

A BLUNT* INSTRUMENT FOR USE BY LOW-LITERATE PARTICIPANTS IN SUMMATIVE AND FORMATIVE EVALUATIONS OF ADULT EDUCATION AND DEVELOPMENT PROGRAMS

ADRIAN BLUNT

University of Saskatchewan, Canada

This article reports the development of an evaluation technique that uses adjective checklists as instruments for use by low-literate and marginally literate adults to provide assessments of their learning experiences in adult education and development programs. Vignettes of activities from three stages of the development process depict the range of groups and learning events used to refine the technique and to establish the reliability and validity of instrument scores. Recommended procedures are outlined for application of the technique, including the selection of adjectives, obtaining responses, summarizing results, and calculating scores and instrument statistics.

Keywords: *adult basic education evaluation; evaluation instrument design; life-skills program evaluation; marginalized learner program evaluation; low-literate learner program evaluation*

This project has its origins in the 1970s in work with minimally literate adult learners. At that time, in my various roles as a community developer, instructor, and adult education program manager, I wished to respect the rights of these learners to participate directly in the formal evaluations of their education and development programs. Literate adult education participants commonly provide individual assessments of their learning experiences that when combined with the responses of others, can provide a program manager and stakeholders with important

ADRIAN BLUNT is a professor in adult and continuing education at the University of Saskatchewan, Canada (e-mail: adrian.blunt@usask.ca).

*blunt 5. abrupt of speech or manner; plain spoken; curt; without delicacy; direct; unceremonious. 1599 Shakespeare, Henry V, IV, vii, 185 "By his blunt bearing he will keepe his word." (*Oxford English Dictionary*, 1970, p. 948)

ADULT EDUCATION QUARTERLY, Vol. 55 No. 2, February 2005 129-149

DOI: 10.1177/0741713604272375

© 2005 American Association for Adult and Continuing Education

information. Yet the commonly used pencil and paper evaluation instruments required relatively high levels of literacy to read and understand their completion instructions and individual scale items. The learners I worked with were unable to complete such forms accurately, and their resistance to such tasks was high. A second area of concern was that the most popular type of instrument required respondents to apply cognitive rules for the completion of psychometric scales, a task that required experience working with abstract measurement concepts that these learners lacked.

An additional methodological concern was that scaling techniques, such as the widely used Likert-type scale, present respondents with precise boundaries between numerical categories. Yet what is being measured, the concepts we think with, are not themselves so precisely delimited. The concepts that are useful and meaningful to us for thought and communication purposes must have some degree of imprecision or generalization for communication to occur. Otherwise, we would have to invent a new concept and expressions for each new experience that we encounter (Bollinger, 1980). When data collectors impose inappropriate numerical operations assumptions on participants and a rigidity on affective responses, they fall prey to the "precision fallacy" (Polkinghorne, 1984). For all of these reasons, the available instruments that required high-level numeracy and literacy skills were, in my opinion, inappropriate.

The groups I worked with included First Nations' youth and young adults, chronic unemployed, early school-leavers, mildly mentally challenged, English-language-challenged immigrants, and street people from the urban core. Frequently, the programs I managed had been designed for learners who were members of several of these categories simultaneously. Typically, the programs were oriented toward low-level adult basic education (ABE), life skills, and social survival skills. What I needed, to facilitate meaningful participation in program evaluations by these learners, was an instrument that required minimal levels of literacy and numeracy.

Purpose of the Study

This article reports the long-term development of a technique for the construction of program evaluation instruments for use by low-literate and marginally literate adults. The technique sought had to generate learner assessment instruments for use in a broad range of ABE and development program contexts, formative and summative evaluations, and empirical research to yield scores with acceptable levels of reliability and avoid the sin of precision fallacy (Polkinghorne, 1984).

THEORETICAL FRAMEWORK

The conceptualization of this project is grounded broadly in semiotics (Hervey, 1982; Kristeva, 1986; Noth, 1990), the discipline that according to its founder,

Ferdinand de Saussure, “tries to make sense of things” (Culler, 1976, p. 113), and in semantics and pragmatics in particular. Semantics as the study of the meaning of words and sentences and pragmatics as the study of language use extend the study of language from “formal object through to language as practical activity” (Forrester, 1996, p. 45). Words are tools whose function is to support communication, both oral and written (Anglin, 1970); when written or spoken, a word conveys meaning, a meaning developed and sustained within a particular social context. Words work when they mean the same thing to different speakers and writers of a language. Communication with understanding is possible in a particular social context, when for all the speakers and listeners, writers and readers, the relation between a word and its referent and the relation between the oral and written word are the same or very similar (Anglin, 1970; Bollinger, 1980). The referent is the physical or imagined object of a word (Bollinger, 1980). All languages have a term similar to the English term *meaning*, and all language users actively seek consensual understandings of meanings to communicate with others. Opportunities to communicate occur as soon as consensual understandings are achieved. Recognizing and using words as referents is an essential early step in the development of communications skills. Words act as signs to convey meanings and are open to interpretation (Morris, 1963). However, the signs do not have meanings in themselves but rather, only in relation to agreement about their use (Frawley, 1992; Thibault, 1997). Beyond this simple understanding of word use is a more complex aspect of language, as meaning has been demonstrated to be, at least, bimodal. Meaning has an attachment to a word as a signifier of an object and a second attachment conveying the value or significance of the object to the communicator. The meaning of any sign “is an indissoluble association between signifier and signified” (Forrester, 1996, p. 39; see Figure 1).

To the evaluator, differences in the role of words are highly important. When an adult educator uses the word *lecture* in a professional sense, the referent may be an instructional technique rather than a method (Verner, 1962).¹ However, to the evaluator, the word’s value is enhanced by the distinction between the signifier and the signification when the word is used by a learner who says, “The instructor gave me a lecture.”

The purpose of evaluative words is to give or deny value to objects that they are applied to, including persons, experiences, and social phenomena. Evaluative words are used to praise, blame, or commend by persons who wish to share with others their valuing, deprecation, or criticism of selected objects. Language, therefore, is directly associated with psychological processes and an important subject for evaluation research. As early as the 1920s, Hartshorne and May (1930) developed an adjective checklist to measure four types of personal conduct. Shortly afterwards, Allport and Odbert (1936) enumerated 17,953 English words in a study to establish personality trait names. In the 1940s, Cattell (1946) used factor analysis of adjectives to establish the primary source traits of personality. One widely used instrument established through experimental work with adjectives is the Adjective

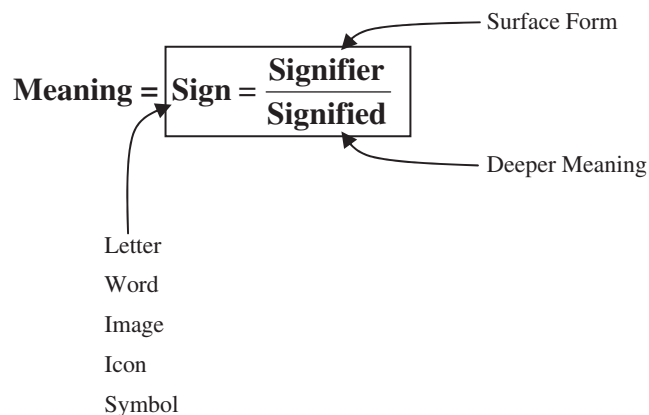


Figure 1. The Components of Meaning

SOURCE: Forrester (1996, p. 39). Adapted with permission from Sage Publications.

Check List (Gough & Heilbrun, 1965). It uses 300 adjectives to generate scores on 24 scales and indices including, for example, Defensiveness, Self-Control, Achievement, Autonomy, and Deference. Perhaps the best-known scaling technique, and the most methodologically sophisticated, to have emerged from this broad line of inquiry into the value of underlying meanings of words is the Semantic Differential Scale (Osgood, Suci, & Tannenbaum, 1957). Osgood et al. (1957) used factor analysis to determine whether signs fell into groups (factors) that shared underlying connotative meanings in common. Their analyses confirmed that certain words, such as *good* and *true*, clustered together to form an evaluative factor. Other words, including *strong* and *hard*, clustered to form a potency factor, and others, such as *fast* and *active*, clustered as an activity factor. The Semantic Differential Scale studies empirically confirm that groups of words perform different functions and persons' responses to those words hold unique content for evaluation and social science studies.

Whether language users prefer to use the written or oral word makes no difference to a word's evaluative utility. The challenge for instrument development in this project was to ensure that the relation between a word and its referent remained constant as the user shifted from speaking a word to using visual recognition to select the same word for use in a literacy context. Furthermore, the goal was to construct a means to express learners' mutual beliefs. To study the meaning of adult education experiences to learners it is necessary to study the words selected by learners.

METHOD

The evaluation technique reported here was developed by trial and error at intervals during a period of almost three decades. The first trials were conducted in the 1970s with low-level ABE participants. Broader applications with a wider range of groups and programs occurred in the 1980s. In the 1990s, development focused on psychometric issues regarding quantification of scale scores, reliability and validity, and use of data in applied research projects. Representative vignettes of activities from the three development periods are presented below.

First Trials

The first effort to develop an instrument was with a group of male offenders in a correctional center in northern British Columbia. I first consulted the Adjective Check List (Gough & Heilbrun, 1965) and selected from its 300 adjectives 10 that I thought the participants might use to describe their instructor. I selected another 10 words to describe the course itself—words that I thought participants might use in a conversation about their satisfaction with the course. Among the adjectives selected to assess the instructor, I chose *knowledgeable*, *thoughtful*, *boring*, and *confusing*. Of the 10 words, 5 held positive connotations and 5 held negative connotations. Among those I selected to obtain responses to the course itself were *exciting*, *valuable*, *chaotic*, and *shallow*. Again, I chose 5 with positive and 5 with negative connotations. I avoided polar pairs of adjectives such as *good* and *bad*. To administer the checklist, I typed the word *Teacher*, underlined and centered, on the top line of a sheet of paper. Below, I typed the 10 words, left justified, in random order, in lower-case letters, as a list, 1 word per line. I repeated the procedure to create a second list for *Course*. Next, I made photocopies, one for each learner, and one overhead projector transparency of each of the two lists.

In a group meeting, I distributed the first list to the participants, placed the transparency copy on the overhead projector, and pointing at each word on the list in turn, asked the group to “circle [the word on the list] if you think it gives your opinion about your teacher. Leave it alone if you disagree with using this word to describe your teacher.” I repeated the procedure with the second transparency, asking the class to “circle the word if it says what you think about the course. Leave it alone if you disagree with using this word to describe your course.” After collecting the response sheets, I tallied the number of words circled, placed the words in numerical frequency order, retyped the lists, printed new overhead transparencies, and returned to class 1 hour later to discuss the results.

With time, I learned that low-literate students have more severe problems with the grammatical construction of sentences than with recognizing individual printed words on a list and that respondents understood the meanings of the great majority of adjectives I selected. On those occasions when a problem with word meanings

did occur, it was resolved through discussion. Learners valued using the assessment instruments, and I was able to use the results to make important program decisions about, for example, rehiring instructors and amending instructional activities and course objectives.

Wider Applications Trials

The adjective checklist evaluation technique was refined and tested with a wide range of groups and programs. With one group of urban First Nations youth, I explained the need to evaluate their instructor's performance and gather information to help make a decision about her reappointment. Rather than provide a list, I asked the group to brainstorm the selection of adjectives. On this occasion, I agreed to use the learners' preferred words, and a list of 20 words, arrived at through student voting, was compiled to be the assessment instrument. The list included several vernacular words and phrases, including *tight-ass*, *earth-mother*, *ball-buster*, and *wonder-woman*. From that trial, I learned to respect learners' interests and to obtain the most useful data possible by asking participants to choose words most appropriate for their intended use. In future trials, with this and similar groups, I asked participants to "use words and language that are OK with your grandmother, and remember they are going to be read and understood by the people who approve programs for you." As Forrester (1996) has stated, "Until we learn to speak in appropriate ways our utterances can not really be taken as 'meaningful' in an everyday sense" (p. 56).

The range of evaluative objects was extended to include program content and processes. Learners rarely encountered difficulties when a discussion of the specific evaluation object preceded the data collection. For example, when an adjective checklist was used to gather participants' assessments of the teaching processes used in a workshop or course, the objective of the evaluation was visually presented as *Teaching Process* and orally defined as

the class discussions intended to help you to understand others' points of view. The occasions you asked questions and talked about topics that were important to you. How well the lectures were presented. All these activities are part of the teaching process.

The technique also demonstrated its value in formative assessments. At one weekend workshop, participants were not fully engaged in the group's activities. Tensions were evident among participants, and some participants and resource persons were not working well together. Using the now standard procedures, we asked participants to complete a feedback form consisting of adjectives selected to assess the learning climate. When the results were compiled and graphically displayed, the frequencies produced a list of alternating positive and negative adjectives. To make clear the extent of the differences in experiences being reported, we high-

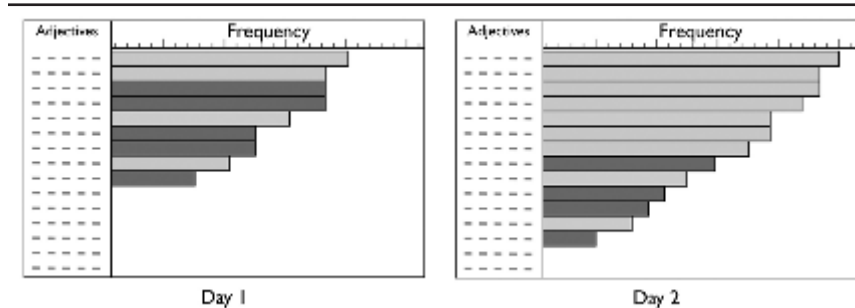


Figure 2. Simulated Frequency Distributions of Formative Adjective Checklist Responses for 2-Day Workshop

lighted the positive items in yellow and the negative items in red. Next, the items were reviewed one word at a time. For example, participants were asked, “OK, you say the process is tight, what would make it loose? Most of you think the session was dull, what will make tomorrow’s session interesting?” The results were discussed with the resource persons in the same format as the participants had reviewed it, and the participants’ comments were used to plan changes to the process and planned activities. At the close of the second day, we repeated the process and compared the graphic summaries (see Figure 2). Shifts in the magnitude and order of item frequencies demonstrate that the instructional processes had changed for the better. Discussions triggered by the instrument’s use became an opportunity to reinforce participants’ ownership of the workshop goals, processes, and outcomes and to guide the facilitator’s and resource persons’ implementation decisions.

In a formal college-based ABE program, groups of learners shared a number of instructors, and evaluation data were used to compare instructors on learners’ estimates of their teaching styles and effectiveness. With time, the technique contributed to building instructors’ teaching portfolios and program files for quality improvement and development purposes. The data contributed to decisions made about instructor retention, teaching assignments, program objectives, instructional activities, and learning resources. From these and similar summative and formative applications, I learned that the technique can be used effectively in short periods of time, with minimal resources, without direct supervision, and to shape the processes and dynamics of learning events. In these situations, the technique supports learner democracy and empowerment and meets the criteria to be considered a popular education technique. Trials were also conducted satisfactorily with several highly literate groups including college instructors, civil servants attending professional development workshops, mature students in a diploma-level business education program, adult education graduate students, and learners enrolled in public, general interest adult education courses.

Psychometric Properties

During the 1990s, development focused on psychometrics: item analysis, individual and group scoring, and score reliability and validity. Checklists generate nominal-level scale data, therefore, nonparametric tests are appropriate for item analysis and test statistics. The advantage of nonparametric statistics is that they are used with small sample sizes, are simple to calculate, and require minimal training for users to understand and calculate score statistics.²

Item pool. The first item pool was developed from the Adjective Check List (Gough & Heilbrun, 1965). Through trial and error applications and consultations with expert panels, a pool of approximately 280 items has been compiled (see Appendix). Items have been added to and deleted from the pool based on their demonstrated efficacy following item analyses described later in this section. The items in the Appendix are intended to be a resource for instrument construction. When additional local-context and learner-appropriate items are to be included in an instrument, it is important that they be subjected to item analysis and matched with pool items for selection likelihood (this process is discussed below).

Content validity. One indicator of the content validity of a psychometric instrument is the extent of agreement among knowledgeable judges that each individual item is appropriate for its intended use. Two item-selection workshops were conducted to determine whether a consensus existed among adult educators with regard to evaluative adjectives for use in program assessments. An expert panel of 9 graduate students and 2 faculty members at the University of Saskatchewan was asked to

imagine an ABE course for downtown residents where students are working at or below the Grade 6 level. List 10 adjectives that would describe this course as one of the best you have observed and 10 that describe it as one of the worst you have ever observed.

Next, the panel was asked to

think of the instructional practices, learning resources, and learning activities in this course and list 10 adjectives that you might use to assess the instruction of this ABE course positively and 10 you might use to assess the instruction negatively.

A third activity required the panel to “list 10 adjectives that identify the strengths of the instructor and a further 10 that identify the instructor’s weaknesses.” When all the items had been compiled, the panel voted on the retention of each to form an ABE program item pool. At the University of Regina, 16 college instructors enrolled in an undergraduate-level course on postsecondary education were asked to review approximately 350 items and to categorize by forced choice each item as “a

word I might possibly use, or definitely would *not* use, to describe an adult education program.” The process was repeated at 1-week intervals with Adult Education Instructor and Adult Education Learning Climate as the assessment objects. The response frequencies were tabulated, and any item categorized by 8 or more respondents as “use” or “not use” was either included or deleted from the pool.

Sixteen adult education graduate students and 4 faculty members at the University of British Columbia were asked to

imagine a typical, public, general interest adult education program. Write a brief, one-sentence description of the program and below it, list five adjectives that you might use in a conversation to describe the program in positive terms and five adjectives to describe the activity in negative terms.

Next, the panel was asked to

think about the same adult education activity in terms of its instructional processes, resources, and learning activities. List five adjectives that you might use in a conversation to assess the instructional activities positively and five that you might use to assess the instruction negatively.

Of the adjectives selected, more than three quarters matched those in the existing item pool. In a second activity, each panel member was given a list of 20 randomly selected items from the pool and asked to categorize each through forced choice as “a word I have used on some occasion, or might use, to describe an adult education program, instructor, or learning climate or a word I have never used, and would not use for such a purpose.” When 10 or more panel members agreed on the categorization of an item, it was either retained or deleted from the item pool.

Item analysis. Item analysis is an essential process in psychometrics to achieve acceptable levels of score reliability and validity. A good achievement test item is one likely to be selected by half of all respondents, that is, one with a difficulty level of 0.5 (Gronlund, 1988). A similar approach to ascertaining the likelihood that an adjective will be selected (accepted) was adopted in this project. Items that are rarely or never selected by respondents serve no useful purpose and bias the calculation of individual scores and the coefficient of assessment (discussed later in this section). Only items that fall within a selection range of 0.20 to 0.80 during several trials on similar objects have been retained in the item pool. The recommended procedure for confirming the acceptance level of checklist items is the nonparametric binomial test (see Siegel, 1956, pp. 36-42). Each item has only two possible response categories, a respondent will either select it or not. The binomial test answers the question, What is the probability that the distribution of selections and rejections occurred by chance?

There is a possibility that the selection of adjectives for inclusion in a checklist will be biased. Positive words, for example, may be more likely to be selected than

negative words because of relative differences in their positive and negative connotative meanings. For example, respondents may find it easier to select *good* than *awful* or to accept *helpful* than *useless*. To ensure a balance in the positive and negative items, for each positive item selected, a negative item with a broadly comparable acceptability level ought to be included. When it is important to obtain individual scores and a coefficient of assessment (described in the following section), it is recommended that the number of items be oversampled and following item analysis, weak pairs of positive and negative items be deleted prior to calculating individual and total group scores.

Instrument scores. A person's score is obtained by calculating the algebraic total of the words selected. Each adjective with a positive connotation has a value of +1, and each adjective with a negative connotation has a value of -1. The positive adjectives selected are summed and the sum of the negative adjectives is subtracted to give a total score:

$$\text{Individual Score} = \Sigma \{+n\} - \{-n\}.$$

When the number of negative adjectives selected exceeds the number of positive adjectives, a negative total score results. In a 30-item instrument, the possible range of scores is 31: from +15 to -15, including 0. An overall assessment can be calculated and reported as a coefficient with values ranging from 1.0 (perfectly positive) to -1.0 (perfectly negative). The coefficient of assessment is calculated by the formula

$$C = \frac{\Sigma \{+n\} - \{-n\}}{1/2 K \times Ss}.$$

The numerator in the equation is the difference between the total number of positive adjectives selected (+n) by the total group and the total number of negative adjectives (-n) selected. The denominator is one half of the total number of items on the checklist (K) multiplied by the total number of respondents in the group (Ss). For a group of 23 learners, who complete a 30-item checklist by selecting 253 positive and 42 negative adjectives, the coefficient of assessment is

$$C = \frac{253 - 42}{15 \times 23}$$

$$C = \frac{211}{345}$$

$$C = 0.61.$$

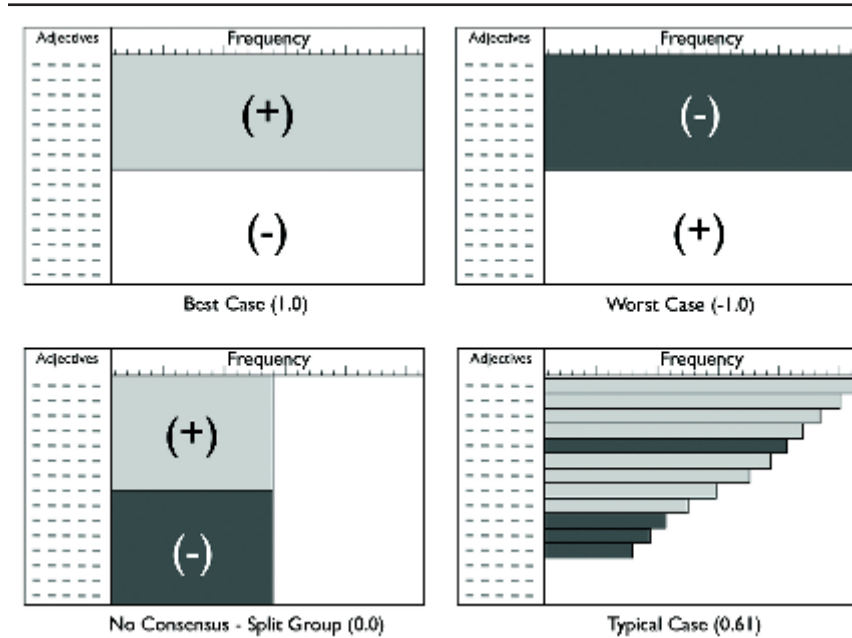


Figure 3. Simulated Frequency Distributions for Four Coefficients of Assessment

Figure 3 presents a graphic representation of coefficient values for four different cases.

Because the number of items selected by each respondent affects the calculation of individual scores and the coefficient of assessment, it is very important, when these statistics are needed, to ensure that the data collection instructions include a request that each respondent select a minimum of approximately half the words on the checklist.

Score reliability. Internal consistency reliability estimates have been computed on many individual instruments. Kuder-Richardson (KR21) estimates, a conservative estimate of unidimensionality, have spanned from lows in the range of 0.73 to 0.79 to highs in the 0.83 to 0.91 range. Spearman Brown split-half reliability coefficients have frequently been obtained in the 0.82 to 0.92 range.

Test-retest estimates of reliability have also been calculated on several occasions with groups other than ABE participants. The use of these participants was convenient to gather data for test-retest analysis and for trials of the instrument with a broader range of respondents. The following results are typical. A group of 21 adult education graduate students at the University of Saskatchewan attending a 3-hour

workshop on use of the *Publication Manual of the American Psychological Association* (American Psychological Association, 1983), completed a 30-item evaluation instrument at the close of the workshop and completed the same instrument again at intervals of 2 and 4 weeks. The objects of the evaluation were the workshop content and instructional activities. A total of 20 students completed the 2-week interval retest ($R = 0.89$) and 17 students, including 1 who did not complete the 2-week retest, completed the 4-week interval retest ($R = 0.83$).

Another test-retest trial was conducted at the University of Regina with 27 postsecondary education instructors studying learning achievement test construction. A shorter, 24-item instrument was used to assess midcourse satisfaction with the course's content and assignments, with the retest occurring after a 2-week interval. The stability of scores for 26 respondents on the 24-item form during 2 weeks was 0.86. A third reliability test of the form, on this occasion reduced to 20 items, was conducted with a group of 23 learners in a basic-level ABE course. The first application occurred in the last class immediately before the Easter break, and the retest was given 6 days later in the first class of the day. The test-retest correlation coefficient obtained for the 20-item checklist for 19 respondents during 6 days was 0.82.

One additional reason for testing the reliability of instruments with a range of items from 30 to 20 was to assess the effect of the number of items on KR21 coefficients. KR21 reliability estimates are known to be sensitive to the number of items in an instrument: the more items, the higher the KR21 coefficient obtained (Niemi, Carmines, & McIver, 1986). It was concluded that a range of 20 to 30 items usually yields satisfactory KR21 coefficients; no substantial improvements are achieved by increasing the number of items beyond 30, and lower than desirable KR21 estimates may occur with fewer than 16 items.

Score validity. The validity of checklist scores has been studied on a number of occasions. In one trial of the adjective checklist, three groups of learners' assessments of their one common instructor were collected together with their responses on the Purdue Rating Scale of Instruction (Remmers & Weisbrodt, 1965). Correlations between scores for each of the three groups on the two instruments ranged from 0.69 to 0.74. Furthermore, a six-person program advisory committee also completed the adjective checklist, and a comparison of the frequencies confirmed a strong consensus on the instructor's qualities and practices.

Faced with a list of words, a respondent might simply check (select) words at random and, thus, provide invalid scores. Although the words on a checklist are always presented in random order, their selection when the assessment is done as intended will reflect a consensual understanding of the object of assessment in terms of the adjectives selected. The majority of respondents typically agree on their interpretation of an object, and the sequence of items selected, therefore, emerges in a pattern that is not random. By testing the order of the words selected, it is possible to determine whether a respondent's set of selections is random and invalid or logi-

cally sequenced and valid. The nonparametric one-sample runs test (see Siegel, 1956, pp. 52-58) can be used to consider the number of runs in a list of adjectives and to indicate the probability that a random selection of words has occurred. A run is a succession of selected words that is followed or preceded by words that were not selected. For example, the following order of selected (Yes) and nonselected (No) words may occur:

Yes Yes No Yes No No No Yes No Yes Yes No No Yes Yes No Yes Yes No No

This sample of word selections begins with a run of 2 Yes's, followed by a run of 1 No, followed by a run of 1 Yes, followed by a run of 3 No's. There is a total of 12 runs in the list of 20 selections. If few runs occur, some process is at work to influence the selections. Similarly, if many runs occur, the selections will also be nonrandom. The one-sample runs test calculates the probability that any one respondent's selections are random (invalid) and need to be deleted from the assessment. The unique value of the runs test is that it is based on the order of selections and can detect randomness among scores that cannot be detected by a test of score frequencies.

What is most important about the value of the technique, from the perspective of validity, is that within an adult education institution, when evaluation instruments are used consistently and regularly over time, the data collected can be used to demonstrably improve overall program quality. For example, data from the assessments can be used to inform instructor-rehiring decisions; to identify areas of course weakness, including the quality of resource materials; to improve instructional activities and the learning climate; and to build learner-client relationships. In this context, confirming the validity of the technique is not a complex process. What is required are regular and systematic comparisons between the checklist profiles and other program indicators including external reviews, learner follow-up surveys, program advisory committee reports, and other empirical instruments with known reliability and validity to determine whether scores from the adjectival checklists and overall program subjective assessment coefficients predict and/or confirm program outcomes.

RECOMMENDED PROCEDURES

The selection of adjectives (items) for an instrument begins with a statement of the purpose of the intended assessment. Although certain adjectives may be selected for a number of evaluative objects, such as *the instructor* or *group participation*, in each case a claim to the content validity of the word, as an item in the instrument, is made based on its selection for a particular evaluative purpose. Having stated the purpose of the assessment (e.g., "To obtain learners' assessments of the quality of their learning experiences in the voter-registration workshop") the next step is to state, for oral communication, to the learners the context for their selection

of an adjective (item). A number of statements are needed to serve this purpose. They need to be written and read aloud at the time of data collection to build a consensual understanding and avoid misunderstandings that extemporaneous statements might induce. For example,

Think about all the learning activities in the workshop—the talk about who can and can't vote, the use of the map to show where registered voters live, the task to locate where on the map class members live, the research questions about what the map told the class, the small group discussion, and the guest speaker's talk about how to register to vote in an election. Think about all those activities as the "whole workshop learning experience." Do you have any questions about what I mean when I say the "whole workshop learning experience?" OK, now imagine your most serious friend from school missed the workshop because she was sick. Later she asks you, "How did the workshop go? Was it any good? Did I miss a really good session?" You answer, using the list of words, "The class was really . . ." "I think it was . . ."

Depending on the overall purpose of the adult education method (again Verner's [1962] definition of *method* is intended), its explicit and implicit intended learner outcomes, there are a number of instrument construction process options. In a life-skills program where participatory and popular adult education instructional techniques are commonly used, adjectives may be selected through a brainstorming activity. After discussing the purpose of the assessment, learners may propose adjectives from their personal vocabularies and arrive at a final list through some consensus-seeking means. The brainstorming can be supplemented by the instructor contributing adjectives for consideration by the group. Program advisory committee members, resource persons, and instructors may also suggest words. Good results have been obtained when instructors, prior to the start of a course or program, are asked, "What words would you like to hear students use to describe their experience in your course? What words would you wish them not to use?" When an instructor selects his or her own items, the assessment results carry a weight and meaningfulness for that person beyond that generated when items are selected by others. The list of items in the Appendix is a resource for program stakeholders. Item pools are best developed over time following item analysis procedures and trial-and-error applications in local contexts.

Experience indicates that extreme evaluative words, such as *lousy* and *ridiculous*, and culturally and politically insensitive words and high-affect loaded words, such as *racist* and *dumb*, are best avoided. Polar pairs of adjectives ought not to be selected. With each single word (item) the opposite meaning to the word is partially addressed by the respondent's decision to select the word or not. If the word *good* is selected, logically its opposite *bad* is rejected, so the inclusion of *bad* in the list is a wasted item. If *good* is not selected, *not-good* is implied, and although *bad* is not available for selection, other words congruent with and sharing some of *bad*'s connotative meaning can be available for selection. Technical words from other contexts, such as sport or popular entertainment and culture, are usually not very

helpful; for example, instructor assessments are not well informed by words such as *hot*, *cool*, or *menacing*. Occasionally, adjectival phrases and hyphenated words may be included where they offer a unique meaning to be introduced into the assessment, are in common usage in learners' discourses, or where stakeholders prefer them, for example, *planned-poorly*, *difficult to understand*, and *beyond my level*. Everyday language of the school, workplace, home, or community is rich with meaningful evaluative adjectives and phrases.

As explained earlier, an equal number of positive and negative words are needed, and a range of 20 to 30 is recommended to ensure the reliability of scores. Validity also increases with a larger number of items as the web of connotative meanings among items is expanded. The number of items should be as representative as possible of all the meaningful subjective judgments that can be made about the object and small enough in number to be as efficient as possible in practice. Completion time is also an important factor in deciding on the number of items, as respondent fatigue or attention spans may influence respondent performance.

Present the instrument to respondents using an overhead projector or computer-assisted projection. The object of the assessment—for example, the learning experiences at the voter-registration workshop—should be printed in lower-case letters at the top of the page, centered, and underlined. Below, left justified, the adjectives should be listed vertically, in random order, one word per line, in a suitably sized font, with line spacing appropriate for clear and easy perception. A photocopy of the instrument projected on the screen is given to each respondent. It is a good idea, particularly with lower literate groups, to mask the list and expose one word at a time on the screen. After reading a statement on the purpose of the assessment, the completion instructions can be read aloud:

If you think the word describes the workshop, mark it with your highlighter (or circle the word). Mark (or circle) each word that you would use to answer questions such as, How did the workshop go? What was it like? If you do not think you would use the word, simply leave it alone. Choose words that say what you think about the workshop.

As each word is revealed on the screen, particularly at the start of the list, the evaluator reads the word aloud, as in a spelling test, and provides a sentence for completion, "Interesting—Was the workshop interesting?" It is important to write the statement of the context and the instructions for completion in language appropriate for reading aloud. Oral language requires a sentence structure, timing, and plain-language vocabulary that is congruent with the task respondents will perform. Experience suggests that oral rehearsal of the instructions helps to ensure a problem-free data collection process.

The final instruction to respondents addresses the question of how many word selections need to be made. The calculation of individual scores and the coefficient of assessment take into account the total number of items selected. Their magnitudes are reduced when the number of words selected falls below 50%. Instruc-

tions, therefore, need to emphasize that the instrument works best when approximately half the words are selected:

Please check (circle) all the words that you agree with. It will be most helpful for the assessment if you check a minimum of about half of the words on your list. It is OK if you choose a few words more or a few less. Still, only check a word if you agree with it.

Typical evaluation data collection procedures are needed to protect the rights of respondents and ensure the validity of scores. Instructors and program planners, for example, ought not to be present during instrument development processes and data collection when there is a possibility that their presence may introduce bias, distortion, or coercion. As with all adult education evaluation efforts, the sole purposes appropriate for use of this technique are to gather information to make informed program planning and instructional management decisions and to be accountable to stakeholders. In institution-based adult education programs, the technique can be extended to include learner assessments of program services and facilities including, for example, counseling services, registration procedures, library resources, levels of personal threat and learner comfort in the social space, and contributions of support staff.

DISCUSSION: WORDS VERSUS NUMBERS AND OTHER CONCEPTUAL ISSUES

This project relied on semiotics, semantics, and pragmatics rather than attitude-scaling theory to develop a technique for the construction of evaluation instruments for use by low-literacy learners. The technique is an alternative to psychometric approaches and confirms prior research that persons' vocabularies can yield meaningful and valid research data, particularly when mutual beliefs in particular social contexts are important (Clark & Marshall, 1981). The approach makes sense to learners and practitioners. As adult educators, we know what we want the educational processes and outcomes of our programs to be. We say and write what we intend to achieve without relying on metrics to aid our thinking and communication. We do not say, "OK, let's aim for a 6.4 on the evaluation scale for this workshop."

Respondents faced with Likert-type scales for the first time frequently express their frustration with the limits imposed by the boxes or spaces. They want their rating to be one quarter of the distance between the 2 and the 3 or, for example, they say, "Under some circumstances my response is a 2, in other circumstances it's a 3." Data collectors may respond with statements such as, "Do not worry too much; the statistical procedures can take these minor preferences into account; do your best with the categories provided; these scales are designed scientifically." Evaluators frequently dismiss respondents' statements as evidence of inexperience,

defensiveness, procrastination, anxiety, pedantry, antiquantitative attitudes, or anti-social noncompliance, yet the problem may be that the measurement technique selected for some groups is itself inappropriate.

When evaluators use Likert-type-scaling procedures, they rarely do more than report mean item and total scale score results. Many program planners and stakeholders are unable to make use of these results for program quality improvement purposes. The most useful information Likert-type scales generate is, in my opinion, box-and-whisker plots of distributions and standard deviations of scores, yet these statistics are rarely reviewed. Furthermore, mean Likert-type scale scores can provide unhelpful and misleading results. For example, consider the following case: On a 7-point Likert-type scale overall assessment of program quality, four persons select 7 (*excellent*), six select 4 (*average*), eight select 2 (*poor*), and two select 1 (*mediocre*). These responses result in a group mean of 3.5:

$$(4 \times 7) + (6 \times 4) + (8 \times 2) + (2 \times 1) = 70/20 = 3.5$$

It suggests an “average” rating by respondents and no urgent need for intervention. Checklist results, however, would reveal a different story. They would show that half the respondents think the course is poor or mediocre and in urgent need of attention and only one fifth of learners were engaged in a quality learning activity. Because Likert-type scale scores are susceptible to the effects of extreme values, in effect, the opinions of some respondents have a greater influence on group means than the opinions of others. Where democratic principles and popular education values are important aspects of a program, the adjective checklist is to be preferred to a Likert-type scale instrument.

The value of numbers in research and evaluation work is not found in their efficiency as signs for respondents’ understandings and meanings. Nor can numbers lay claim to greater validity in the communication of respondents’ understandings and meanings, because numbers are always translated to and from words and the opportunity for error enters into every such translation. Numbers are invaluable in certain aspects of evaluation and research because without them, item analysis for instrument improvement and statistical analyses of evaluation data is impossible. Today, all too frequently, and unnecessarily, as most evaluations do not require statistical analyses, the tail wags the dog. Learners often wish to express themselves in more direct and sometimes more subtle ways than scaling procedures allow and words, with their great range of meanings, when considered in well-defined contexts, are often the best means to express understandable personal preferences.

Language removes the constraints on expressions of values, perceptions, and beliefs imposed by numbers. For the highly literate, test and evaluation scale scores are always approximations of the true nature of the phenomenon being reported. For the lower literate, numbers are more gross approximations of the phenomenon and on occasion, like text, numbers can suppress the reporting of their views entirely. Simply stated, the evaluative adjectives selected by learners through this

technique are criteria according to which the worth of the learners' subjective experiences can be determined and communicated to others. Engaging low-literacy learners in the evaluation of their programs is an important activity leading toward the achievement of an important unstated educational outcome. Using evaluative words and phrases in a formal transparent process makes the user accountable for what she or he says. The learners are placed in the position of being responsible for what they say. They must say what they mean and mean what they say. This project confirms the findings of semantics and pragmatics research that leads research "away from 'in the head' concerns and onto issues of participation with others. We move from the individual to the social, from thinking to talking" (Forrester, 1996, p. 57), and for the adult education evaluator, from talk to action.

CONCLUSIONS

The research challenge faced in this project was to develop an instrument to elicit natural, thoughtful, and genuine everyday responses that individuals give when asked about their experiences in an adult education or development activity. The adjectival checklist technique is an alternative to number-based, precision-fallacy-riddled approaches. Experience confirms that this technique can be used to obtain valid and reliable evaluation data quickly with minimal resources and democratically with wide stakeholder participation.

The foundational assumption of the "blunt-instrument technique" is that natural language, in its simplest form as words, is the ultimate depository of meanings about the value of personal experience, and metrics are not required to mediate the communication of low-literate persons' experiences in adult education and development programs. The value of a numerical score lies in its contribution to statistical analysis, and the role of statistical analysis is to seek understandings that cannot be achieved by other means, such as causation and the probability that certain phenomena may occur. These uses are far less frequent than one might imagine given the widespread popularity of psychometrics. In discussions about program quality, a set of learner-chosen adjectives likely holds greater value for the majority of adult education participants and providers than scale scores that few use and interpret properly.

APPENDIX
ADJECTIVE POOL FOR ASSESSMENT OF
ADULT EDUCATION AND DEVELOPMENT PROGRAMS

abstract	complex	efficient	hostile
academic	complicated	elementary	humanistic
accepting	comprehensive	emotional	humorous
active	conceptual	encouraging	hurried
adaptable	concrete	energetic	idealistic
adult-oriented	confusing	engaging	imaginative
advanced	conscientious	enjoyable	immature
adventurous	conservative	enlightening	important
aggressive	considerate	entertaining	impractical
alert	conventional	enthusiastic	inappropriate
alienating	creative	excellent	inclusive
ambitious	daring	exciting	incomplete
analytic	defensive	exemplary	incomprehensible
apathetic	deliberate	exhausting	individualistic
appropriate	demanding	experiential	industrious
artificial	democratic	fair-minded	ineffective
artistic	demoralizing	familiar	informal
assertive	dependable	first-rate	informative
authoritarian	depressing	flexible	inhibited
autocratic	developmental	focused	innovative
average	different	forceful	insensitive
awful	difficult	formal	insightful
awkward	disappointing	fragmented	inspiring
balanced	dishonest	frank	intellectual
basic	disjointed	friendly	intelligent
boring	disorganized	frivolous	intense
calming	dissatisfying	frustrating	interesting
capable	distant	general	intolerant
cautious	distracting	good	introductory
challenging	doctrinaire	great	invigorating
chaotic	dogmatic	hard-headed	involving
childish	dull	hard-hearted	irrelevant
clear-thinking	dynamic	hard-working	irresponsible
closed	easy	helpful	irritating
common	effective	honest	lengthy

(continued)

APPENDIX (continued)

limited	planned-badly	satisfactory	thoughtless
limiting	planned-well	satisfying	thought-provoking
lively	pleasant	scattered	tight
logical	poor	second-rate	timely
long	positive	sensitive	tolerant
loose	powerful	serious	tough
mature	practical	severe	trusting
meaningful	pragmatic	shallow	typical
meaningless	predictable	significant	unclear
mechanical	prepared-badly	simple	unconventional
methodical	prepared-well	simplistic	undemocratic
mindless	progressive	sincere	understandable
mind-numbing	provocative	skillful	understanding
mismanaged	purposeful	slow	unemotional
moderate	rational	smooth	unfriendly
motivating	realistic	sociable	uninhibited
narrow-minded	reasonable	sophisticated	uninteresting
natural	redundant	spontaneous	unrealistic
negative	reflective	stable	unsatisfactory
new	rejuvenating	stimulating	unsuitable
nice	relaxed	straightforward	useful
obscure	relevant	strange	useless
old	reliable	strong	valuable
optimistic	repetitive	structured	versatile
orderly	repressive	superficial	warm
organized-badly	reserved	supportive	waste of time
organized-well	resourceful	surprising	weak
original	responsible	tedious	weird
outdated	responsive	tense	wonderful
paced-badly	restrictive	terrible	woolly
paced-well	ridiculous	theoretical (too-)	worrying
participatory	rigid	theory-based	worthwhile
patronizing	rough	thorough	
pessimistic	rushed	thoughtful	

NOTES

1. In *A Conceptual Scheme for the Identification and Classification of Processes for Adult Education*, Verner (1962) provided a very useful classification of adult education terms that enables important conceptual distinctions to be made between, for example, adult education methods and techniques.
2. Most statistics texts include one or more chapters on nonparametric testing. However, Siegel (1956) is the classic nonparametric reference text commonly cited in good contemporary texts.

REFERENCES

- Allport, G. W., & Odbert, H. (1936). Trait names: A psycho-lexical study. *Psychological Monographs*, 47(1), 211.
- American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: Author.
- Anglin, J. M. (1970). *The growth of word meaning* (Research Monograph No. 63). Cambridge, MA: MIT Press.
- Bollinger, D. (1980). *Language—The loaded weapon: The use and abuse of language today*. London: Longman.
- Cattell, R. B. (1946). *Description and measurement of personality*. Yonkers-on-Hudson, NY: World Book.
- Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, I. Sag, & B. Webber (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge, UK: Cambridge University Press.
- Culler, J. (1976). *Ferdinand de Saussure*. London: Harvester.
- Forrester, M. A. (1996). *Psychology of language: A critical introduction*. Thousand Oaks, CA: Sage.
- Frawley, W. (1992). *Linguistic semantics*. Hillsdale, NJ: Lawrence Erlbaum.
- Gough, H. G., & Heilbrun, A. B., Jr. (1965). *The Adjective Check List manual*. Palo Alto, CA: Consulting Psychologists Press.
- Gronlund, N. E. (1988). *How to construct achievement tests*. Englewood Cliffs, NJ: Prentice Hall.
- Hartshorne, H., & May, M. A. (1930). *Studies in the nature of character: III. Studies in the organization of character*. New York: Macmillan.
- Hervey, S. (1982). *Semiotic perspectives*. London: Allen and Unwin.
- Kristeva, J. (1986). Semiotics: A critical science and/or a critique of science. In T. Moi (Ed.), *The Kristeva reader* (pp. 74-88). New York: Columbia University Press.
- Morris, C. (1963). *Signification and significance: A study of the relations of signs and values*. Cambridge, MA: MIT Press.
- Niemi, R. G., Carmines, E. G., & McIver, J. P. (1986). The impact of scale length on reliability and validity: A clarification of some misconceptions. *Quality and Quantity*, 20(4), 371-376.
- Noth, W. (1990). *Handbook of semiotics*. Bloomington: Indiana University Press.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.
- Oxford English dictionary* (Vol. 1). (1970). Oxford, UK: Clarendon.
- Polkinghorne, D. E. (1984). Further extensions of methodological diversity for counseling psychology. *Journal of Counseling Psychology*, 31(4), 416-429.
- Remmers, H. H., & Weisbrodt, J. A. (1965). *Manual of instruction for the Purdue Rating Scale for instructors* (Rev. ed.). West Lafayette, IN: University Book Store.
- Siegel, S. (1956). *Nonparametric statistics in the behavioral sciences*. London: McGraw-Hill.
- Thibault, P. J. (1997). *Re-reading Saussure: The dynamics of signs in social life*. London: Routledge.
- Verner, C. (1962). *A conceptual scheme for the identification and classification of processes for adult education*. Chicago: Adult Education Association.