Student evaluations of teaching: effects of the Big Five personality traits, grades and the validity hypothesis

Carol Lynn Patrick*

Department of Psychology, Fort Hays State University, 600 Park Street Hays, KS 67601, USA

The purpose of the current study was to examine whether the Big Five personality traits and expected student grades relate to student evaluations of teachers and courses at the college level. Extraversion, openness, agreeableness and conscientiousness were found to be personality traits favoured in instructors, whereas neuroticism was not. A significant correlation was found between the students’ expected grades in the course and student evaluations of the course, but not the evaluations of the instructor. When the effect of students’ perceived amount of learning was taken into account, no significant effect of grades was found on teacher ratings. Personality explained variance in teacher and course evaluations over and above grades and perceived learning.

Keywords: teaching; evaluation; personality; Big Five; grades

The effects of student and instructor personality on teaching ratings and course evaluations

Teaching evaluations are important on college campuses. For professors, evaluations may partially determine tenure, promotion or merit. For students, evaluations may help determine the faculty the university will retain to teach them in the future. This will affect the quality of their education. Much concern has been expressed about the possible sources of bias that may affect student evaluations of teaching (SETs). Because 85% of colleges/universities utilise SETs as a part of the instructor evaluation process, and 50% use SETs solely for instructor evaluation, it is essential to study sources of bias in the SET process (Seldin 1993).

Sources of bias in SETs

A number of factors have been suggested as possible sources of bias in student evaluations. They may bias evaluations by having effects on the ratings given, yet being out of the control of the instructor who is being evaluated. Discussed below are the most commonly researched possible biases. In this review: (1) those variables where the effects on SETs are fairly well established in the literature will first be described, followed by; (2) those where findings have inconsistent results or have been studied infrequently.

*Email: cpatrick@fhsu.edu
**Strong support**

The majority of the literature supports the idea that instructors who teach required courses are rated less favourably by students than instructors who teach elective courses. Although a few studies have found no relationship (e.g. Hildebrand, Wilson, and Dienst 1971), the majority do (e.g. Divoky and Rathermel 1988; Feldman 1978; Petchers and Chow 1988; Scherr and Scherr 1990).

The research also consistently reports that instructors teaching higher-level courses are rated more favourably than those teaching lower-level courses (e.g. Goldberg and Callahan 1991; Moritsch and Suter 1988). However, it is important to remember that freshmen and sophomores are more likely to be taking required courses, rather than elective ones, and this may in part explain these results.

Studies consistently show that social sciences and humanities instructors are rated more favourably by students than math and science instructors (e.g. Cashin 1992; Centra 1993). The reason for this is not fully understood, but the results are consistent.

**Limited/no support**

The time of day the class is held has been suggested as a possible factor biasing student ratings. There are few studies on this topic, but they consistently find no relationship (e.g. Feldman 1978).

Although sources of bias on SETs with either strong support or limited support have well-established results, they are not the focus of the present study. Instead, sources of bias with inconsistent results or inadequate study are the variables to be examined. The effect of grading practices and personality of the instructor are among these variables and are discussed in the next two sections.

**Grading**

Perhaps the most studied (and certainly the most disputed) possible source of bias in teaching evaluations is grading practices. There is a generally held belief that instructors who give higher grades (or whose students expect they will give higher grades) receive higher ratings. Does the research support this belief? While there is some limited support, analyses of the research that examines correlations between student grades (or expected grades) and SETs show minimal to no relationships. In fact, the average correlation between grades and student ratings across all studies is essentially zero. Twenty-four individual studies report a zero correlation between grades and student ratings; whereas 37 studies report a positive relationship and one a negative relationship. Of those reporting positive correlations, the mean correlation was .18 (with a standard deviation of .16), generally considered to be a minimal relationship. And, in the studies where positive correlations are found, they are more likely to be when ‘homemade’ student ratings forms (in which the reliability and validity of the instrument are not established) are used (Arreola 2000).

So, there appears to be little strong evidence that reliable and valid student evaluation instruments are affected a great deal by grades. If they are affected, it is generally believed to be a student endorsement of teaching leniency; that is, that students are rating the teachers highly, not because they are ‘good’ teachers, but because they give easy grades. However, other explanations have been offered for the grade/ratings correlation, indicating that the situation may be more complex. Possible
explanations when there is a correlation between grades and student evaluations (Marsh 1987):

- The leniency hypothesis – the professor’s leniency in assigning grades favourably influences student evaluation scores.
- The validity hypothesis – students who have learned more in the class will receive higher grades and will naturally rate the professor more highly because of the knowledge they have gained in the course.
- The prior characteristics hypothesis – particular student or course factors (such as student motivation or class size) will affect student learning, grades received and actual teaching quality. For instance, students who have a high level of motivation will probably learn more, receive higher grades, and therefore give higher ratings to their professors.

Although the second and third hypotheses have not been studied extensively, limited evidence has been found to support each (Marsh and Roche 2000). In summary, then, the research does not strongly support a strong relationship between grades and student ratings of instructors, although results are mixed. Also, there may be other plausible explanations besides grading leniency when such relationships do exist. Thus, the validity hypothesis is a focus of the current study.

**Instructor personality and SETs**

One of the sources of bias affecting the accuracy of teaching evaluations has been suggested to be instructor personality. That is, students’ view of teaching effectiveness may be more swayed by the instructors’ personality characteristics than by other teaching characteristics (including how much students learn). For example, instructors who are perceived as warm, kind, funny, enthusiastic or entertaining may receive higher student ratings, regardless of their level of knowledge in the subject being taught (e.g. Ahmadi, Helms, and Raiszadeh 2001; Feldman 1986; Widmeyer and Loy 1988; Wilson 1998). Warm interpersonal style, personal charisma, self-assurance and professionalism in teaching approach have been indicated as important qualities of an ideal instructor (Bennett 1982).

It is important to mention that teachers’ own assessments of their personalities have not been found in the past to relate to SETs (Feldman 1986), although student perceptions of the teachers’ personalities have been found to relate (e.g. Clayson and Sheffet 2006). Research examining the reliability and validity of personality judgements shows that assessments made by a single acquaintance or a group of acquaintances is more accurate than self-reports (Kolar, Funder, and Colvin 2006).

Personality is arguably an amorphous concept, and certainly a multifaceted one, with much variance in how it is conceptualised and evaluated. A number of different personality tests have been examined to ascertain their relation to SETs. Several are described here.

For example, in a study utilising the Myers-Briggs Type Indicator, it was found that extraverted teachers received significantly higher SETs than introverted teachers (Hart and Driver 1978). Also, intuitive professors were generally preferred more than sensing professors, and feeling more than thinking. When the California Psychological Inventory was used to measure teacher personality, it was found that being affective, sensitive, rational, practical and independent related to higher SETs (Jenkins and Downs 2001).
The Big Five personality traits

**NEO PI-R**

Currently, one of the major theories of personality is based on the Revised NEO Personality Inventory (NEO PI-R) (Costa and McCrae 1992). The NEO PI-R measures personality on five domains (also known as ‘the Big Five’): neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. Although this theory is currently one of the most favoured theories of personality (on Google scholar the Big Five personality measures appear in more scholarly articles than either the 16 Personality Factor (16 PF) or the California Psychological Inventory), and it is parsimonious in encapsulating personality in only five domains, few studies exist that examine the relation between its domains and SETs.

**Neuroticism** is defined as an individual’s emotional stability, and includes the facets of anxiety, hostility, depression, self-consciousness, impulsiveness and vulnerability. Those with high scores in neuroticism are more likely to experience psychological distress and show maladaptive coping. Those with low scores tend to be calm and relaxed and cope effectively with stress. They are more likely to have irrational ideas and exhibit impulsiveness.

A combination of interpersonal interaction skills, positive affect and energy level make up the domain of **extraversion**. Warmth, gregariousness, assertiveness, activity level, excitement-seeking behaviours and positive emotions are the facets of extraversion. Low scorers tend to prefer to be alone rather than with others, and are more reserved and quiet than high scorers. Those who score high on extraversion are more social, energy-filled and optimistic.

The **openness to experience** domain describes a person’s interest in new experiences and ideas in different areas. These areas are: fantasy, aesthetics, feelings, actions, ideas and values. High scorers are open to new experiences in these areas; low scorers tend to prefer what is familiar to them.

**Agreeableness** is the domain that examines the ways in which one interacts with others in terms of level of trust, straightforwardness, altruism, compliance, modesty and tender-mindedness. Trust, sympathy, helpfulness and compassion are typical of high scorers, whereas low scorers are more distrustful, self-centred and antagonistic and prefer to compete rather than collaborate with others.

The domain of **conscientiousness** encompasses the ability to control impulses, and to be organised and motivated. Its facets are competence, order, dutifulness, achievement striving, self-discipline and deliberation. This domain measures one’s ability to control impulses and delay gratification. A high scorer could be described as goal-oriented, organised, punctual and dependable, and a low scorer as impulsive and unmotivated.

The Big Five and SETs

Research has shown that the Big Five personality traits have been associated with job-related variables. For example, Holland occupational interest scores have been found to correlate with Big Five scores (Costa, McCrae, and Holland 1984), and extraversion, one of the Big Five personality traits, has been useful in predicting job performance (Organ and Ryan 1995). But do the Big Five relate to teaching evaluations?

The NEO PI-R was used in a study regarding personality traits of high school teachers and their relation to SETs (Cutchin 1999). Self-ratings of personality by
teachers were utilised. Different facets of the five domains were found to be important to students in their ratings of teachers. One facet of extraversion (gregariousness) was significantly and positively related to SETs, as were two facets of openness (openness to ideas, openness to values), one facet of agreeableness (trust) and two facets of conscientiousness (order and achievement striving). Unlike Cutchin’s study of high school teachers, college instructors’ own assessments of their personalities have not been found in the past to relate to SETs (Feldman 1986).

In a college sample, Clayson and Sheffet (2006) examined the relation of the Big Five personality traits to SETs. In this study, student perceptions of teacher personality were measured on a five-point Likert scale (e.g. rating level of agreeableness on a scale from 1 to 5 with the term ‘disagreeable’ at one end of the scale and ‘agreeable’ at the other). Student ratings of teaching were also gathered. At the end of the term, it was found that the five personality scales together explained 64% of the variance in SETs, with each individual Big Five trait accounting for between 12% and 55% of the SET variance.

Therefore previous research has shown that students’ perceptions of the Big Five personality characteristics relate to their ratings of teaching effectiveness. However, there has only been one study conducted at the college level, and it did not utilise an instrument that has been shown to be reliable and valid at measuring the Big Five. It also appears to have completely, or predominantly, used a sample of business teachers and students, which may not be generalisable to universities as a whole.

The purpose of the current study was to examine whether Big Five personality traits, measured in a reliable and valid manner, related to individual teacher ratings and course evaluations at the college level. Specifically, the relation of students’ self-expressed personality traits and students’ assessments of instructor personality traits to teacher and course ratings were measured. It was hypothesised that there would be a positive relation between SETs and extraversion, openness, agreeableness and conscientiousness; that is that instructors who rate higher on the indicated personality traits would tend to receive higher SETs. A negative relation between SETs and neuroticism was expected, with teachers having low neuroticism scores tending to receive higher SETs.

The most controversial relation between more studied sources of bias in SETs, grades, reviewed at the beginning of this manuscript, was also examined. A small correlation was expected between students’ expected grades and SETs. It was expected, according to Marsh’s (1987) validity hypothesis that, once the effects of amount students perceive they learned in the course is removed, little to no effect of grades on SETs would remain.

Method

Participants

Participants were 176 students from seven general education courses (both upper and lower division) and 34 majors at a small Midwestern university. The courses were taught by five teachers. Because a sample size of five is not adequate to analyse the instructors’ data with any reasonable power, and because instructor ratings of their own personalities have not been found generally to relate to teaching ratings (Feldman 1986), the instructors’ data were not used in the analyses. The students’ mean age was 21.22 (SD = 3.38), and ages ranged from 17 to 52. Seventy-three were males (41.5%)
and 103 were females (58.5%). Twenty-one percent of the sample were freshmen, 27.8% sophomores, 24.4% juniors and 36.7% seniors.

**Instruments**

*Big Five Inventory*

Students filled out two copies of John, Donahue, and Kentle’s (1991) Big Five Inventory (BFI): one assessing their own personality traits, and another assessing their instructors’ personality traits. Course instructors also completed the BFI about themselves, although the data were not used in the current study.

The BFI is a 44-item questionnaire designed to measure the Big Five personality traits of neuroticism, extraversion, openness, agreeableness and conscientiousness. It was selected for use in the current study, rather than the NEO PI-R (Costa and McCrae 1992), because it is much shorter than the NEO PI-R, but still has good reliability and convergent validity with that instrument.

Internal consistencies of the five BFI scales range from .75 to .90, averaging above .80. Test–retest reliabilities have been found to range from .80 to .90. Convergent validity with each corresponding scale of the NEO ranges from .83 to .97, with a mean of .92 (Pervin and John 1999).

*Demographics and teaching questionnaire*

Students also completed an author-created demographics and teaching questionnaire. This survey asked students to evaluate their teacher and the course (on a four-point scale), the amount they learned in the course (on a three-point scale), the grade they expected to receive, and their major, age, gender and year in school. The self-constructed scale to measure teaching ratings was used because ‘homemade’ measures of teaching effectiveness are more prone to bias than highly examined, reliable and valid measures (Arreola 2000). Yet they are all too frequently utilised at universities. Therefore, just such a homemade measure was chosen for the current study to adequately represent the student ratings instruments at many universities, and the worst case scenario of susceptibility to bias. Amount learned was measured on a three-point Likert-type scale, rather than the more typical five- or seven-point Likert scale, because students, as novices to areas of study, may not be able to make finely tuned judgements about how much they have learned. Therefore, it was believed that they could decide whether they were learning more, less or the same as they did in typical classes.

**Procedure**

Student participants were tested at the end of class periods during the last three weeks of the semester, by permission of their instructor. The instructor left the classroom to fill out a BFI measuring their personality traits. After signing an informed consent, students were given a packet of surveys to complete. The packets contained a BFI for students to complete about their own personalities, and one to rate their instructor’s personality. These two surveys were randomly counterbalanced as to which of the two were presented first. The final survey was the demographics and teaching questionnaire.
Results
Pearson correlations were computed among the teacher rating score, the course evaluation score, and the three different measures of the Big Five personality traits.

Students’ assessments of instructor personality traits
Correlations were observed at the $p < .01$ level among all Big Five personality traits (as measured by students’ assessments of instructor personality traits), course evaluation scores and teacher ratings (see Table 1). Higher course evaluations and teacher ratings were found with higher levels of extraversion, openness, agreeableness and conscientiousness, and lower course evaluations and teacher ratings with higher levels of neuroticism.

Students’ self-expressed personality traits
The only correlation found between students’ self-expressed personality traits and teacher ratings or course evaluations was a positive correlation between student agreeableness and rating of the instructor’s teaching ability (see Table 1). This result indicates that students who rate higher in the personality trait of agreeableness tend to rate their instructor as more effective.

Grades and SETs
The correlation between expected grades and overall rating of the course was significant ($r = .21, p < .01$), indicating that the grades students expected to receive in the course explained only about 4% of the variance in the overall ratings of the course. However, the correlation between expected grades and the overall rating of the effectiveness of the teacher was not significant ($r = .12, p = .13$), indicating that the grade the student expected to receive in the course did not significantly impact their opinion of the effectiveness of the teacher.

Table 1. Correlations between overall teaching ratings and Big Five domain scores ($N = 176$).

<table>
<thead>
<tr>
<th>Students’ rating of instructor on:</th>
<th>Students’ rating of class</th>
<th>Students’ rating of instructor’s teaching ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroticism</td>
<td>$-0.31^{**}$</td>
<td>$-0.23^{**}$</td>
</tr>
<tr>
<td>Extraversion</td>
<td>$0.27^{**}$</td>
<td>$0.27^{**}$</td>
</tr>
<tr>
<td>Openness</td>
<td>$0.43^{**}$</td>
<td>$0.33^{**}$</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>$0.30^{**}$</td>
<td>$0.23^{**}$</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>$0.35^{**}$</td>
<td>$0.33^{**}$</td>
</tr>
<tr>
<td>Students’ rating of self on:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>$0.09$</td>
<td>$0.03$</td>
</tr>
<tr>
<td>Extraversion</td>
<td>$-0.09$</td>
<td>$0.01$</td>
</tr>
<tr>
<td>Openness</td>
<td>$0.14$</td>
<td>$0.14$</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>$0.12$</td>
<td>$0.15^{*}$</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>$0.07$</td>
<td>$0.13$</td>
</tr>
</tbody>
</table>

* $p < .05$; ** $p < .01$. 
The validity hypothesis, personality and SETs

The validity hypothesis (Marsh 1987) states that, when a correlation between grades and SETs occurs, one explanation could be that students who learn more in a course will: (1) receive a higher grade; and (2) rate the teacher more favourably. In order to test this hypothesis, two hierarchical multiple regressions were computed. In one, the criterion was the student rating of the course overall, and in the second, the student rating of the instructors’ effectiveness. In both regressions, predictors were entered in three steps. In the first step, the amount the students indicated they learned was the only variable included. In the second step, the expected grade variable was added, to see if it explained variance over and above that of the amount learned alone. The second step is the test of the validity hypothesis. In the third step, the student ratings of the Big Five personality variables in their instructors were added, to understand whether personality explains any variance over and above that of amount learned and expected grade.

Hierarchical regression – course rating

In the first step, amount learned explained a significant amount of variance in the overall course rating, $R^2 = .23$, $F(1,171) = 49.70$, $p < .001$. When expected grade was added in Step 2 [$R^2$ chg = .03, $F(1,170) = 7.66$, $p < .01$], a significant amount of additional variance in course rating was noted, as it was in Step 3 when personality variables were added [$R^2$ chg = .12, $F(5,165) = 6.34$, $p < .001$]. Openness was the personality variable that added significantly to the explanation of variance in course ratings in Step 3 (see Table 2).

Table 2. Hierarchical multiple regression statistics: overall rating of class regressed on student ratings of amount learned in course, expected grade in course and student ratings of instructor personality characteristics on the BFI.

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SEB</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount learned</td>
<td>.564</td>
<td>.080</td>
<td>.475**</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount learned</td>
<td>.550</td>
<td>.079</td>
<td>.463**</td>
</tr>
<tr>
<td>Expected grade</td>
<td>.148</td>
<td>.053</td>
<td>.183**</td>
</tr>
<tr>
<td>Step 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount learned</td>
<td>.483</td>
<td>.075</td>
<td>.406**</td>
</tr>
<tr>
<td>Expected grade</td>
<td>.105</td>
<td>.052</td>
<td>.130*</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>-.004</td>
<td>.012</td>
<td>-.026</td>
</tr>
<tr>
<td>Extraversion</td>
<td>-.009</td>
<td>.010</td>
<td>-.072</td>
</tr>
<tr>
<td>Openness</td>
<td>.037</td>
<td>.010</td>
<td>.325**</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.006</td>
<td>.012</td>
<td>.045*</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.006</td>
<td>.011</td>
<td>.047*</td>
</tr>
</tbody>
</table>

*p < .05; **p < .01.

Note: Step 1: $R^2 = .23$ ($p < .001$), adjusted $R^2 = .22$; Step 2: chg $R^2 = .03$ ($p < .01$); Step 3: chg $R^2 = .12$ ($p = .001$). Measures of personality in Step 3 are all based on student ratings of instructor personality on the BFI.
Hierarchical regression – teacher rating

In the first step, amount learned explained a significant amount of variance in the overall instructor rating, $R^2 = .08$, $F(1, 171) = 15.85$, $p < .001$. When expected grade was added in Step 2 [$R^2_{chg} = .01$, $F(1, 170) = 1.82$, $p > .05$], a significant change was not noted, indicating that expected grades did not explain a significant amount of variance in instructor ratings. A significant change of explanation of variance in instructor ratings was noted when, in Step 3, personality variables were added [$R^2_{chg} = .11$, $F(5, 165) = 4.36$, $p = .001$]. Openness and conscientiousness were the personality variables that added significantly to the explanation of variance in instructor ratings in Step 3 (see Table 3).

Discussion

Personality of the instructor, when assessed by the students, seemed to have an effect on student ratings of both the course and the effectiveness of the teacher. Extraversion, openness, agreeableness and conscientiousness were personality traits favoured in instructors, whereas neuroticism was not. Courses of teachers who exhibited the same traits were also rated higher. This result was expected based on previous research utilising a different instrument for measuring Big Five personality traits (Clayson and Sheffet 2006). Therefore, the current research supports the concept that personality relates to SETs and may be a source of bias in evaluations.

When students’ ratings of their own personalities were correlated with SETs, only one significant association was found. Agreeable students tended to report that their teachers had better teaching skills. Given that agreeable people are sympathetic and compassionate, it makes sense that as students they would be more likely to supportive of their teacher (if not the course itself).
The association between expected grades and course ratings was significant, whereas the association between expected grades and ratings of instructors was not. This indicates that the grade the student expected to receive in the course did not significantly impact their opinion of the effectiveness of the teacher, even though it had a small effect on whether they liked the class.

In the test of the validity hypothesis (Marsh 1987), it was found that when the effect of the amount students perceived they learned in the course was removed ($\beta = .46$), only a small effect of expected grades ($\beta = .18$) was found on the rating of the course. This provides some support for the validity hypothesis, that the often-found relationship between grades and evaluations may be because when students learn more, both grades and their opinion of the course/teacher will be higher. When it came to teacher ratings, expected grade ($\beta = .10$) explained no variance over and above that of the effect of amount student perceived they learned ($\beta = .29$), again supporting the validity hypothesis. In both regressions, there was an effect of instructor personality (as judged by students) over and above that of amount learned and grades, once again supporting the importance of personality in SETs.

In conclusion, it is reasonable to suggest that the traits of neuroticism, extraversion, openness, agreeableness and conscientiousness affect teaching ratings and evaluations of courses. However, whether personality traits are a source of bias in SETs that we should seek to minimise, or whether they actually improve instructional ability, is an issue yet to explore. Research examining the actual learning of students over the course of a semester (rather than their own perceived amount of learning) and its relations with teacher personality and SETs are warranted.

Notes on contributor
Carol L. Patrick is an associate professor of psychology at Fort Hays State University. She received her doctoral degree from Southern Illinois University in Lifespan Developmental Psychology, and also has a master’s degree from Wichita State University in counselling.

References


