J Comput High Educ (2012) 24:58–69
DOI 10.1007/s12528-012-9053-3

# Online versus paper evaluations: differences in both quantitative and qualitative data

**William B. Burton · Adele Civitano · Penny Steiner-Grossman**

Published online: 3 March 2012


**Abstract** This study sought to determine if differences exist in the quantitative and qualitative data collected with paper and online versions of a medical school clerkship evaluation form. Data from six-and-a-half years of clerkship evaluations were used, some collected before and some after the conversion from a paper to an online evaluation system. The quantitative data consisted of a composite score based on the average of several Likert-type items; the qualitative data consisted of open-ended comments about the clerkships. Clerkship ratings were more positive in the online version. Students made significantly longer comments about both strengths and weaknesses on the online form than on the paper form. In addition, comments made on the online form were judged to be more informative and showed less evidence of "negativity" than those made on the paper form. The findings suggest that both quantitative and qualitative data obtained with online evaluation forms can differ in important ways from data collected with paper forms.

**Keywords** Student feedback · Course evaluation questionnaires · Qualitative data · Inter-rater reliability · Factor analysis

## Introduction

In 2004, the medical school that served as the setting for this study began moving from a paper-based course and clerkship evaluation system to a predominantly online system. The process of entering data had become increasingly burdensome, making it difficult to provide feedback in a timely manner. Moreover, it was believed that the existing system in which students completed paper course

W. B. Burton (✉) · A. Civitano · P. Steiner-Grossman
Office of Educational Resources, Albert Einstein College of Medicine, 1300 Morris Park Ave., Belfer 211, Bronx, NY 10461, USA
e-mail: william.burton@einstein.yu.edu

evaluations immediately after examinations, while yielding high response rates, limited the extent to which students could provide meaningful and constructive feedback.

As evaluators, the authors envisioned several benefits of making the transition to an online system, but they soon learned that other stakeholders did not share their optimism. Many feared the change would be accompanied by a large drop in response rates, thereby compromising the quality of the data, and some course leaders and clerkship directors also predicted that their ratings would fall. They argued that disgruntled students would be more highly motivated than other students to fill out their evaluations and, assuming a less than perfect response rate, this differential motivation would bias the results in a negative direction.

Given these concerns, the literature review was focused around two questions: First, are the data, both quantitative and qualitative, obtained from online surveys comparable to data obtained from paper surveys? Second, are the psychometric properties of online surveys similar to those of paper versions of the same surveys? The authors chose to limit the literature search to evaluation studies.

Eighteen published studies from the evaluation field were located that investigated the comparability of quantitative data collected with paper and online surveys. Of these, 14 reported minimal or no differences in the mean or median ratings obtained with the two methods (Ardalan et al. 2007; Avery et al. 2006; Dommeyer et al. 2004; Donovan et al. 2006; Ernst 2006; Gamliel and Davidovitz 2005; Handwerk et al. 2000; Heath et al. 2007; Layne et al. 1999; Liegle and McDonald 2004; Paolo et al. 2000; Rice and Van Duzer 2005; Smither et al. 2004; Thorpe 2002). Of note, though Ernst (2006) found that mean item ratings were identical in the paper and online conditions, extremely high and low ratings were more common in the online condition. He interpreted this to mean that students with strong opinions are more likely than other students to complete online evaluations. In two of the four studies that did report differences (Carini et al. 2003; Tomsic et al. 2000), students gave more positive ratings if they completed the form online, as opposed to on paper. Neither of these studies used random assignment. Chang (2005) found that college students made more positive ratings of instruction when they filled out paper forms than when they used online forms. Kasiar et al. (2002) reported that 29% of the Likert items in the evaluation differed significantly between the online and paper versions, but they did not report the direction(s) of those differences.

Thirteen evaluation studies addressed the comparability of qualitative data (i.e., open-ended comments) collected with paper and online surveys. Eight of the studies reported that respondents were more likely to provide comments on an online form than on a paper form (Anderson et al. 2005; Ballantyne 2004; Donovan et al. 2006; Handwerk et al. 2000; Heath et al. 2007; Johnson 2003; Kasiar et al. 2002; Layne et al. 1999), and eight of the studies reported that the comments provided on online forms were longer than those provided on paper forms (Anderson et al. 2005; Ardalan et al. 2007; Ballantyne 2004; Bullock 2003; Donovan et al. 2006; Hmieleski and Champagne 2000; Rice and Van Duzer 2005). Donovan et al. (2006) also reported that comments collected online were judged to be less favorable and more useful or informative than comments collected on paper. In contrast, Heath

et al. (2007) found that comments provided online were more favorable than those written on paper. Only one study failed to find differences in either the percentage of students who provided comments on paper and online forms, or in the length of comments provided on the two versions.

Five of the evaluation studies cited above provided evidence regarding the psychometric properties of the paper and online instruments they used. Four studies conducted factor analyses, and all four reported that identical factor structures emerged out of the data obtained with the two forms (Chang 2005; Handwerk et al. 2000; Layne et al. 1999; Smither et al. 2004). Chang (2005) also found that the two forms had high convergent validity and similar internal consistency reliabilities. The remaining study (Gamliel and Davidovitz 2005) found that the test–retest reliability of their paper evaluation form was greater than that of the online form, but they interpreted this as an artifact caused by differences in formatting.

The purpose of this study is to determine if differences exist in the quantitative and qualitative data collected with paper and online versions of a clerkship evaluation form. This is a very timely topic, given that that medical schools and other institutions of higher learning are converting to online evaluation systems at a fast pace. It is surprising, however, that only one study of this topic was found that was conducted in the context of medical education (Paolo et al. 2000).

Based on the findings of the literature review, it was hypothesized that the quantitative data collected with paper and online surveys would be comparable, both in terms of their means and their psychometric properties. In contrast, it was hypothesized that there would be systematic differences in the open-ended comments collected using the two methods. Specifically, it was expected that respondents using online forms would make more comments and longer comments compared to respondents using paper forms. However, it was predicted that respondents using online and paper forms would not differ in their tendency to make informative comments or to show evidence of negativity, as defined in the study by Donovan et al. (2006).

## Methods

### Participants

Quantitative data from 4,873 third-year medical student clerkship evaluation forms were used for the study: 2,141 of these were collected with paper forms, and 2,732 were collected with online forms. Qualitative data from 929 third-year medical student clerkship evaluation forms were used for the study: 558 of these were collected with paper forms, and 371 were collected with online forms.

### Instruments

Evaluations of required third-year clerkships were collected using a standard form containing 23 Likert-type items, with response choices ranging from 1 = strongly disagree to 5 = strongly agree. The questions asked students to rate the clerkship in

four general areas: (1) organization/orientation—e.g., "I received an adequate orientation to the functioning of the site" and "My role in the management of patients was clearly defined"; (2) teaching—e.g., "Conferences and seminars were valuable learning experiences" and "Opportunities were provided for performance of supervised physical exams"; (3) evaluation/feedback—e.g., "I received sufficient ongoing feedback about my performance" and "I had a clear understanding of the criteria on which I was being evaluated"; and (4) global assessment—e.g., "This experience furthered my growth and development as a physician" and "Overall, this clerkship was a positive learning experience." All questions were phrased so that a higher level of agreement indicated a more positive response. The bottom of the evaluation form included spaces for students to write two open-ended comments, one regarding perceived strengths ("What were the greatest strengths of this clerkship?") and one regarding weaknesses ("What were the greatest weaknesses of this clerkship?"). The online evaluation form was designed to resemble the paper form as much as possible.

The clerkship evaluation form utilized in this study has been in use since the 1994–1995 academic year. An aggregate score derived from the average of the 23 items collected in the 2000–2001 and 2001–2002 academic years was found to have excellent internal consistency reliability (i.e., Cronbach's alpha = 0.95 and 0.94, respectively).

## Procedures

During the 2004–2005 academic year, the online system was piloted in three of the six required third-year clerkships at the medical school, while evaluation data for the remaining three clerkships were collected with the paper form. Following the success of the pilot test, the remaining three clerkships were added to the online system at the beginning of the 2005–2006 academic year.

For the numeric ratings, data from a six-and-a-half year period were selected for inclusion in the analyses: in addition to the 2004–2005 academic year (described above), data from the 2002–2003 to 2003–2004 academic years were collected entirely with paper forms, and data from the 2005–2006, 2006–2007, 2007–2008 years, along with the first half of the 2008–2009 academic years, were collected entirely with online forms.

For the qualitative data, the authors selected 6 months of clerkship evaluation data from two separate years: the second half of the 2003–2004 academic year, when all six clerkships were evaluated using the paper form; and the second half of the 2005–2006 academic year, when all six clerkships were evaluated using the online form. These two six-month periods were selected because they were the ones closest together in time in which the data were collected with different methods, thereby minimizing potential history effects. While it would have been preferable, in terms of representativeness, to have selected two periods of a full 12 months each, it was felt that this would have been too burdensome for the raters.

All numeric data, along with the comments from the online form, were scanned or downloaded into an Excel database. Comments written on the paper forms were hand-entered into the same database. No corrections or changes in spelling,

grammar, or punctuation were made to the comments about clerkship strengths or weaknesses, with the exception that comments of "none" (or its equivalent) were removed.

## Data analysis

Data from the 23 Likert-type items on the evaluation form were averaged to form an aggregate score. Internal consistency reliability was determined by calculating Cronbach's alpha separately for the aggregate scores derived from data collected with the paper and online forms. Separate factor analyses were also conducted to determine if the 23 items collected with the paper and online forms loaded on similar factors. We used a decision rule of eigenvalues greater than 1.0.

If a difference in ratings occurred as a result of switching to the online system, it should be seen 1 year earlier in the clerkships changing to the online system in 2004–2005, compared to the clerkships changing to the online system in 2005–2006. To test for this, an analysis of variance (ANOVA) was run with clerkship group (early change vs. late change) and academic year as class variables. Evidence for this 1-year lag in the change in ratings would require that a significant interaction between clerkship group and academic year be found. A longitudinal graph of aggregate scores also was inspected to see if changes coincided with the timing of the change to the online system. Within each clerkship group, *t* tests were run to see whether pre- and post-change aggregate scores differed. Effect sizes for quantitative measures were determined by calculating Cohen's d scores (standardized mean differences), defined as the difference between two means divided by the pooled standard deviation.

The qualitative data were analyzed in several ways. First, the numbers of students who offered comments about strengths and weaknesses were tallied. Chi-square tests were used to determine if these numbers differed based on group (i.e., paper vs. online). Second, the numbers of characters typed or written in were calculated separately for the strengths and weaknesses questions. Third, after first being randomized to allow for blinded review, the comments about clerkship weaknesses were classified as being high or low in informativeness. This term was defined as the extent to which a comment provided formative feedback (Donovan et al. 2006)—i.e., made specific reference to something that needed to be changed and/or provided detailed suggestions for how the clerkship could be improved. In a similar blinded manner, comments about clerkship weaknesses also were classified as to whether or not they showed evidence of negativity, which was defined as expressions of anger, frustration, sarcasm, disrespect, or profanity within the comment. In order to be so classified, the negativity had to be present in the comment itself (i.e., comments that merely described unpleasant incidents or experiences in the clerkship did not qualify).

Ratings of informativeness and negativity were made independently by two of the authors, who had a combined total of 28 years of experience in analyzing qualitative evaluation data and communicating with faculty about their interpretation. These two qualities are extremely important in teaching evaluations in general (Donovan et al. 2006; Krupat et al. 2011), and consistently draw the most

attention from faculty and course leaders. The informativeness of comments influences how useful the feedback is for helping teachers improve their courses, and the negativity of comments can influence whether or not a faculty member chooses to return as a teacher the following year. Disagreements between raters were settled by consensus. To provide an estimate of inter-rater agreement, Cohen's kappa coefficients were calculated on the paired ratings of informativeness and negativity. Effect sizes for these two measures were determined by calculating odds ratios.

## Results

Across the six clerkships and the six-and-a-half academic years, response rates ranged between 60 and 85% for the online evaluations, and were about 95% for the paper evaluations. The internal consistency reliabilities of the composite scores were very high, and were equivalent for the paper and online versions (Cronbach's alpha = 0.95 in both conditions). The factor structure of ratings was strikingly similar for the paper and online versions of the evaluation form. For the paper form, the eigenvalue associated with the first factor was 10.7 and it explained 83% of the variance. For the online form, the eigenvalue associated with the first factor was 10.8 and it explained 84% of the variance. In both factor analyses, no other eigenvalue exceeded 1.0. Table 1 shows that the factor loadings of the 23 items on the first factor were essentially identical for the paper and online versions of the evaluation form.

Figure 1 displays the mean aggregate scores across the period of the study. A significant main effect on the ANOVA was found for academic year ($F[6,4846] = 15.9$, $p = 0.0001$), as was a significant interaction between clerkship group and academic year ($F[6,4846] = 2.8$, $p = 0.009$). In combination, these findings indicate that ratings improved following the change to the online evaluation system. In addition, a significant main effect was found for clerkship group ($F[1,4846] = 12.1$, $p = 0.0005$), which may be due to the considerable separation in the ratings of the two clerkship groups in academic years 2006–2007 and 2007–2008. It is unknown what accounts for that temporary difference. A visual inspection of Fig. 1 appears to show that a change in ratings occurred 1 year earlier in the clerkships that made the early switch to the online system. If one looks at each clerkship group separately, the mean aggregate score was significantly lower when the evaluations were completed on paper than when they were completed online. That is, for the clerkships that switched early, the five aggregate post-change scores were higher than the two aggregate pre-change scores (4.2 vs. 3.9, $t = 9.9$, $df = 2{,}351$, $p = 0.0001$). Likewise, for the clerkships that switched later, the four aggregate post-change scores were higher than the three aggregate pre-change scores (4.1 vs. 3.9, $t = 3.9$, $df = 2{,}171$, $p = 0.0001$). For both clerkship groups, all aggregate pre-change scores were below 4.0, whereas all aggregate post-change scores were above 4.0. The effect size for the difference in aggregate scores between the online and paper conditions was relatively small ($d = 0.18$).

**Table 1** Factor loadings of individual items with factor number 1

| Item | Paper (n = 558) | Online (n = 371) |
|---|---|---|
| 1 | 0.63 | 0.63 |
| 2 | 0.66 | 0.63 |
| 3 | 0.75 | 0.73 |
| 4 | 0.49 | 0.51 |
| 5 | 0.64 | 0.59 |
| 6 | 0.44 | 0.48 |
| 7 | 0.49 | 0.57 |
| 8 | 0.57 | 0.61 |
| 9 | 0.64 | 0.66 |
| 10 | 0.59 | 0.60 |
| 11 | 0.71 | 0.73 |
| 12 | 0.63 | 0.57 |
| 13 | 0.75 | 0.77 |
| 14 | 0.72 | 0.75 |
| 15 | 0.76 | 0.78 |
| 16 | 0.72 | 0.74 |
| 17 | 0.76 | 0.75 |
| 18 | 0.69 | 0.69 |
| 19 | 0.66 | 0.68 |
| 20 | 0.77 | 0.74 |
| 21 | 0.78 | 0.74 |
| 22 | 0.80 | 0.81 |
| 23 | 0.85 | 0.84 |

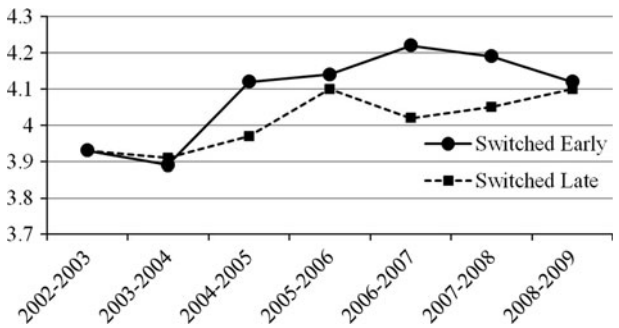


**Fig. 1** Mean clerkship scores by period of study

Several differences were found in the qualitative data that were collected with the paper and online forms (Table 2). Although no difference was found in the percentage of students who provided comments about clerkship strengths ($\chi^2 = 0.11$, $df = 1$, $p = 0.74$), students were significantly more likely to provide

**Table 2** Frequency and length of comments regarding clerkship strengths and weaknesses

| | Paper (n = 558) | Online (n = 371) | *p* value |
|---|---|---|---|
| Strengths | | | |
| Students who commented[a] | 364 (65.2%) | 245 (66.3%) | 0.74 |
| Characters per comment[b] | 98 (64.5) | 145 (76.6) | 0.0001 |
| Weaknesses | | | |
| Students who commented[a] | 330 (59.1%) | 193 (52.3%) | 0.039 |
| Characters per comment[b] | 134 (80.0) | 187 (73.2) | 0.0001 |

[a] Numbers in the paper and online columns refer to number and (percentage)

[b] Numbers in the paper and online columns refer to mean and (standard deviation)

comments about clerkship weaknesses on the paper form than on the online form ($\chi^2 = 4.25$, $df = 1$, $p = 0.039$). In addition, students made significantly longer comments about both strengths (t = 7.59, $df = 608$, $p = 0.0001$, d = 0.42) and weaknesses (t = 6.96, $df = 522$, $p = 0.0001$, d = 0.46) on the online form than on the paper form. For strengths, the average length of comments collected with the online form was 145 characters, compared to 98 characters for the paper form (an increase of 48%). For weaknesses, the average length of comments collected with the online form was 187 characters, compared to 134 characters for the paper form (an increase of 40%).

Inter-rater reliability was high on the ratings of both informativeness (kappa = 0.80) and negativity (kappa = 0.76). As shown in Table 3, comments judged to be high in informativeness were significantly more likely to be made on the online form than on the paper form (90.7 vs. 74.5%, respectively; $\chi^2 = 19.1$, $df = 1$, $p = 0.0001$, odds ratio = 3.15). This is not a surprising finding given the significant differences in length reported earlier, but there were many cases of short comments that were judged to be informative, and lengthy comments that were judged to be uninformative. Table 3 also shows that negativity was significantly more prevalent in comments written on paper forms than in those typed into online forms (17.0 vs. 10.4%, respectively; $\chi^2 = 4.4$, $df = 1$, $p = 0.035$, odds ratio = 1.78).

## Discussion

The findings of this study suggest that both quantitative and qualitative data obtained with online evaluation forms can differ in important ways from data

**Table 3** Informativeness of and negativity in comments regarding clerkship weaknesses [a]

| | Paper (n = 329) | Online (n = 193) | *p* value |
|---|---|---|---|
| Informativeness | 245 (74.5%) | 175 (90.7%) | 0.0001 |
| Negativity | 56 (17.0%) | 20 (10.4%) | 0.035 |

[a] Numbers in each row indicate the number and percentage of respondents who wrote comments rated to be informative or negative

collected with paper forms. With respect to the quantitative data, the hypothesis that ratings obtained with paper and online surveys would be comparable was not supported by the findings. Specifically, the finding that numeric ratings were more favorable when the online form was used is consistent with only 2 of the 18 studies reviewed (Carini et al. 2003; Tomsic et al. 2000). Nevertheless, it must be noted that the effect size for this difference was small.

With respect to the qualitative data, it was hypothesized that respondents using online forms would make more comments and longer comments compared to respondents using paper forms, but that respondents using online and paper forms would not differ in their tendency to make informative comments or to show evidence of negativity. These hypotheses were partly supported by the findings. In line with the majority of other studies, the length of comments about both strengths and weaknesses was greater on the online form than on the paper form. Comments collected online were found to be more informative than those collected on paper, which is consistent with the findings of Donovan et al. (2006). In addition, two aspects of the present findings seem consistent with the findings of Heath et al. (2007), who reported that comments provided online were more favorable than those written on paper. First, students in the present study were less likely to provide comments about clerkship weaknesses on the online form than on the paper form. Second, instances of negativity were less common on the online form than on the paper form. Finally, in contrast to most previous research, students were equally likely to provide comments about clerkship strengths on both the online and paper forms.

In attempting to explain these findings, the answer may have more to do with the setting in which evaluations are completed than with the method itself (i.e., paper vs. online). After finishing a high-stakes exam, students are likely to be exhausted and perhaps upset, and interested in leaving the exam room as quickly as possible. In contrast, being able to complete an evaluation at home provides some time for students to recover from the stress of the exam, and perhaps reflect on the issues they will be asked to rate and comment on in the evaluation.

In support of this interpretation, Ravelli (2000) found that students felt that the online evaluation system allowed them to make more thoughtful comments than did the in-class, paper evaluation system. In addition, Dommeyer (2006) found several differences between evaluations completed inside and outside the classroom, which he attributed to the greater privacy and time allotted to evaluators outside the classroom.

In asking medical students to evaluate their courses, clerkships, and teachers, we always struggle to minimize any "ceiling effects" in the data. Therefore, it could be argued that converting to an evaluation system in which the ratings are even more favorable than they were before the changeover is an undesirable outcome. However, we feel that a strong argument could be made that the data (both the numeric ratings and the comments) are artificially depressed in terms of favorability when students are surveyed immediately after an exam. The opportunity to complete an evaluation in a more "neutral" setting seems to us a better choice, and may produce data that are more "valid" than in an exam room setting.

The major limitation to our study is the fact that students were not randomly assigned to complete the online and paper evaluation forms. Instead, we used a

convenience sample in which the students who completed their evaluations on paper and those who completed their evaluations online were members of different cohorts. It is possible that the differences we reported were the result of maturation or history effects (e.g., the increased experience of clerkship directors, or institution-wide improvements to the clinical curriculum) rather than to changes to the evaluation system. We would argue, however, that splitting the six required clerkships into those that changed early to the new evaluation system and those that changed later allowed us to look for differences that could not be attributed to history or maturation effects. We are aware of no institution-wide changes to the clinical curriculum that are consistent with the timing of the changes in the ratings that we found.

Another limitation is the fact that response rates for the in-class, paper evaluations were higher than those for the online evaluations. In spite of this, response rates for the latter were still quite respectable (60–85%), so we believe that the differences between conditions in both the quantitative and qualitative data cannot be attributed to differences in the response rates.

A final limitation is the fact that online administration of an evaluation is by nature less standardized than administering it to all students in an exam room. This may increase unwanted variability in the resulting data. However, we believe that the many advantages of the online system outweigh this potential disadvantage.

This study was exploratory in nature and thus many of the findings are intriguing, but inconclusive. Nevertheless, this is only the second study to investigate the impact of switching from a paper to online evaluation system in a medical school setting (Paolo et al. 2000).

With the advent of online evaluations of courses and clerkships, we have the opportunity to take much greater advantage of the qualitative data collected alongside the quantitative data in our evaluations. Even without any further changes, we already have found that with online administration students tend to write much more detailed comments than they did with the paper forms, and the fact that these comments are typed makes them much easier for course and clerkship leaders to analyze and act upon. If these comments are also more informative and contain fewer instances of negativity, as our findings indicate, teachers are likely to find them more useful in terms of improving future teaching and learning.

## References

Anderson, H. M., Cain, J., & Bird, E. (2005). Online course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education, 69*(1), 34–43.

Ardalan, A., Ardalan, R., Coppage, S., & Crouch, W. (2007). A comparison of student feedback obtained through paper-based and web-based surveys of faculty teaching. *British Journal of Educational Technology, 38*(6), 1085–1101.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does on online delivery system influence student evaluations? *Journal of Economic Education, 37*(1), 21–38.

Ballantyne, C. S. (2004). *Online or on paper: An examination of the difference in response and respondents to a survey administered in two modes*. Paper presented at the annual international conference for the Australasian Evaluation Society, October 13–15, Adelaide, South Australia.

Bullock, C. D. (2003). Online collection of midterm student feedback. *New Directions for Teaching and Learning, 96*, 95–102.

Carini, R. M., Hayek, J. C., Kuh, G. D., Kennedy, J. M., & Ouimet, J. A. (2003). College students responses to web and paper surveys: Does mode matter? *Research in Higher Education, 44*(1), 1–19. doi:0361-0365/03/0200-0001/0.

Chang, T. S. (2005). The validity and reliability of student ratings: Comparison between paper-pencil and online survey. *Chinese Journal of Psychology, 47*(2), 113–125.

Dommeyer, C. J. (2006). The effect of evaluation location on peer evaluations. *Journal of Education for Business, 82*(1), 21–26.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29*(5), 611–623.

Donovan, J., Mader, C., & Shinsky, J. (2006). Constructive student feedback: Online vs. traditional course evaluations. *Journal of Interactive Online Learning, 5*(5), 283–296.

Ernst, D. (2006). *Student evaluations: A comparison of online vs. paper data collection*. Paper presented at the annual conference for Educause, October 9–12, Dallas, TX.

Gamliel, E., & Davidovitz, L. (2005). Online versus traditional teaching evaluation: Mode can matter. *Assessment & Evaluation in Higher Education, 30*, 581–592.

Handwerk, P. G., Carson, C., & Blackwell, K. M. (2000). *On-line vs. paper-and-pencil surveying of students: A case study*. Paper presented at the annual forum for the Association for Institutional Research, May 21–24, Cincinnati, OH.

Heath, N. M., Lawyer, S. R., & Rasmussen, E. B. (2007). Web-based versus paper-and-pencil course evaluations. *Teaching of Psychology, 34*(4), 259–261.

Hmieleski, K., & Champagne, M. V. (2000). Plugging into course evaluation. The technology source archives at the University of North Carolina. http://technologysource.org/article/plugging_in_to_course_evaluation/.

Johnson, T. D. (2003). Online student ratings: Will students respond? *New Directions for Teaching and Learning, 96*, 49–59.

Kasiar, J. B., Schroeder, S. L., & Holstad, S. G. (2002). Comparison of traditional and web-based course evaluation processes in a required, team-taught pharmacotherapy course. *American Journal of Pharmaceutical Education, 66*, 268–270.

Krupat, E., Pelletier, S. R., & Chernicky, D. W. (2011). The third year in the first person: Medical students report on their principal clinical year. *Academic Medicine, 86*(1), 90–97.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*(2), 221–232. doi:0361-0365/99/0400-0221.

Liegle, J., & McDonald, D. S. (2004). *Lessons learned from online vs. paper-based computer information students' evaluation system*. Paper presented at the annual conference for Information Systems Educators, November 3–7, Newport, RI.

Paolo, A. M., Bonaminio, G. A., Gibson, C., Partridge, T., & Kallail, K. (2000). Response rate comparisons of e-mail and mail-distributed student evaluations. *Teaching and Learning in Medicine, 12*(2), 81–84. doi:10.1207/S15328015TLM1202_4.

Ravelli, B. (2000). *Anonymous online teaching assessments: Preliminary findings*. Paper presented at the annual national conference for the American Association for Higher Education, June 14–18, Charlotte, NC.

Rice, L. S., & Van Duzer, E. V. (2005). A comparison of three modes of student ratings of teacher performance. Educational Resources Information Center (ERIC) #ED490478. http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/c1/b3.pdf.

Smither, J. W., Walker, A. G., & Yap, M. K. T. (2004). An examination of the equivalence of web-based versus paper-and-pencil upward feedback ratings: Rater- and ratee-level analyses. *Educational and Psychological Measurement, 64*, 40–61. doi:10.1177/0013164403258429.

Thorpe, S. W. (2002). *Online student evaluation of instruction: An investigation of non-response bias*. Paper presented at the annual forum for the Association for Institutional Research, June 2–5, in Toronto, Ontario.

Tomsic, M. L., Hendel, D. D., & Matross, R. P. (2000). *A World Wide Web response to student satisfaction surveys: Comparisons using paper and Internet formats*. Paper presented at the annual forum for the Association for Institutional Research, May 21–24, in Cincinnati, OH.

## Author Biographies

**William B. Burton** received the Ph.D. degree in environmental psychology from the City University of New York in 1998. From 1999 to 2010, he was an Assistant Professor in the Department of Family and Social Medicine at the Albert Einstein College of Medicine. From 2003 to the present he has been Associate Director of the Office of Educational Resources at the Albert Einstein College of Medicine. Since 2010, he has been an Associate Professor at the same institution. He is currently the Membership Chair of the Society of Directors of Research in Medical Education. His research interests are focused on the assessment of student performance and the use of technology in teaching and learning.

**Adele Civitano** joined the Office of Educational Resources at the Albert Einstein College of Medicine in 1994 and became the office's Data Administrator in 2008, after earning her B.S. degree from Iona College in 2007. Her major areas of responsibility have been data input and the administration of evaluations of the medical school curriculum.

**Penny Steiner-Grossman** earned a Master of Public Health degree in 1981 and a Doctor of Education degree in 1993, both from Columbia University, and joined the faculty of the Albert Einstein College of Medicine, also in 1993. As an associate professor in the Department of Family and Social Medicine, Dr. Grossman moved to the Office of Medical Education in 1997, first as Director, then as Assistant Dean in the Office of Educational Resources. She has expanded the office's responsibilities beyond curriculum evaluation to include targeted faculty development activities and assistance to faculty preparing for promotion. She also has served as co-coordinator of Einstein's LCME Institutional Self-Study.