Introduction

National research reveals that student evaluations are the most widely used source of information for assessing college course and instructor effectiveness (Seldin, 1993). If student development programs are to be truly perceived, not as “extracurricular” activities, but as co-curricular experiences that are equally important for promoting student learning as the academic curriculum, then it makes good sense to assess student development programs with methods of evaluation that are comparable to those used to evaluate college courses. Since student ratings represent the most frequently used and extensively researched strategy for evaluating courses in the curriculum, it would be useful to capitalize on this substantial body of students’ course evaluations, and explore how these findings may be adopted or adapted to improve the reliability, validity, and utility of co-curricular evaluation.

One major advantage of using student ratings to evaluate co-curricular programs is that their reliability and validity have probably received more empirical support than any other method of assessment; more than 1300 articles and books have been published that contain research on student ratings (Cashin, 1988, 1995). Despite perennial criticisms of student evaluations by some faculty and publication of some isolated studies that purport to refute their validity, when the results of all studies are viewed collectively and synthesized, they provide strong support for their reliability, validity, and utility of student ratings (Sixbury & Cashin, 1995). If a student-rating survey or questionnaire is well constructed and properly administered, it can be an effective and efficient vehicle for assessing student experiences, attitudes, perspectives, and self-reported outcomes, whether they may be curricular or co-curricular in nature.

Nevertheless, not all student-rating surveys are created equally. Their degree of reliability, validity, and utility can vary and may be influenced by (a) the instrument’s form and content, (b) the process in which the instrument is administered or delivered, and (c) how data generated by the instrument are analyzed, summarized, and reported. This article represents part one of a two-part series, and will focus on top-priority strategies for improving the quality of a student evaluation instrument’s content and form. In the second part of this article, issues relating to instrument administration, data analysis, and reporting results will be addressed.

Why Should We Conduct Student Evaluations of Co-Curricular Programming?

This is the first question that needs to be addressed in the assessment process, because its answer guides and drives decisions about the content, from, and intended use of the evaluation instrument. As Upcraft and Schuh (1996) note, “Perhaps the most important of all assessment principles is contained in the question, ‘Why are we doing this study’?” (p. 316). When answering this question, the classic distinction between “formative” and “summative” evaluation (Scriven, 1967) is an important one to keep in mind. If the
The purpose of assessment is *formative*, its goal is to obtain information that can be used as feedback to improve or fine-tune an existing program. Arguably, this is the most important purpose of program assessment, namely, to promote positive change in co-curricular programming that, ultimately, promote positive change in the students who experience them. As Scriven notes, “Student ratings are not only a valid, but often the only valid, way to get much of the information needed for most evaluations” (1988, p. 16).

In contrast to formative evaluation, *summative* evaluation is designed to “sum up” a program’s overall value or impact—for the purpose of making bottom-line decisions about its continuation or expansion. Co-curricular programs often lack the long history of acceptance and existence as ensconced courses in the traditional college curriculum, thus student evaluation of co-curricular programs may be intentionally designed and used for summative purposes, i.e., to generate the type of “value-added” evidence needed to ensure their long-term survival.

There is also a more tacit advantage of *student involvement* in program evaluation that is highly congruent with the mission of the student development profession. Namely, asking students to evaluate co-curricular experiences serves to validate their personal experiences, sending them the message that we genuinely care about their feelings, perceptions, and concerns. Student evaluation of co-curricular programming may be viewed as a *learner-centered* process that has the potential to engage and empower students, by providing them with an opportunity to gain a greater sense of involvement, ownership or control of their college experience.

**Recommendations on the Content and Form of an Evaluation Instrument**

1. Cluster individual items comprising the program-evaluation instrument into *categories* that represent important program *objectives or components*.

   Items comprising the evaluation instrument should include the *stated objectives* of the program because, in effect, these are the *intended outcomes* of the program. The instrument could be organized by grouping similar objectives into the same section, or by clustering together items relating to each of the following key components of programming: (a) program *planning and design* (e.g., questions pertaining to overall program organization or structure, and clarity of program objectives); (b) program *content*—what subject matter the program “covers” (e.g., topics and subtopics); and (c) program *process*—manner of program delivery (e.g., lecture-like presentations, large or small-group discussions, experiential learning activities).

   A major advantage of this organizational strategy is that the separate sections of categories can function as signposts or retrieval cues for the designers of the survey, ensuring that the items selected for inclusion in the instrument reflect a well-balanced sample of all key program dimensions (design, delivery, and content) that can affect the nature and quality of the students’ co-curricular learning experience. Another advantage of grouping items under section headings is that it can serve as a cue or signal to *students* that they should remain mindful of different dimensions to the program while they are evaluating it. This, in turn, will help students *discriminate* among important components of program effectiveness, thereby increasing the likelihood that they will assess them *independently*. 
Lastly, partitioning the instrument into separate sections reflecting separate program dimensions should help to reduce the risk of a general “halo effect”—i.e., the tendency for a student to complete the evaluation instrument by going right down the same column and filling in only “1s” or “5s” on all items, depending on whether they generally liked or disliked the program.

- **Avoid** evaluation items that are comprised of *compound sentences* that ask students to evaluate *two* different aspects of the program *simultaneously* (e.g., “The program content was both stimulating and useful.”)

  This practice forces respondents to give the same rating to both aspects of the program, even if they are more satisfied with one aspect than the other. For example, a student may feel the program content was “useful,” but not very “stimulating.”

- **Include** one or two *negatively worded* items that require students to reverse the rating scale (e.g., “I did *not* receive useful information in this program.”).

  Such items serve two purposes: (a) They encourage students to read and rate each item carefully, serving to reduce the frequency of “positive response-set” mistakes, which can occur when the respondent routinely goes straight down a rating column and fills in a uniformly high rating for all items (Arreola & Aleamoni, 1990). (b) They allow identification of evaluations forms that have not been completed carefully and may need to be deleted from the data analysis. For example, students who have responded thoughtlessly by filling in all positive or all negative ratings may be identified by their failure to reverse their response bias on the negative-worded item(s).

- **Provide a rating scale that allows five-to-seven choice points or response options.**

  A wider range of numerical options can result in mean (average) ratings for individual items that display a wider spread in absolute size or value. For instance, a 6-point scale may be superior to 4-point rating scales because the latter may yield mean ratings for separate items which vary so little in absolute size that the small mean differences between items may be dismissed as insignificant or inconsequential. For example, with a 4-option rating scale, a program might receive mean ratings for different items on the instrument that range from a low of 2.8 to a high of 3.3. Such a narrow range of differences in mean ratings can lead to the interpretation that these minuscule differences merely represent random “error variance” or students’ failure to respond in a discerning or discriminating manner.

  An expanded 6-point scale has the potential to produce larger mean differences across individual items, thus providing more discriminating data. In fact, research on student evaluations of courses does suggest that a rating scale with fewer than five choices tends to reduce the instrument’s ability to discriminate between satisfied and dissatisfied respondents, while a rating scale with more than seven choices does not add to the instrument’s discriminability (Cashin, 1990).

- **When asking students to rate their degree of involvement or satisfaction with a program, be sure to include a zero or “not used” option.**

  This response alternative allows a valid choice for those students who may have never experienced the event or program in question (Astin, 1991). It makes no sense to force respondents to rate an experience that they have never experienced. This would
only generate unnecessary data to process and invalid data to interpret. If forced to agree/disagree with a co-curricular event they have not experienced, what will probably happen is that respondents will choose the “disagree” or “strongly disagree” option, thus artificially reducing the overall average rating for that item and creating the impression that the item (co-curricular event) was less beneficial than it actually was. Also, students may become confused or frustrated when attempting to respond accurately to an instrument that requires them to rate something they did not experience. This may reduce the likelihood that students take the instrument seriously and complete it carefully and thoughtfully.

Lastly, failure to provide a “0” or "Not Used" (NU) option may result in loss of one useful piece of data about co-curricular programming. This option can provide data information about what programs were least/most likely to be attended, which constitutes information that may be useful for future program planning and improvement. Thus, inclusion of the 0/NU option can serve not only to eliminate invalid ratings from respondents who never experienced the programs being rated, it can also generate important and useful information about program participation rates.

Beneath each item to be rated, print the phrase, “reason for rating,” and provide a small space that students can use for providing written remarks relating to make to their rating.

Written comments can serve to clarify or elucidate numerical ratings and often provide the most useful feedback for program-improvement purposes, especially if such comments are specific. In most surveys, the individual items comprising the instrument are forced-choice rating items, which do not allow respondents any opportunity for open-ended comments that would enable them to explain why they gave that particular rating on that particular item. Allowing students to write comments with respect to each individual item (rather than restricting written remarks to the typical “general comments” section at the very end of the evaluation form), should also serve to increase the specificity of written feedback provided and, consequently, its utility for program improvement. This is also a good illustration of how qualitative data can be simultaneously gathered along with the usual quantitative data (numerical ratings) generated by student satisfaction/ratings surveys.

Include at least two global items on the evaluation instrument that assess summative (overall) program effectiveness or impact.

The following statements illustrate global items that are useful for summative-evaluation purposes:

(a) I would rate the overall quality of this program as: (poor → excellent).
    - Reason for this rating:

(b) I would rate the general usefulness of this program as: (very low → very high).
    - Reason for this rating:

(c) I would recommend this program to other students: (strongly agree → strongly disagree).
    - Reason for this rating:
Responses to these global items can provide an effective summary or “summative” snapshot of students’ overall assessment of the program, functioning like a final program “grade,” which can be readily cited in evaluation reports and used as evidence to support program effectiveness. Inclusion of global items on the evaluation instrument may not only provide a valid snapshot of general program quality, it also allows for the examination of relationships between students’ overall program rating and their ratings on individual items relating to specific program dimensions. Comparing students who gave the program very high global ratings with students who gave it very low overall ratings may provide an answer to the following question: For what particular items on the evaluation instrument do these two groups of students display the largest discrepancy in ratings? Those particular items with the largest discrepancies may be indicators of specific aspects or dimensions of the program that are most important to students, i.e., those that carry the most weight in influencing their overall perception of program quality and general level of program satisfaction. These particular program dimensions may then be targeted and used as focal points for program-improvement efforts.

1) Include an open-ended question asking for general comments about the program’s relative strengths and weaknesses, and how the latter may be improved or rectified. Open-ended questions allow students the freedom to provide a divergent range of responses, which can provide useful information about their overall reaction to the program as well as practical suggestions for program improvement. For example, in response to their participation in a residence-hall program experience, students could be asked to provide a written response to a question that asks them to “describe a positive change (if any) in your attitude or approach to the college experience that resulted from your participation in the program.” Or, they could be asked, “Was there anything missing from the program’s informational content or method of delivery that would have improved its quality?” Students’ written responses to these questions could be aggregated and their content analyzed to identify recurrent themes or common response categories.

One particular qualitative assessment method that can be used to summarize students’ written comments is category analysis, a procedure in which the reader engages in inductive data analysis, by identifying common themes that emerge among the comments as they are read, and then organizing these comments into categories (Lincoln & Guba, 1985). A tally of the number of written comments per category may also be kept and reported along with the identified categories. These category-specific frequency counts can then be used as quantitative data to help summarize and assess how representative or generalizable were individual students’ written comments (qualitative data).

To maximize the validity of category analysis, have two or more independent readers categorize the comments so that inter-reader (inter-classifier) agreement may be assessed. When data appear in the form of written comments, which are not amenable to objective and reliable machine-scored assessment, use of multiple assessors allows for the more subjective assessments of humans to be crosschecked, so that their reliability or consistency can be confirmed. Also, the reading and categorizing of comments should be conducted by someone who has no vested interest in the outcome of the program (an “outside” or “third-party” evaluator) so as to reduce the possibility of evaluator bias.

1) Give students the opportunity to suggest questions that they think should be included on the evaluation instrument.
This opportunity could be cued by a prompt at the end of the evaluation form, such as, “Suggested Questions for Future Evaluations.” This practice has three major advantages: (a) It may identify student perspectives and concerns that the evaluation form failed to address, (b) it shows respect for student input, and (c) it gives students some sense of control or ownership of the evaluation process.
References


STUDENT EVALUATION OF CO-CURRICULAR PROGRAMMING

Part II: Research-Based Strategies for Instrument Administration, Data Analysis, and Reporting Results

Joe Cuseo
Marymount College

Introduction
This article represents the second part of a two-part series. Part I focused on research-based strategies for improving the quality of a student evaluation instrument’s content and form. This part focuses on the process in which the evaluation instrument is administered or delivered, and how data generated by the instrument may be most effectively analyzed, summarized, and reported.

Recommendations on the Process of Instrument Administration
Research indicates that the wording of instructions read to students prior to distribution of evaluation forms can affect the nature of student ratings (Pasen et al., 1978). Effective verbal instructions serve to prepare or prime students for the important role they play in evaluating college programs, and provide them with a positive “mental set” toward the assessment process:

- To increase student enthusiasm for program evaluation and to improve the validity of the results obtained, include some or all of the following information in the instructions read to students prior to program evaluation.
- Explain to students why the evaluations are being conducted—for example, to help program facilitators improve the quality of their presentation and to improve the quality of the program offerings.
- Inform students about what will be done with their evaluations once they have completed them, assuring them that their evaluations will be carefully reviewed and taken seriously by the program director when planning future co-curricular experiences.
- Remind students that program evaluation is an opportunity for them to provide meaningful input that could improve the quality of the program for many future generations of students.
- Remind students that they should avoid the temptation to give uniformly high or uniformly low ratings on every item, depending on whether they generally liked or disliked the program or the program facilitator. Instead, remind students to respond to each item independently and honestly.
- Encourage students to provide written comments in order to clarify or justify their numerical ratings. Emphasize that specific comments are especially welcome because they often provide the most valuable feedback on program strengths and the most useful ideas for overcoming program weaknesses.

The key point being made in the foregoing series of recommendations is that we cannot assume that students already know how to be effective program evaluators; rather, we have to take some responsibility for preparing them to assume this role effectively.

Recommendations for Data Analysis, Summary, and Reporting
In addition to calculating the average (mean) student rating for individual items on the evaluation instrument, also calculate and report the **percentages of students choosing each rating option**.

This statistic will reveal how student responses were distributed across all response options, thus providing potentially useful information about the degree of consistency (consensus) or variation (disagreement) among student ratings for each item on the instrument.

To gain a **reference point** for interpreting student perceptions of a particular co-curricular event, compare their ratings with those provided for **other co-curricular events**.

This recommendation would be most effectively implemented by compiling an overall average (norm) of student ratings for programs in general, against which the rating of individual programs could be compared.

To gain a **longitudinal** perspective on **changes** in student attitudes or behaviors between the start and completion of a program, administer an evaluation instrument before programming begins, then re-administer the instrument after students have experienced the program.

This pre/post (before/after) design can be created by administering an evaluation instrument, or selected items therefrom, to students early in the first term (e.g., during orientation), so that these responses could function as a baseline (pre-test) against which their post-program (post-test) responses may be compared. To increase the likelihood that pre- to post-program changes in student attitudes or behavior can be attributed to the program, rather than to personal maturation over time or to the college experience in general, students’ pre- and post-program responses could be compared with the responses provided by other students (at the beginning and end of the same semester) who did not participate in the program.

By comparing program participants with students of similar age and level of college experience who do not experience the program, the non-participants may serve as an effective control group for the confounding effects of chronological maturation and general college experience. With this evaluation design, if differences are found between program participants and non-participants, a strong argument can be made that program participation had a **direct, causal** effect on student attitudes or behavior.

**Conclusion**

Although this two-part manuscript has focused primarily on the construction and administration of an instrument that provide quantitative data (student ratings), the most valid and comprehensive approach to program evaluation is to use a **balanced blend of both quantitative and qualitative methods**. While acknowledging that quantitative and qualitative research emerge from contrasting philosophical traditions and rest on very different epistemological assumptions (Smith & Heshusius, 1986), the position taken here is that the data generated by these two styles of inquiry provide **complementary** sources of evidence, with the limitations of one method being offset or counterbalanced by the advantages of the other. For instance, students’ written comments on surveys can be used to help interpret the average scores computed for numerical ratings, while the average rating scores can be used to counterbalance the tendency to draw overgeneralized
conclusions from several written comments that happen to be particularly poignant and powerful, but which are not representative of students as a whole.

Historically, surveys and questionnaires have not been considered to be qualitative research methods because they generate quantitative data (numerical ratings). However, written comments made by respondents to clarify their ratings do represent legitimate qualitative data, the content of which can be analyzed and classified systematically. Even the sheer number of positive or negative written responses students make beneath a specific item on a rating survey may itself serve as a measure of the importance or intensity of student feelings about the issue addressed by that item. As the National Orientation Directors Association (NODA) recommends for surveys of orientation programs, “Request individual written comments and provide space on the evaluation for these remarks. Participants with strong opinions about certain activities will state them if adequate space is provided. Summarize written comments in detail; consider indicating the number of times the same negative or positive comments were made” (Mullendore & Abraham, 1992, pp. 39-40).

Focus-group interviews represent another qualitative method that may be effectively used in conjunction with quantitative (ratings) instruments. Succinctly defined, a focus group is a small (6-12 person) group that meets with a trained moderator in a relaxed environment to discuss a selected topic or issue, with the goal of eliciting participants’ perceptions, attitudes, and ideas (Bers, 1989). In contrast to surveys or questionnaires, which solicit individual students’ numerical ratings and written comments, focus-group interviews solicit students’ verbal responses in a discussion-group setting. Verbal responses to orally posed questions often turn out to be more elaborate and extensive than written comments.

Focus-group interviews may be used in conjunction with student ratings’ instruments. For example, focus groups can be conducted as a follow-up to the collection of student ratings, for the purpose of gaining greater insight into the meaning of the survey’s quantitative results. Interview questions may be posed to a focus group, asking the group to offer their interpretation or explanation of ratings given to particular aspects of the program. Or, the order can be reversed, with focus groups conducted first to collect ideas that may later be used to develop specific items for inclusion on surveys or questionnaires.

Another qualitative method that may used to complement and augment quantitative ratings is the “minute paper,” which may be defined as a short (one minute or less) writing activity engaged in after completion of a learning experience—in response to a question posed at the end of a learning experience (e.g., “What was the most memorable or useful thing you learned today?”). The nature of the question can vary, depending on the nature and objective of the learning experience, but the one constant among all types of minute papers is that it provides immediate feedback on how students respond to a specific learning experience. Listed below is a sample of potential minute-paper questions that may be used immediately following co-curricular learning experiences:

>What do you think was the major purpose or objective of today’s program?
>What do you think was the most important point or central concept communicated in today’s program?
>What stands out in your mind or what do you recall most vividly about today’s class?
>What would you say was the most interesting idea or most useful strategy discussed in today’s program?
> What was the most *enlightening example* or most *powerful image* you experienced in today’s program?
> What was the most *convincing argument* (or counterargument) that you heard in today’s program?
> During today’s program, what idea(s) struck you as things you could or should immediately *put into practice*?
> Have you *personally experienced* any of the events that were discussed in today’s program?
> Did you see any *connections* between what was discussed in today’s program and what is being covered in any of your *course(s)*?
> What was the most *surprising* and/or *unexpected* idea expressed in today’s program?
> What do you think was the most *puzzling, confusing, or disturbing* idea that surfaced in today’s program?
> What *helped* and/or *hindered* your understanding of today’s program?
> What *questions remain unanswered* about the issue addressed in today’s program?

Using minute papers to assess specific co-curricular events—immediately after the students experience them—can provide assessment data with high validity because student memory of the experience is vivid and details can be readily recalled. In contrast, the standard, end-of-the-semester or end-of-the-year evaluation requires students to reconstruct and retrieve experiences from long-term memory that may date back to the beginning of the term. In addition to this program-assessment advantage of minute papers, they also encourage students to actively reflect upon and find personal meaning in the event they have just experienced, which should promote deeper learning and integration of the co-curricular experience with the academic experience of writing.

Lastly, collection of minute papers at the end of a co-curricular event can provide an immediate and accurate headcount of how many students participated in the experience, thus generating useful quantitative data in addition to the qualitative data provided by the students’ written comments. Such combined use of qualitative and quantitative data again reinforces the current recommendation that these approaches should be viewed as complementary rather than contradictory methods of evaluation.
References


