
- Not Sure

Do you agree that the difference in quality between the second worst and the third worst is big enough to be a meaningful difference?

- Yes
- No
- Not Sure

Do you agree that the difference in quality between the worst and the third worst is big enough to be a meaningful difference?

- Yes
- No
- Not Sure

Now please rate each professor on a 1 to 5 scale where 1 is poor and 5 is excellent. You can use any kind of decimal (e.g. 3.80) you need to make your rating accurate.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Question 5. Think of two of the better professors you've had – picture them in your mind. Now imagine you had to rank them (no ties allowed!) in order of general quality. Take a minute and think who is best and who is second best.

Now, please rate each professor on a 1 to 5 scale where 1 is poor and 5 is excellent.

You can use any kind of decimal (e.g. 3.80) you need to make your rating accurate.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Now imagine you were a Dean and that you could give $1,000 raise to one of the two professors. You have a single $1,000 raise – it can't be split.

Is the best professor sufficiently better than the second best professor for you to feel justified giving the best professor the raise?

- Yes, the best professor is meaningfully better

- No, the two professors are too close in quality

If participants answered no, they received the next part of the question. If they answered yes, the question was done.

Since you answered no, meaning that the professors are too close in quality, keep thinking of the other professors you've had who are not as good as the best professor.

Keep thinking until you settle on one that is of sufficiently lesser quality such that if you had to choose between the best professor and the lesser quality professor, you WOULD feel justified giving the best professor the raise.

Now please rate that lesser quality professor on a 1 to 5 scale where 1 is poor and 5 is excellent.

You can use any kind of decimal (e.g. 3.80) you need to make your rating accurate.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Question 6. Now we want you to think of two of the worst professors you've ever had. Imagine you had to rank them. Please rank them in your mind and remember that there can be no ties.

Now, please rate each professor on a 1 to 5 scale where 1 is poor and 5 is excellent.

You can use any kind of decimal (e.g. 3.80) you need to make your rating accurate.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Now imagine you were the Dean and you were faced with budget cuts that require you to have to fire one of these professors, but you had too many students to fire more than one professor.

Is the worst professor sufficiently worse than the second worst professor so that you can feel justified firing the worst professor?

- Yes, the worst professor is meaningfully worse
- No, the two professors are too close in quality

Since you answered “no, the two professors are too close in quality,” please think of other professors who are better than the worst professor. Keep thinking until you settle on one that is of sufficiently higher quality such that if you had to choose between the worst professor and the higher quality professor, you would feel justified firing the worst professor.

Now please rate that lesser quality professor on a 1 to 5 scale where 1 is poor and 5 is excellent. You can use any kind of decimal (e.g. 3.80) you need to make your rating accurate.

This was then followed by a 1-5 scale with 1 labeled poor and 5 labeled excellent.

Table 1. Positive Approach Descriptive Statistics

	n	Range	Mean (SD)
Question 1			
Best Professor	163	3.32-5.00	4.7685(.3490)
Less Good Professor	163	1.42-4.90	3.9283(.5467)
Question 3			
Best Professor	112	3.63-5.00	4.7868(.3172)
Second Best Professor	112	3.02-4.79	4.0977(.3983)
Question 5			
Good Professor	120	2.14-5.00	4.6615(.4487)
Less Good Professor	120	2.76-4.90	4.1201(.4323)

Table 2. Negative Approach Descriptive Statistics

	n	Range	Mean (SD)
Question 2			
Worst Professor	163	1.00-3.55	1.5133(.5594)
Less Bad Professor	163	1.06-3.94	2.2637(.5826)
Question 4			
Worst Professor	93	1.00-3.42	1.3227(.4536)
Second Worst Professor	93	1.05-3.67	2.0546(.4952)
Question 6			
Bad Professor	110	1.00-2.81	1.3405(.4318)
Less Bad Professor	110	1.06-3.41	2.0167(.4851)