# 7

# Making Sense of Assessment Data

Imagine that faculty in your department collected assessment data using reflective essays, open-ended survey questions, embedded essay questions, or portfolios. You are sitting in your office, and in front of you is a set of these materials. Your task is to make sense of them. You want to summarize what has been learned, and you want to do this efficiently. Perhaps you are working alone, or maybe you are leading this effort and some faculty volunteers are waiting for your advice on how to proceed.

Collecting data is one thing, but making sense of them is something else. We want to use analytic techniques that are simple, direct, and effective. Faculty should focus on the task at hand. They are accustomed to pulling out red pens and covering student documents with wise advice. Analyzing data for program assessment is a much different task. The goals are to make sense of the information and to summarize it in a way that provides feedback on student mastery of learning objectives or that responds to questions that faculty want answered. In addition, the analysis should provide information that informs faculty as they decide how to respond to results. Focusing on these goals allows faculty to set down their red pens, tune out irrelevant information, and find efficient ways to complete the task--turning raw data into useful information.

Making sense of data takes time. Hastily thrown together data collection procedures may result in data that are not worth examining because results are ambiguous or unresponsive to the issues being addressed. Making sense of data is only part of a well-conceived assessment process. Data collection procedures should be pilot tested to ensure that they result in data that are worth analyzing.

This chapter describes two common ways to approach assessment data: content analysis and the use of rubrics. Content analysis allows us to summarize a communication. For example, we use content analysis to make sense of responses to open-ended survey questions. Rubrics allow us to categorize the quality of products, such as the quality of critical thinking in student essays. This chapter emphasizes basic approaches that faculty from any discipline can apply. While specialists may employ more sophisticated techniques, these simple strategies can yield information that contributes to meaningful, manageable, and sustainable assessment programs.

## CONTENT ANALYSIS

Content analysis involves making sense of the material being reviewed. Most often this material will be verbal, such as interview or survey responses, but we might also want to describe themes that occur in products, such as works of art or political cartoons. Although one person can conduct a content analysis, results might have more credibility if multiple reviewers are involved. Idiosyncratic interpretations and preconceptions are less likely to influence results, and inter-rater agreement can be examined.

Content analysis should be approached with an open mind and a willingness to "hear" what respondents are saying. While everyone enjoys hearing praise, most faculty don't celebrate when they receive criticism. Reviewers should be aware that common responses to criticism are to defend or counterattack. They should avoid "spinning" results to support their personal opinions or denigrating those whose opinions differ from their own.

When approaching this task, faculty should begin with a clear understanding of the goals associated with the review. If they are examining responses to a survey question on the quality of advising, their content analysis should focus on what students report about advising. If they are reviewing student reflections on the impact of campus experiences on their ability to interact within a multicultural environment, the focus should be on identifying the types of experiences that students mention and the consequences of those experiences. Faculty should ask, "Why am I reviewing these materials? What do we want to learn?" The answers will direct the focus of the content analysis.

When conducting the content analysis, reviewers want to summarize common themes and the extent of consensus concerning those themes. Generally, they develop a written coding scheme that describes how individual responses will be categorized. Then they code data by recording instances of each category so they can develop accurate summaries in the reports. Coding results for small studies can be easily accomplished by hand, and a statistical program (e.g., SPSS) or spreadsheet (e.g., Excel) is often used to help organize results for larger studies.

Sometimes coding categories are predetermined. At other times categories emerge as materials are reviewed. For example, if students are asked about the quality of advising, three categories might be predetermined: negative, neutral, and positive. Readers could review each response and categorize it appropriately, and this probably could be done fairly rapidly.

If the content analysis stops at this point, however, readers have thrown away much potentially useful information. In fact, if faculty only wanted to learn how often students categorize advising experiences as negative, neutral, or positive, it would have been easier to use a simple item requesting this rating. Students were asked to write essays because more information was desired, and content analysis involves objectively summarizing this information. As reviewers analyze the content, themes emerge, and these are integrated into the coding scheme. For example, readers might notice that students who report negative experiences often complain about the lack of advisor availability in the evenings, or they may notice that students who report positive experiences often praise the peer advising center. This is information worth including in the summary because it adds to the formative validity of the assessment process.

Once the coding scheme is finalized, it can be applied to the documents. If multiple readers are involved, it is typical to ask them to independently review a few products so checks for inter-rater reliability can be conducted. If disagreements are rare, the process can continue, but if disagreements are common, readers should identify the reasons for the discrepancies and agree on how they should be resolved. In this way, they clarify ambiguities in the coding scheme.

Usually basic student information, such as gender and class level, is coded, too. This allows the reviewer to describe characteristics of the sample in the report, and the reviewer might decide to examine subgroup differences. For example, some themes might occur more often among women or among international students, and this added information may be important for using the results effectively. Try your hand at coding student responses in Figure 7.1.

Once data are coded, the summary should be written with the audience in mind. Program faculty want concise reports that address the questions they want answered. Exemplars should be provided for all themes that are discussed, so their meanings are clear; and sometimes reports provide complete lists of all comments related to each theme so readers can independently judge their meaning. If only exemplars are provided, conclusions should be quantified. For example, "Fifty percent of the students who report negative experiences requested evening advising hours." This quantification (50%) provides a summary that is much less ambiguous than using terms like "many" and "most." When faculty read that "Many students requested evening advising hours," it is not clear if "many" means three, 15, or 50 out of the 100 students who responded. Opinions that are expressed more frequently are more likely to be given serious attention.

Reports should include sufficient detail so that readers can interpret numerical results accurately. For example, if you report student suggestions for improving program advising, you might state that 80% of the students suggest one solution and 40% suggest an alternative solution. The alert reader will notice that these percentages sum to over 100%, and this might be confusing unless you have noted that students were allowed to offer multiple suggestions. In addition, the report should clarify how many responses were considered. For example, if only a fraction of the students offered suggestions, the report might say, "Forty percent of students offered suggestions for improving advising. Among these, 50% recommended that advisors be available in the evening." This is less ambiguous than simply reporting that 20% of students recommended evening advising. If counts are given, rather than percentages, they

FIGURE 7.1
CONTENT ANALYSIS CODING PRACTICE

Here is a coding scheme for student responses to this question: "Faculty want students in our program to develop search skills to identify and secure print and non-print materials relevant to topics in political science. Describe one assignment or activity that you experienced in a political science course that helped you strengthen these search skills."

Faculty asked this question because they wanted to hear students' perspective on which types of assignments and activities most effectively help them develop search skills. Based on reviewing a sample of student responses, faculty developed the following coding scheme:

1) Identification number (assign each student a number, beginning with "1")
2) Student class level (1 = Freshman, 2 = Sophomore, 3 = Junior, 4 = Senior, 5 = Grad)
3) Overall response (0 = no response/question was unanswered, 1 = provided a usable response, 2 = stated/implied that search skills were not strengthened in a P.S. course, 3 = response was not relevant to the question/could not be coded)
4) Positive mention of a structured assignment that leads students step-by-step through the search process (0 = No, 1 = Yes)
5) Positive mention of an assignment that required students to use or develop search skills, but no mention of step-by-step guidance (category 4) or group activity (category 6) (0 = No, 1 = Yes)
6) Positive mention of an assignment or activity that required working collaboratively with peers (0 = No, 1 = Yes)
7) Positive mention of a how-to-search lecture in a course (0 = No, 1 = Yes)
8) Positive mention of an in-class activity related to search skills (0 = No, 1 = Yes)
9) Positive mention of the use of online learning materials (0 = No, 1 = Yes)
10) Positive mention of personal assistance by a faculty member (0 = No, 1 = Yes)
11) Positive mention of personal assistance by a TA (0 = No, 1 = Yes)
12) Positive mention of personal assistance by a librarian (0 = No, 1 = Yes)

Use this scheme to code the responses below. Create a coding sheet with 12 columns (one column for each coding category) and five rows (one row for each student), and enter the appropriate code in each column.

*(continued on page 136)*

1) (a senior) "I think I learned the most when Professor Ruiz required team projects. We had to do a thorough literature review, and most of us didn't know how to start, but my team had a couple of students who were really good at computers, and we all learned a lot. We also really appreciated the help of Cindy, one of the reference librarians. She was always able to answer our questions when we got confused."

2) (a graduate student) "I thought I was pretty good at finding materials until I took Dr. Web's class. She made us break the search into steps: Clarify the research question, identify key terms, etc. This really helped me learn how to structure searches, and now I use these skills every time I need to find some information."

3) (a junior) "I had to do lit reviews for papers in several PS courses and learned how to do searches then. Sometimes I got help from the TA."

4) (a senior) "The first time I had to do a major lit review, I was totally lost. I went to Dr. Meese's office, and he showed me how to use the campus search system. Thanks, Dr. M.!"

5) (a sophomore) "I'm not very good at this. I kinda wander around the library looking at journals and books in the PS section until I find something I can use."

Here is the coding:

```
1 4 1 0 0 1 0 0 0 0 0 0 1
2 5 1 1 0 0 0 0 0 0 0 0 0
3 3 1 0 1 0 0 0 0 0 0 1 0
4 4 1 0 1 0 0 0 0 0 1 0 0
5 2 2 0 0 0 0 0 0 0 0 0 0
```

should be presented in a context that clarifies how many products were reviewed. For example, the fact that 20 students praised the peer advising center has different meaning if they were in a sample of 30 students or in a sample of 500 students.

Sprinkling reports with direct quotes from students–using the student's voice–makes the reports more powerful because the results clearly represent what students are telling us. Instead of wondering if the person doing the summary misrepresented student opinions, faculty are able to assess this themselves. If this is overdone, busy faculty will ignore quotations, but well-chosen quotations can have high impact. Although the following story is not from higher education, it illustrates this point in a memorable way. A friend worked under contract to examine the experiences of spouses of enlisted military personnel as part of a larger effort to

examine operations at a military base. He delivered a long written report sparkling with tables of numbers, and he was invited to address high-ranking officers to summarize his findings. He used PowerPoint slides of his tables to support his main conclusions, then he reported on some focus groups and included a story shared by a young military wife whose baby almost died because ambulance drivers could not find a place to park near their on-base housing. The commanding officer immediately ordered his staff to find a solution, and it was quickly implemented. According to my friend, that one story had more impact on base functioning than all the rest of the report. People like to hear stories, and they often respond to them in different ways than they respond to the same types of information presented in neatly tabulated columns. This experience also points out the need for reviewers to attend to individual responses.

Sometimes an individual's response is unique and does not fit the coding scheme, but it is so important that it deserves recognition in the report. For example, if a disabled student in a focus group expresses concern about the safety of a campus bus ramp, this information should be provided to relevant campus personnel. Reports sometimes include lists of all comments because readers may decide that some of the points are of sufficient importance that they should be addressed, regardless of how often they occur.

The report also should provide other information that affects interpretation. For example, it should summarize characteristics of the sample. Readers will want to know if results are *generalizable*, that is, if they are likely to accurately represent all student opinions. Were the products from a representative sample of students or were they from a special sample, such as honors students or new transfer students? Findings from special samples may not accurately describe the opinions or experiences of other groups. Readers should have information on how the data were collected and how student privacy and confidentiality were addressed. Students might respond differently if they are interviewed by faculty who determine their grades, and they may withhold personal information if they believe responses can be tied to them. Content analysis allows you to summarize what was found, but the entire process should be designed to provide valid information.

## SCORING RUBRICS

Assessment workshop participants often report that the segment on rubrics had the greatest impact on their assessment activities, as well as their teaching. Scoring rubrics make the impossible manageable.

Scoring rubrics are explicit schemes for classifying products or behaviors into categories that are steps along a continuum. These steps generally range from "unacceptable" to "exemplary," and the number of intermediate categories varies with the need to discriminate among other performance levels. Rubrics can be used to classify virtually any product or behavior, such as essays, research reports, portfolios, works of art, recitals, oral presentations, performances, and group activities. Judgments can include self-assessments by students or judgments can be made by others, such as faculty, other students, fieldwork supervisors, and external reviewers. Rubrics are versatile tools. They can be used to provide formative feedback to students, to grade students, and to assess programs. A well-designed rubric should allow evaluators to efficiently focus on specific learning objectives while reviewing complex student products, such as theses, without getting bogged down in irrelevant details.

There are two major types of scoring rubrics: *holistic* and *analytic*. Holistic rubrics describe how one global, holistic judgment is made. Analytic rubrics involve making a series of judgments, each assessing a characteristic of the product being evaluated. Figure 7.2 is a holistic rubric, and Figure 7.3 is an analytic rubric. As you read these rubrics, think about how you might adapt them for your own use.

Notice that the descriptions in the holistic rubric combine various dimensions, such as focus, development, and the use of language. Holistic judgments are made based on reviewing the entire product in sufficient detail so that the classification can be made with confidence. This allows readers to scan complicated products, such as portfolios, paying careful attention only to segments directly related to making this discrimination. Professional readers, such as readers who review essays for test publishers, can make holistic judgments concerning student writing quickly and with substantial reliability and validity. Faculty, with some practice, can develop similar skills.

The rubric in Figure 7.3 engages students in the assessment by asking them to rate peers who developed team projects together. Sharing

FIGURE 7.2
HOLISTIC RUBRIC FOR ASSESSING STUDENT ESSAYS

| | |
|---|---|
| Inadequate | The essay has at least one serious weakness. It may be unfocused, underdeveloped, or rambling. Problems with the use of language seriously interfere with the reader's ability to understand what is being communicated. |
| Developing competence | The essay may be somewhat unfocused, underdeveloped, or rambling, but it does have some coherence. Problems with the use of language occasionally interfere with the reader's ability to understand what is being communicated. |
| Acceptable | The essay is generally focused and contains some development of ideas, but the discussion may be simplistic or repetitive. The language lacks syntactic complexity and may contain occasional grammatical errors, but the reader is able to understand what is being communicated. |
| Sophisticated | The essay is focused and clearly organized, and it shows depth of development. The language is precise and shows syntactic variety, and ideas are clearly communicated to the reader. |

FIGURE 7.3
ANALYTIC RUBRIC FOR PEER ASSESSMENT OF TEAM PROJECT MEMBERS

| | Below Expectation | Good | Exceptional |
|---|---|---|---|
| Project contributions | Made few substantive contributions to the team's final product | Contributed a "fair share" of substance to the team's final product | Contributed considerable substance to the team's final product |
| Leadership | Rarely or never exercised leadership | Accepted a "fair share" of leadership responsibilities | Routinely provided excellent leadership |
| Collaboration | Undermined group discussions or often failed to participate | Respected others' opinions and contributed to the group's discussion | Respected others' opinions and made major contributions to the group's discussion |

this rubric with students early in the term gives them guidance on faculty expectations, and sometimes such rubrics are developed in consultation with the students themselves. Programs often have interpersonal skills among their learning objectives, but many faculty are not comfortable trying to teach these skills because they have no special training in this area. A rubric such as this could help faculty structure expectations and provide students with appropriate feedback, and it could help faculty align the curriculum with this objective and collect data for program assessment. Rubrics have many uses.

Analytic rubrics can be thought of as collections of holistic rubrics that assess different aspects of the product being assessed. These aspects should be linked to learning objectives, and they may be developed through primary trait analysis (Walvoord & Anderson, 1998). Primary trait analysis is the process of determining the primary traits (criteria) that faculty use to evaluate products. The rubric in Figure 7.3 assesses three primary traits: project contributions, leadership, and collaboration.

The previous examples describe categories, but rubrics also can be used to generate scores. Figure 7.4 is a rubric that could be used for grading oral presentations. The rubric includes possible score ranges for each cell, and total scores can range from 0 to 30. If a rubric like this is embedded in lower-division and upper-division courses, faculty could track student growth in these skills. Using rubrics as embedded assessment tools should not interfere with faculty control of their classes or grading practices. As was mentioned in Chapter 1, grading requires more precise measurement than assessment. Although faculty may agree to share a common rubric, each might add additional course-specific traits and each might assign points to the categories in different ways. The scores on this rubric could be used for grading, but the categories that were used might provide sufficient detail for program assessment.

Rubrics can be created for examining almost any product, including portfolios. Figure 7.5 is a generic rubric for assessing learning objectives using portfolios. Notice the emphasis on evidence, rather than conjecture. This rubric might provide sufficient guidance for faculty reviewers, or they may need more detailed criteria for some learning objectives. Perhaps a generic rubric would be useful as a starting point, when faculty first review portfolios, and discussion will lead to the development of a more detailed rubric. If faculty always wait until they perfect procedures, little assessment would occur!

FIGURE 7.4
ANALYTIC RUBRIC FOR GRADING
ORAL PRESENTATIONS

| | Below Expectation | Satisfactory | Exemplary | Score |
|---|---|---|---|---|
| Organization | No apparent organization. Evidence is not used to support assertions.<br>(0-2) | The presentation has a focus and provides some evidence that supports conclusions.<br>(3-5) | The presentation is carefully organized and provides convincing evidence to support conclusions.<br>(6-8) | |
| Content | The content is inaccurate or overly general. Listeners are unlikely to learn anything or may be misled.<br>(0-2) | The content is generally accurate, but incomplete. Listeners may learn some isolated facts, but they are unlikely to gain new insights about the topic.<br>(5-7) | The content is accurate and complete. Listeners are likely to gain new insights about the topic.<br>(10-13) | |
| Style | The speaker appears anxious and uncomfortable, and reads notes, rather than speaks. Listeners are largely ignored.<br>(0-2) | The speaker is generally relaxed and comfortable, but too often relies on notes. Listeners are sometimes ignored or misunderstood.<br>(3-6) | The speaker is relaxed and comfortable, speaks without undue reliance on notes, and interacts effectively with listeners.<br>(7-9) | |
| Total Score | | | | |

Program assessment using rubrics can be conducted in a number of ways. Faculty can use rubrics in classes and aggregate the data across sections, faculty can independently assess student products and then aggregate results, or faculty can participate in group readings in which they

FIGURE 7.5
GENERIC RUBRIC FOR ASSESSING PORTFOLIOS

| | Unacceptable: Evidence that the student has mastered this objective is not provided, unconvincing, or very incomplete. | Marginal: Evidence that the student has mastered this objective is provided, but it is weak or incomplete. | Acceptable: Evidence shows that the student has generally attained this objective. | Exceptional: Evidence demonstrates that the student has mastered this objective at a high level. |
|---|---|---|---|---|
| Learning Objective 1 | | | | |
| Learning Objective 2 | | | | |
| Learning Objective 3 | | | | |

review student products together and discuss what they found. Group readings can be very effective. Faculty often develop deeper understanding of the learning objectives and refine the rubric as they work, and they can immediately discuss the implications of results with the evidence fresh in their minds. Field work supervisors or community professionals also may be invited to assess student work using rubrics, and sometimes students are invited to do self or peer assessments.

Rubrics should be pilot tested, and evaluators should be "*normed*" or "*calibrated*" before they apply them. This is often accomplished through training using sample products that are established exemplars of different levels of performance. If group readings are conducted, they generally begin with everyone independently reviewing a few products so that differences can be discussed and resolved. If two evaluators apply the rubric to each product, inter-rater reliability can be examined. Once the data are collected, faculty discuss results to identify program strengths and areas of concern, then they close the loop by identifying and making changes to improve student learning.

Sometimes faculty readers require special training to overcome old habits. Faculty who are accustomed to grading on a curve might have difficulty making judgments based on the criteria stated in a rubric. They need to know that it is not essential to use all rubric levels and that they should not be concerned about how often each category is used. Some learning objectives are easier to achieve than others or are better aligned with the curriculum, so it is possible to find extensive use of higher categories for some objectives and lower categories for other objectives. The assessment should uncover student strengths, as well as limitations, and conclusions should be based on an objective application of the criteria stated in the rubric. Readers also should be told to be careful to rate each category in analytic rubrics separately, avoiding a *halo effect*. A halo effect occurs when judgments are influenced by each other. For example, if students are asked to use the peer assessment rubric in Figure 7.3, they might be tempted to classify an uncooperative student into the lowest category on all three dimensions, even if only one is appropriate.

Although this book does not focus on grading, rubrics often are embedded in courses because they are so useful for this purpose. As was mentioned in Chapter 1, learner-centered teaching often uses collaborative and cooperative learning models to help students develop, and these models work best when students are not penalized for working together. Rubrics allow faculty to assign grades based on the achievement of objectives rather than on how well students perform compared to each other, so their use encourages students to help each other learn. Figure 7.6 offers some suggestions for using rubrics in courses. Faculty can get double duty out of rubrics by using one rubric for grading *and* program assessment.

Holistic rubrics can be created using seven steps:

1) Identify what you are assessing (e.g., critical thinking).
2) Identify the characteristics of what you are assessing (e.g., appropriate use of evidence, recognition of logical fallacies).
3) Describe the best work you could expect using these characteristics. This describes the top category.
4) Describe the worst acceptable product using these characteristics. This describes the lowest acceptable category.
5) Describe an unacceptable product. This describes the lowest category.
6) Develop descriptions of intermediate-level products and assign them to intermediate categories. You might decide to develop a

### FIGURE 7.6
### SUGGESTIONS FOR USING RUBRICS IN COURSES

- Hand out the grading rubric with the assignment so students will know your expectations and how they'll be graded. This should help students master your learning objectives by guiding their work in appropriate directions.
- Use a rubric for grading student work and return the rubric with the grading on it. Faculty save time writing extensive comments; they just circle or highlight relevant segments of the rubric. Some faculty include room for additional comments on the rubric page, either within each section or at the end.
- Develop a rubric with your students for an assignment or group project. Students can then monitor themselves and their peers using agreed-upon criteria that they helped develop. Many faculty find that students create higher standards for themselves than faculty would impose on them.
- Have students apply your rubric to some sample products before they create their own. Faculty report that students are quite accurate when doing this, and this process should help them evaluate their own products as they are being developed. The ability to evaluate, edit, and improve draft documents is an important skill.
- Have students exchange paper drafts and give peer feedback using the rubric; then give students a few days before the final drafts are turned in to you. You might also require that they turn in the draft and scored rubric with their final paper.
- Have students self-assess their products using the grading rubric and hand in the self-assessment with the product; then faculty and students can compare self- and faculty-generated evaluations.

scale with five levels (e.g., unacceptable, marginal, acceptable, competent, outstanding), three levels (e.g., novice, competent, exemplary), or any other set that is meaningful.

7) Ask colleagues who were not involved in the rubric's development to apply it to some products or behaviors and revise as needed to eliminate ambiguities.

These steps can be repeated to generate an analytic rubric. The first time you try to create a rubric is the hardest, but, like any skill, it becomes easier with practice.

Developing a rubric takes time, and sometimes it is easier to adapt one that already exists. Fortunately, many examples are available. Walvoord and Anderson's (1998) book, *Effective Grading*, and Wiggins's (1998) book, *Educative Assessment*, include many rubrics, and the information literacy rubrics developed by the Colorado Department of Education (http://www.cde.state.co.us/cdelib/download/pdf/inforubr.pdf) cover a wide range of important learning objectives. Many others are available on the web, and links to them are available on the California State University Student Learning Outcomes site (http://www.calstate.edu/acadaff/sloa/links/rubrics.shtml).

As you look for rubrics to adapt, keep an open mind. Look for general formats, relevant concepts, and key words that you can incorporate into your work. Don't restrict yourself to rubrics created for your discipline. Faculty in an economics program might find that a rubric designed for assessing students' mastery of basketball dribbling is almost perfect. They just have to replace "dribble the ball" with "apply Keynesian economics." In addition, don't ignore the potential of rubrics developed for primary and secondary schools. Many can be easily adapted to higher education by strengthening requirements. Your goal when examining others' rubrics is not to find one that you can apply directly to your work; look for ideas that help you customize a rubric to match your needs.

## INTER-RATER RELIABILITY

We generally want to verify that scores based on subjective judgments have inter-rater reliability. Inter-rater reliability indicates the extent of agreement among different reviewers. Without this information, we might wonder if summaries are accurate depictions of what was examined. For example, if we knew that a content analysis was done by a faculty member who is skeptical about the value of virtual courses, we might be suspicious of this person's summary of student opinions on this topic. If other raters agree, though, we have more confidence in the conclusions. This section describes some simple ways to examine inter-rater reliability. As with content analysis and rubrics, basic strategies that could be applied by anyone are described. Experts may conduct more sophisticated analyses, but usually these are not necessary if the only purpose is to verify that results are reasonably reliable.

When we do content analyses, subjective judgments are involved in two steps: establishing the themes to be coded and doing the coding. Probably the best way to have confidence in the selected themes is to have at least two people independently develop them, compare notes, and come to agreement. This process can be summarized in the report, and readers could be told how difficult it was to determine the final categories. To examine inter-rater reliability of coding judgments, we need to have at least two readers independently apply the coding scheme to a set of documents so we can compare their judgments. Content analyses frequently involve dichotomous decisions: Is the variable being coded present or not? In this case, the percentage of agreements is a good indicator of inter-rater reliability. Do the raters agree 10% of the time or is this percentage closer to 90%? Ten percent is too low; most would agree that 90% is acceptable, although even higher would be better. Figure 7.7 demonstrates this calculation. The finding of 70% agreement in this example suggests that the criteria for identifying the presence of this theme are too loosely defined and further work is needed. Coding schemes should be pilot tested and checked for inter-rater reliability *before* faculty invest time in coding the entire data set.

An alternative approach is to have pairs of raters work together on all decisions, collaborating as they go to resolve discrepancies. Although judgments probably are made more slowly, raters generally are very confident about their accuracy. To examine this process, you might ask pairs of raters to keep track of how often they disagreed in their original judgments. If this percentage is low, you probably could reduce the workload by having only one rater analyze each product. You also, of course, could have two pairs of raters analyze the same set of documents and calculate the percentage of their agreements. In this way, you treat each pair of raters as if they were one person.

You may recognize that these calculations have a flaw. You could get perfect inter-rater agreement if both raters always use one rating category, that is, always code that the theme is absent or that the theme is present. This could happen when themes occur in almost no products or in almost every product. The purpose of the assessment is to uncover findings that will inform decision-making. It is important, of course, to report information that has considerable consensus, and sometimes it is useful to report a rare, but significant event (such as the story of the baby at the military base). The percentage-of-agreement summary statistic is most reasonable when many opportunities for disagreement occur, and this happens when we examine themes that are common but not universal.

FIGURE 7.7
CALCULATING THE PERCENTAGE OF AGREEMENTS

Here are coding results for two raters who made dichotomous decisions about the presence of a theme in ten products. A "0" indicates that the theme is absent; a "1" indicates that the theme is present.

| | |
|---|---|
| 0 | 0 |
| 0 | 0 |
| 0 | 1 |
| 0 | 1 |
| 1 | 0 |
| 1 | 1 |
| 1 | 1 |
| 1 | 1 |
| 1 | 1 |
| 1 | 1 |

Overall, the raters agreed seven times. They agreed two times that the theme was absent, and they agreed five times that the theme was present. They disagreed three times. They agreed on seven out of the ten judgments, so their percentage of agreement is 70%.

If judgments involve more than two categories, as is common in the use of scoring rubrics, the situation is a bit more complex. Two approaches are possible. First, one could assign scores to categories (e.g., "1" for the lowest of five categories and "5" for the highest), and then calculate the correlation between two raters who have reviewed a set of materials. This might work fine, but would fail if there is insufficient variability. Correlations summarize how much variability in one variable can be explained by the other. If most ratings use a single category, there may be insufficient variability to explain, and the correlation will be small–not because raters disagreed, but because the products being evaluated were too similar. An alternative is to ask how many of the ratings were identical, one point apart, two points apart, etc. For example, if a scoring rubric has five categories, judgments can be, at most, four points apart (one rater classifies the product in the lowest category and the other classifies it in the highest category). Say raters are identical on 80% of their ratings, within one point on 18% of their ratings, and within two points on the remaining 2% of the ratings. Most probably would find this acceptable. As with dichotomous judgments, raters could work in pairs, resolving discrepancies as they go, and you could examine the percentage of discrepancies requiring discussion or the degree of agreement between independent pairs of raters.

Figure 7.8 illustrates these calculations. In this example, raters gave identical ratings 75% of the time, differed by one point 20% of the time, and differed by two points 5% of the time, that is, ratings were within one point 95% of the time. Sometimes the distinction between selected categories is particularly important. For example, ratings of "1" and "2" might identify products that fall below minimum standards, and ratings of "3" and "4" might identify products that meet minimum standards. With only one exception (the product categorized as a "2" by one rater and as a "4" by the other rater), the raters agreed every time if a product met minimum expectations. Based on this analysis, the raters might agree to revisit this one product and come to consensus on it. The correlation for these data is .79. This summarizes the relationship in a different way, and the .79 is a moderately high correlation. Perfect agreement would generate a correlation of 1.0, and strong inter-rater reliability would be indicated by correlations in the .90s. These correlations must be interpreted loosely because they are affected by the amount of variability in scores. When using a 4-point rubric, variation might be low, which could reduce the size of the maximum possible correlation that could be observed.

## MANAGING THE REVIEW

Sometimes one person does a content analysis or develops and applies a rubric, and other times these are team efforts. One of the most valuable aspects of program assessment is that it provides a forum for faculty discussion of student learning. At some point, all or most faculty in the program should find out what was learned and should discuss the implications for program functioning. It is not essential that everyone become involved in a content analysis or use a rubric, but the participation of more than one person can be useful. Their collective wisdom might make the process more reliable and valid, and each might learn from the experience.

Content analysis is often done by one person. Sometimes this person is a neutral outsider, such as an assessment consultant, but sometimes the task requires insider knowledge. In this case, a well-supervised undergraduate or graduate student may be able to do the job well. Having a large group create the coding scheme probably is a waste of time because most content analyses do not require this amount of collaboration, but inviting a second person to independently develop themes or pilot test the coding makes good sense. If coding criteria do not require high-level professional judgments, students may be able to do the coding with appropriate supervision. Whichever way it is done, it is important that the summary accurately reflects what was being analyzed and provides information with formative value. Faculty time is valuable, and their time may be best spent reflecting on a well-written summary and deciding what follow-up action is required.

### FIGURE 7.8
### SUMMARIZING THE DIFFERENCES BETWEEN RATINGS

Here are coding results for two raters who classify 20 products into four categories (labeled from "1" to "4").

| | |
|---|---|
| 1 | 1 |
| 1 | 1 |
| 1 | 2 |
| 2 | 1 |
| 2 | 2 |
| 2 | 2 |
| 2 | 2 |
| 2 | 4 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 3 | 3 |
| 4 | 3 |
| 4 | 3 |
| 4 | 4 |
| 4 | 4 |
| 4 | 4 |

In reviewing these 20 documents, the raters gave identical ratings 15 times (75%), disagreed by one point four times (20%), and disagreed by two points one time (5%). The correlation between the two sets of ratings is .79.

Rubrics must be clearly written so they can provide reliable, valid information. If rubrics will be embedded within courses, involvement by faculty who offer these courses makes good sense, although most will serve on a consultative basis rather than as the rubrics' primary author. In addition, if the rubric is designed to assess a program learning objective, all program faculty have some interest in its development. This process should be inclusive, with open invitations for input by all relevant faculty.

Faculty have developed a number of strategies for applying rubrics to products. Sometimes faculty work alone. If assessments are embedded within courses, faculty may apply rubrics to student work within their own courses and send their results to someone who tabulates the data. Faculty also can independently apply rubrics to collected products. For example, program faculty may divide a collection of portfolios, with each faculty member assessing a portion of them. This can work well, but validity would be threatened if individuals apply standards differently. Training and norming generally are necessary to ensure the integrity of this process.

Rubrics often are applied by groups who come together for this purpose. For example, program faculty may agree to devote a half day to this task. Before the meeting, the rubric must be finalized, the products to be examined must be collected, and a data collection procedure must be developed. The reading should begin with an orientation to clarify the nature of the task and to norm the reviewers. Figure 7.9 offers some suggestions to the person who facilitates this orientation.

Group ratings can be collected in several ways. If the reliability of the rubric is known to be high, it may be reasonable to have only one reader analyze each document, but it generally is preferable to use two readers. Inter-rater reliability can be examined and errors can more easily be identified and corrected. When two readers work independently, the second reader may be allowed to peek at the first rater's judgments. Readers often are curious about other's opinions, and no harm is done if the first rater's scores are hidden until after the second opinions have been recorded. Sometimes results are monitored as they are turned in, and documents are given to a third reader when necessary to resolve discrepancies. For example, the facilitator may send any document that has a scorer difference of more than one point to a third reader who deter-

### FIGURE 7.9
### SCORING RUBRIC GROUP ORIENTATION

1) Invite readers who offer and control the curriculum and who have the capacity to make informed judgments about student learning.

2) Describe the purpose for the review, stressing how it fits into program assessment plans. Explain that the purpose is to assess the program, not individual students or faculty, and describe ethical guidelines, including respect for confidentiality and privacy.

3) Describe the nature of the products that will be reviewed, briefly summarizing how they were obtained.

4) Describe the scoring rubric and its categories. Explain how it was developed.

5) Explain that readers should rate each dimension of an analytic rubric separately, and they should apply the criteria without concern for how often each category is used.

6) Give each reviewer a copy of several student products that are exemplars of different levels of performance. Include, if possible, a weak product, an intermediate-level product, and a strong product, and you also might include a product that appears to be particularly difficult to judge. Ask each volunteer to independently apply the rubric to each of these products, and show them how to record their ratings.

7) Once everyone is done, collect everyone's ratings and display them so everyone can see the degree of agreement. This is often done on a blackboard, with each person in turn announcing his or her ratings as they are entered on the board. Alternatively, the facilitator could ask raters to raise their hands when their rating category is announced, making the extent of agreement very clear to everyone and making it very easy to identify raters who routinely give unusually high or low ratings.

8) Guide the group in a discussion of their ratings. There will be differences, and this discussion is important to establish standards. Attempt to reach consensus on the most appropriate rating for each of the products being examined by inviting people who gave different ratings to explain their judgments. Usually consensus is possible, but sometimes a split decision is developed. For example, the group may agree that a product is a "3-4" split because it has elements of both categories. Expect more discussion time if you include a hard-to-rate example, but its consideration might save time and prevent frustration during the subsequent review. You might allow the group to revise the rubric to clarify its use, but avoid allowing the group to drift away from the learning objective being assessed.

9) Once the group is comfortable with the recording form and the rubric, distribute the products and begin the data collection.

mines which rating is more accurate. Sometimes readers work in pairs, independently rating each document, then jointly resolving all disagreements. They may be asked to discuss only the ratings that differ by some amount, such as at least two units. When two raters disagree, faculty must decide which rating will be used in the analysis, or they may decide to use both (e.g., Allen, Moe, & Roberts, 2000; Noel, 2001a). Whatever the decision, the project report should document how data were generated.

Faculty often are in the best position to apply scoring rubrics, and they may be more likely to take results seriously if they conducted the analysis themselves. Others, however, may be invited to participate. Sometimes student self assessments have credibility, especially if based on objectively defined rubrics. Peers who have firsthand knowledge of relevant information also can apply rubrics, such as the peer rubric in Figure 7.3. Others may have expertise or firsthand knowledge, such as graduate students, fieldwork supervisors, alumni, or other professionals. Program faculty also might consider working with colleagues from institutions with similar missions. They could assess student work together, perhaps examining products from both campuses simultaneously, or they could review each other's student work. This might create opportunities to develop fresh ideas about curriculum and pedagogy, and it might lead to other collaborations, such as faculty exchanges, co-authored publications, or jointly sponsored theses. If they share rubrics, they also may be able to provide benchmark data that would be useful when interpreting findings. The need to provide training and to examine inter-rater agreement is important when people who did not participate in rubric development will assess student work. Rubrics are simple tools. They have many uses, and their value is limited only by the imaginations of those who use them.