Meaningful Measurement of Theory-Based Service-Learning Outcomes: Making the Case with Quantitative Research

Robert G. Bringle and **Julie A. Hatcher**
Indiana University-Purdue University Indianapolis

Research is most beneficial when the design of research is guided by a theory and when the information that is gained through data collection is relevant to supporting, developing, refining, and revising a theory. The practice of service-learning will be improved when we understand the conditions that increase the likelihood of service-learning classes reaching intended educational outcomes. This article provides recommendations for generating meaningful information about service-learning that include evaluating hypotheses derived from theory, using multiple-item measures of theoretical constructs, using designs that allow causal inferences to be made, and making appropriate theoretical and practical generalizations from research.

The growth in the number of service-learning classes in higher education during the 1990s has been a remarkable example of pervasive educational reform. This expansion, however, is somewhat paradoxical because educators and administrators have invested time and resources developing service-learning courses and campus infrastructure when there is a paucity of research evidence documenting the effectiveness of service-learning in reaching educational objectives of the course, the curriculum, and the institutional mission. The increase in the number of service-learning courses without supporting evidence is all the more remarkable because it has occurred during a decade that has witnessed increased emphasis on assessment and accountability in higher education.

This article advocates that service-learning practitioners must devote more resources to conducting systematic, scientific assessment of service-learning outcomes across students, faculty, institutions, and communities. Conducting systematic scientific research with meaningful indicators of educational outcomes represents a public, peer-reviewed, and replicable type of information gathering that is important for increasing confidence among practitioners, providing a justification to those in positions to support the expansion and recognition of service-learning, and developing theory to enhance our understanding of practice.

The acute need for research on service-learning has not gone unnoticed. Wingspread conferences were conducted in 1991 and 1993 to develop a research agenda for service-learning. More recently, Giles and Eyler (1998), the Research Advisory Council convened by Campus Compact during 1997-1998, and the Campus Compact Presidents' Declaration on the Civic Responsibility of Higher Education provide a set of issues, topics, and questions for research. However, there has been limited guidance on how to conduct rigorous quantitative and qualitative research on service-learning through examples in the literature, graduate student training, professional development, and collaboration with researchers in cognate areas.

Reflective practitioners need to assess their practice in a manner that is structured, systematic, and meaningful (Hatcher & Bringle, 1997). Cambridge (1999) defines assessment as "generating and using information about performance that is fed back into the system from which it comes to improve that system" (p. 176). There are five important types of assessment (reflection, process evaluation or monitoring, outcome evaluation, correlation, and experiment) that are particularly relevant to providing feedback to educators to improve practice (see Table 1). In general, reflection and monitoring are more qualitative methods of assessment; correlational and experimental methods are more quantitative.

A number of factors shape when each of these different assessment approaches is preferred. First, the nature of the question being asked and answered can guide the type of assessment method that is selected. Some questions are better suited to qualitative methods (e.g., "How are my students reacting to their first day of community service?"), whereas others benefit from the use of quantitative methods (e.g., "How does the perceived efficacy of students in service-learning classes compare to students in traditional classes?"). Second, use of the information is important in guiding the selection of the method of assess-

TABLE 1
Types of Assessment

Reflection	Process Evaluation or Monitoring	Outcome Evaluation	Correlational	Experimental
Activities producing information oriented toward the self-assessment of persons *who* are engaged in an experience (e.g., student reflection, faculty teaching portfolio)	Activities producing information about *how* a class, course, or program was implemented (e.g., use of classroom assessment techniques; Did orientation occur at the service site?)	Activities producing information about *what* outcomes occurred as a result of a class, course, or program (e.g., student learning, changes in moral development, community impact)	Activities producing information about *when* a relationship exists between aspects of a class, course, or program (e.g., Are learning outcomes related to prior service experience?)	Activities producing information about *why* a specific outcome occurred (e.g., Does increasing the amount of structured reflection increase retention of course content?)

ment. Qualitative methods provide valuable information when compiling teaching portfolios, conducting program reviews, and providing narrative for publications. In contrast, quantitative methods may be more desirable when designing and conducting a program evaluation for a grant, publishing research for peer-reviewed publications, conducting research on theory-based hypotheses, and conducting multi-campus studies. A third consideration in selecting an assessment strategy is the nature of the audience. Stories and anecdotes may be particularly relevant to some audiences (e.g., presentations to a Board of Trustees, persuading students to enroll in service-learning classes), whereas quantitative data and systematic research may be more influential in other situations (e.g., accreditation documentation, presenting program data to the president before budget allocations).

We are especially interested in promoting quantitative research on service-learning that systematically collects information by following the prescriptions of scientific research. Other types of assessment are important and meaningful; however, scientific research holds great potential for improving practice, building a conceptual basis for understanding the strengths and limitations of service-learning, sustaining the growth and institutionalization of service-learning, and convincing colleagues about the merits of service-learning. The key components of this approach (see Figure 1) include developing theories that articulate relationships between constructs, deductively translating theoretical constructs into replicable and meaningful variables in order to evaluate theory-based hypotheses (construct validity), structuring data collection so that causal inferences can be made (internal validity), conducting appropriate statistical analyses that permit inferences about patterns of results (statistical validity), and making appropriate inductive inferences about the theoretical and practical implications of the

results (external validity). This overview highlights the importance of the relationship between theory and research in conducting research on service-learning. All research, both quantitative and qualitative, is most beneficial when the design of research is guided by a theory and when the information that is gained through data collection is relevant to supporting, developing, refining, and revising a theory.

Theory

Theories represent cognitive and linguistic templates that are laid upon phenomena. Yet they are not "merely theories" in the sense that they are inconsequential. McGuire poignantly states the indispensability of theories.

> What makes theorizing a tragedy is not that our theories are poor but that, poor as they are, they are essential, for we cannot do without them. The ubiquity of formal and informal theorizing

FIGURE 1
The Relationship between Theory and Research
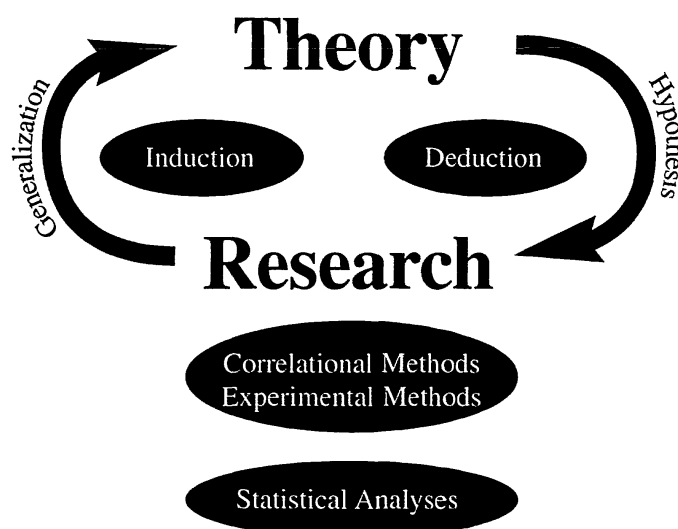
demonstrates its indispensability. To cope with reality we must reduce it to the oversimplified level of complexity that our minds can manage and distort it into the type of representations that we can grasp. We are reduced to groping for theories that are happy instances of brilliant oversimplification whose elected ignorances and distortions happen to be incidental to the matter under consideration, so that within the momentary situation the theory's apt focusing of our creative and critical appraisal yields gains that outweigh the losses caused by its oversights and distortions. (1980, p. 54)

Theories provide conceptual grounding for research. They articulate and detail (a) the nature of constructs (e.g., civic responsibility, motivation to volunteer, racial tolerance, active learning), (b) the relationships between a construct and observables (e.g., persons with racial tolerance display these characteristics...), and (c) relationships between constructs (e.g., racial tolerance is correlated with a motivation to promote social justice). When research is derived from theory and evaluates theory-based hypotheses, the work is more systematic and has broader implications. To provide merely a description of a specific service-learning program or an isolated service-learning class is severely limited because of the context-specific and idiosyncratic nature of the case description. To show how the design and implementation of the service-learning program or class was guided by theory, and to demonstrate how the experiences and outcomes are consistent or inconsistent with expectations derived from a theory, provides a much richer basis from which conceptual generalizations can be understood and a basis upon which lessons learned can be applied to other settings. Theories also provide a basis for integrating and differentiating disparate observations and findings and a basis for understanding boundary conditions for phenomena (e.g., when the theory or principle does not apply).

Giles and Eyler (1994) note that, "service-learning, as a relatively new social and educational phenomenon, suffers from the lack of a well articulated conceptual framework" (p. 77). The closest service-learning comes to having a theory of its own is Dewey's educational theory (Dewey, 1916, 1933), and Kolb's (1984) experiential learning theory based upon Dewey's work. Giles and Eyler (1994) and Hatcher (1997) provide excellent examples of how Dewey's educational theory generates testable hypotheses about the design, processes, and outcomes of service-learning classes. Theories can also be adapted from disciplines and other areas of research. For example, Stukas, Snyder, and Clary (1999) effectively use social psychological theory to

generate testable hypotheses about the influence of mandatory service on motives for future service. McEwen (1996) explores how student development theories relate to service-learning. The American Association for Higher Education's *Series on Service-Learning in the Disciplines*[1] is a rich resource for examples of how discipline-based theories can enrich and broaden the conceptual analysis of service-learning and guide research.

All five types of assessment described in Table 1 can be situated within and related to theory. Reflection activities can be guided by theory. For example, students can engage in reflection activities that are framed by theories and conceptual frameworks that deal with faith-based perspectives, political ideology, theories of career exploration, student development theory, or psychological theory. Similarly, monitoring program implementation can be guided by the theoretical constructs that provided the rational for program design.

Recommendation. Both quantitative and qualitative research on service-learning should include a theoretical context from which hypotheses are generated and within which results and implications are interpreted.

Meaningful Measurement

Theoretical constructs make sense out of observations and experiences. For example, critical thinking is a construct that gives coherence to differences that are observed in students' verbal remarks, behaviors, experiences, and written work. In order to execute scientific research, each construct must be operationalized, which refers to specifying (a) a set of procedures for implementing an intervention (e.g., students wrote five 3-page reflection papers during the semester that related the service activities to the course content and their personal lives), or (b) a set of procedures for measuring a construct (e.g., changes in motivation were assessed using a pre-, post-test with Clary and Snyder's Volunteer Functions Inventory, Clary & Snyder, 1991). A good criterion for an operationalization of a construct is the ability of another researcher to take the description of the operationalization and replicate the intervention or measure. Operationalizations are much more specific than the construct they represent and, as such, are samples of the construct's conceptual domain. For example, there are many conceptualizations of motivation; Clary and Snyder (1991) provide one that measures six motives for volunteering, one of which is an altruistic motive. Their scale contains five items that sample the domain of altruistic motives for volunteering. Although an operationalization of an intervention (e.g., service-learning class) is important, our assessment of how mea-

sures have been operationalized in research on service-learning identifies a critical deficiency that deserves attention: too much service-learning research has been based on single-item indices of unknown integrity rather than multiple-item measures with documented properties that establish their meaningfulness.

A measure is assumed to correspond to a single construct. When this is the case for a construct other than descriptive variables (e.g., sex, age, marital status), a single indicator measure has two potentially significant limitations: (a) it runs the risk of being an inadequate sample from the conceptual domain (i.e., a single indicator measures only a small portion of the construct), and (b) a single indicator possesses idiosyncratic qualities that make it an imperfect measure (i.e., it measures other constructs in addition to the target construct). The most direct means to overcome the shortcomings of a single-indicator index is to increase the number of indicators in the measure. Multiple indicators should provide a more representative sample of the construct's domain. For example, the five items that measure altruistic motives on Clary and Snyder's (1991) scale each measure somewhat different aspects of the motive. In addition, the summary index of a multiple-item measure averages across and minimizes the idiosyncratic qualities of each specific indicator.

A discussion of multiple-indicator indices of constructs most readily applies to paper-and-pencil scales, but, it is not limited to only that type of measurement. Single- and multiple-indicator indices can be taken from observations of behaviors, reports by key persons (e.g., site supervisors), samples of journal entries, physiological measures of reactions to stimuli to assess empathic responses, ratings of job and community service performance by clients, and archival measures (e.g., institutional data). In all cases, these indices are improved by having multiple indicators. For example, Epstein (1979) demonstrated that observations of students' punctuality at class increased in reliability when they were taken over multiple class meetings in comparison to when they were based on punctuality for only one class meeting.

One prerequisite for generating a set of meaningful and representative indicators is a clear articulation of the construct's attributes so that indicators can be adequately identified and sampled. For example, when developing a scale to measure a construct (e.g., career motives for volunteering), a pool of items should be generated that represents the construct's salient attributes according to a particular theoretical statement. However, it is possible to identify or produce good and bad indicators to represent a construct's conceptual domain. Pilot testing should be conducted on the preliminary items to assess their

psychometric properties and inadequate items should be discarded. One way to evaluate the quality of items is to evaluate the degree to which the set of items is coherent and uni-dimensional. Items that do not correlate with other items (or items that do not correlate with the scale total) should be discarded. The degree to which the resulting scale is unidimensional can be evaluated with a measure of internal consistency (e.g., coefficient alpha, split-half reliability) or through measures derived from factor analysis (e.g., theta, omega). All of these measures vary from .00 to 1.0 and the numerical value is a function of (a) the average inter-item correlation and (b) the number of items.

According to Nunnally (1967), coefficient alpha is an estimate of the correlation between the scale and a hypothetical alternative form of the scale of the same length. Alternatively, it is an estimate of the correlation of the scale with the construct's true score. Coefficient alpha is also an estimate of the degree to which the scale's scores are free of error. An important principle that is related to coefficient alpha is that, other things being equal (e.g., item quality), the more items a scale contains, the more reliable, coherent, and error free it will be. Practice indicates that there can be rapid gains in alpha when the number of items is increased from two to five or more items. Clearly, there is a point of diminishing returns; adding more and more indicators ceases to increase alpha and, in the case of paper-and-pencil scales, may burden respondents with too many items. The optimal number of indicators that are used, however, should be empirically determined from an item analysis.

Although there is a predisposition to think of an index (e.g., scale) as being reliable, it is better to consider reliability to be an attribute of the scores that are being analyzed. If the scale lacks internal consistency in a particular data set, then this will preclude subsequent analysis even if the scale has been found by others to be internally consistent. Lack of reliability undermines the ability to detect systematic effects in the results. For that reason, coefficient alpha (or an equivalent measure) for that data set should accompany every multiple-indicator index of a construct that is reported in research. In addition, it may be appropriate to assess the temporal stability of a measure when the attribute being measured can be considered constant across time (e.g., self-esteem) and not malleable (e.g., attitudes). This is most typically done with a test-retest measure of reliability. When making observations, assessing inter-rater agreement is important to establishing the integrity of the information.

Finally, not only is it desirable to use multiple-indicator measures that are unidimensional and reli-

able, but it is also important that the index provide a meaningful measure of the intended construct (Messick, 1989). The validity of an index can be assessed by comparing scores of persons known to differ on the attribute (known groups validity; e.g., experienced volunteers score higher on motives to volunteer than do inexperienced volunteers), correlations of scores with other measures of the same or related construct (convergent validity; e.g., a measure of moral development correlates with other measures of moral development), and nonsignificant correlations of scores with constructs posited to be unrelated to the target construct (discriminant validity; e.g., scores on a scale are uncorrelated with a measure of the social desirability response bias).

Recommendation. Research on service-learning should incorporate psychometrically sound, multiple-item measures of constructs for which there is evidence supporting their construct validity.

Research Designs for Making Causal Inferences

Most theoretical propositions involve a causal relationship between two constructs (e.g., structured reflection results in higher academic performance than unstructured reflection). Such causal relationships also provide practical guidance for designing courses (e.g., structured reflection can be expected to produce higher academic performance than unstructured reflection). Therefore, the most useful information to collect through research will evaluate causal relationships. Most research conducted on service-learning is severely limited by two factors that confound the ability to make causal inferences: (a) the use of correlational designs, and (b) the self-selection of participants into groups that produces pre-existing differences in groups. Regarding correlational designs in which two naturally occurring constructs are measured and the nature of the covariation is assessed (e.g., a Pearson product-moment correlation coefficient), it is typically very difficult to infer which variable caused the covariation in the other, or if the covariation is spurious and due to some other unmeasured construct.

In the case of pre-existing group comparisons in which, for example, the outcomes of students in a service-learning class are compared to students in a non-service-learning class, the ability to make causal inferences is limited by the possibility of pre-existing differences due to the self-selection of students into those types of classes. Eyler, Giles, and Braxton (1997) found pervasive pre-existing differences; students enrolled in service-learning classes scored higher on pre-test measures of citizenship confidence, citizenship values, citizenship skills, and perceptions of social justice, when compared to students in non-service-learning classes.

There are two ways of dealing with pre-existing group differences. First, pre-existing differences can be measured and their influence statistically removed as covariates. Sax and Astin (1997) and Bringle and Kremer (1993) used statistical control of covariates as a means for controlling pre-existing differences. Unfortunately, this solution is limited to the set of covariates that is measured and there is no assurance that all relevant covariates have been measured.

Another solution is to randomly assign participants to conditions in order to make it unlikely that there are significant differences on any relevant variables. It may be possible under some circumstances to randomly assign students to groups (Stukas, Snyder, & Clary, 1999); however, assigning students to classes is typically impractical. Another approach has students select sections of a course without knowing ahead of time about differences in course requirements associated with each. Markus, Howard, and King (1993) and Osborne, Hammerich, and Hensley (1998) successfully used this approach. Although this procedure does not fit the strict definition of random assignment, it can be a reasonable approximation. In order to preserve the integrity of this procedure (a) movement of students between service-learning and non-service-learning sections must be prohibited or monitored, analyzed, and reported and (b) differential dropout from the two types of sections must be monitored, analyzed, and reported.

Recommendation. Research on service-learning should use research designs that increase the confidence with which causal inferences can be made between two constructs.

Statistical Validity

The ability to make appropriate statistical inferences about the pattern of results depends on several factors. First, the appropriate statistic must be selected according to the nature of the variables in the research and the question being answered. All statistical procedures have assumptions. Using statistics that clearly violate assumptions (e.g., parametric statistics on nominal data) can lead to erroneous results being interpreted. However, in some circumstances, violation of assumptions is a matter of degree (e.g., deviation from normality) and some compromises on assumptions are tolerable. One area in which structuring the statistical analysis becomes particularly tricky is when change scores are being monitored across time. Data collected from repeated measure designs require suitable statistics. Analyzing algebraic difference scores between pre-, post-tests is problematic because the difference score is less reliable than the constituent measures.

Second, the research study must have a sufficient number of subjects to detect an effect if it is present. In general, larger sample sizes are advantageous because the statistic will have increased power (i.e., the ability to detect a significant effect that is present in the population) and it is possible to conduct more refined internal analyses of the data set (e.g., comparing males and females, majors and nonmajors). Some research on service-learning has been constrained by small samples and lack of power. Wilkinson (1999) provides excellent advice about these and additional issues related to statistical validity in research.

Recommendation. Research on service-learning needs to use appropriate statistics to make inferences about the implications of results and needs to have sufficient sample size for adequate power.

Generalizing Research Results

The primary purpose of conducting research is to gain knowledge about a phenomenon and communicate that knowledge so that others may benefit in their work. Induction is the generalization of broad conclusions or theoretical propositions from specific facts and findings. The quality of induction relies on evaluating the degree to which the specific information forms an appropriate basis for a broad conclusion. An evaluation of the accuracy of a broad conclusion will be a matter of degree or likelihood because the conclusion necessarily exceeds the information upon which it is based. The degree to which research findings generalize is called external validity.

One issue of external validity is the nature of the sample that was used in the research. Generalizing from a sample (e.g., students in a psychology service-learning class at a liberal arts college) to broader populations (e.g., all first year students at that institution, all students in psychology classes, all students in liberal arts colleges) depends on the similarity between the sample and the population. Random sampling and its variations (e.g., stratified random sampling) is one approach to ensuring similarity. A related issue is the heterogeneity of the sample: the less redundant the sample is, the more broadly the generalization can be. If service-learning students from only one type of class or one type of institution are studied, generalizations that can be confidently drawn are more limited than if service-learning students from different types of classes and different types of institutions are included (e.g., Eyler, Giles, & Braxton, 1997; Sax & Astin, 1997). A final issue that influences the confidence with which generalizations can be made is that, all things being equal (e.g., the quality of the sample), larger samples lead to greater confidence in generalizations.

Not all research has as its primary purpose making accurate generalizations to a specific population of subjects. Mook (1983) points out that some research is conducted for other reasons, such as to know if something is possible, to study unusual cases, and to test a theoretical proposition. In these cases, generalizing to a broader sample is less important than conducting high quality research that has construct validity and internal validity. Most social science research relies on accidental and convenience samples, rather than randomly selected samples, because it is testing theoretical propositions that should apply to any reasonable sample. Being able to generalize the results to a theory (construct validity of causal inferences) is often more important than being able to generalize to a population of subjects.

Although these examples have focused on the subjects who are participating in the research, that is only one aspect of external validity. The same principles apply to other aspects of research, including different types of operationalizations, settings, conditions, sources of data, and time periods. For example, the issues of heterogeneity and sample size also apply to how constructs are operationalized (e.g., Is there a set of heterogeneous, yet theoretically representative operationalizations that is large in number and that provide converging results?). Confidence in making generalizations from findings is increased when results replicate across maximally heterogeneous operationalizations (typically across different studies). It is for this reason that researchers, when possible, measure the same construct (e.g., self-efficacy) using different measurement procedures (e.g., self-report, archival records) and different sources of information (e.g., teacher reports of student learning, student reports of learning), some of which may be qualitative.

A deficiency we have noted in research on service-learning is the tendency to report specific findings, most typically from case studies (e.g., one class, one program, one institution) without making justified generalizations about practice, theory, and policy. In part, this shortcoming is due to the absence of a theoretical context within which the course or program was developed and from which broad inferences can be made. Qualitative research is assumed to contribute strongly to the inductive side of the relationship between theory and research because of the hope that it will cast a broad net for effects and identify important constructs and relationships for subsequent research. However, too often, qualitative research is not situated in theory and fails to make significant and appropriate contributions to developing and refining theory. In the absence of a pre-existing theory, it is important that qualitative research

generate appropriate broad principles that can be subsequently tested with either qualitative and/or quantitative methods.

Recommendation: Research on service-learning, both quantitative and qualitative, should make appropriate and meaningful theoretical and practical generalizations from research results.

Conclusions

Scientific research on service-learning provides a significant and even necessary component of work that practitioners in service-learning must more frequently use to (a) develop theory that explains the process and outcomes of service-learning, (b) improve the practice of service-learning courses and programs, (c) facilitate developing a culture of evidence and assessment on campuses, (d) offer a justification for increased allocation of campus resources to service-learning, and (e) provide a basis for developing policy associated with the institutionalization of service-learning in higher education. To the degree that this presents a daunting task, it will be necessary to establish collaborative research teams, provide ongoing professional development to service-learning educators, and garner internal and external resources to support the research.

Systematic, scientific, theory-based research with reliable and valid operationalizations has many strengths, but it is sometimes criticized as being incomplete and falling short of capturing the rich experiences of practitioners, students, and community partners involved in service-learning. Although using multiple-item scales to measure a construct is systematic, deciding to use a particular scale may mean that other important aspects of what else is occurring are not being measured. In addition to being rather narrow, using scales to measure constructs is predicated on the outcomes being identified ahead of time and leaves little room for unexpected outcomes. However, it must be acknowledged that no information gathering procedure can measure everything and that both quantitative and qualitative research methods have strengths and weaknesses. The solution to this conundrum may reside in sequencing or mixing quantitative and qualitative methods (Driscoll, Holland, Gelmon, & Kerrigan, 1996; Eyler & Giles, 1999). When sequencing methods, we recommend that qualitative methods (e.g., focus groups, interviews, student journals) be used to identify relevant constructs and to map conceptual domains that become the basis for quantitative analysis. In either case, qualitative research that provides quotable material, stories, conceptual insight, and diverse information can complement the more focused and systematic quantitative findings. Replicability of results and extract-

ing principles from which other practitioners can learn should be important to both quantitative and qualitative research.

The stakes are high for establishing what service-learning can and cannot do effectively. One of the most persuasive methods for establishing the value of service-learning is through quantitative research from which casual inferences about meaningfully measured outcomes can be made. For as Thorton (1999) notes,

> Science is one of the glories of our world. It is the work of the analytic mind in love with life, and its poetry is mathematics. Science continuously reveals wonder, if we let it ... (p. 26)

Notes

[1] For more information about this series visit the AAHE web page [www.aahe.org}.

References

Bringle, R. G., & Kremer, J. F. (1993). An evaluation of an intergenerational service-learning project for undergraduates. *Educational Gerontology, 19,* 407-416.

Cambridge, B. (1999). Effective assessment: A signal of quality citizenship. In R. G. Bringle, R. Games, & E. A. Malloy (Eds.), *Colleges and Universities as Citizens* (pp. 173-192). Needham Heights, MA: Allyn & Bacon.

Clary, E., & Snyder, M. (1991). A functional analysis of altruism and prosocial behavior: The case of volunteerism. In M. Clark (Ed.), *Review of Personality and Social Behavior,* vol. 12 (pp. 119-148). Newbury Park, CA: Sage.

Dewey, J. (1916). *Democracy and education.* New York: Macmillan.

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process.* Boston: D.C. Heath.

Driscoll, A., Holland, B., Gelmon, S., & Kerrigan, S. (1996). An assessment model for service-learning: Comprehensive case studies of impact on faculty, students, community and institution. *Michigan Journal of Community Service Learning, 3,* 66-71.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097-1126.

Eyler, J., & Giles, D. E., Jr. (1999). *Where's the learning in service-learning?* San Francisco: Jossey-Bass.

Eyler, J., Giles, D. E., Jr., & Braxton, J. (1997). The impact of service-learning on college students. *Michigan Journal of Community Service Learning, 4,* 5-15.

Giles, D. E., Jr., & Eyler, J. (1994). The theoretical roots of service-learning in John Dewey: Toward a theory of service-learning. *Michigan Journal of Community Service Learning, 1,* 77-85.

Giles, D. E., Jr., & Eyler, J. (1998). A service learning research agenda for the next five years. In R. A. Rhoads & J. P. F. Howard (Eds.), *Academic service learning: A pedagogy of action and reflection* (pp. 65-72). San Francisco: Jossey-Bass.

Hatcher, J. A. (1997). The moral dimensions of John Dewey's philosophy: Implications for undergraduate education. *Michigan Journal of Community Service Learning, 4,* 22-29.

Hatcher, J. A., & Bringle, R. G. (1997). Reflection: Bridging the gap between service and learning. *Journal of College Teaching, 45,* 153-158.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development.* Englewood Cliffs: Prentice-Hall.

Markus, G. B., Howard, J. P. F., & King, D. C. (1993). Integrating community service and classroom instruction enhances learning: Results from an experiment. *Educational Evaluation and Policy Analysis, 15,* 410-419.

McEwen, M. K. (1996). Enhancing student learning and development through service-learning. In B. Jacoby & Associates (Eds.), *Service-learning in higher education* (pp. 53-91). San Francisco: Jossey-Bass.

McGuire, W. J. (1980). The development of theory in social psychology. In S. Duck & R. Gilmour (Eds.), *The development of social psychology* (pp. 53-80). London: Academic Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-387.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw Hill.

Osborne, R. E., Hammerich, S., & Hensley, C. (1998). Student effects of service-learning: Tracking change across a semester. *Michigan Journal of Community Service Learning, 5,* 5-13.

Sax, L. J., & Astin, A. W. (1997). The benefits of service: Evidence from undergraduates. *Educational Record,* Summer/Fall, 25-32.

Stukas, A. A., Snyder, M., & Clary, E. G. (1999). The effects of "mandatory volunteerism" on intentions to volunteer. *Psychological Science, 10,* 59-64.

Thorton, J. (1999). *A field guide to the soul: A down-to-earth handbook of spiritual practice.* New York: Bell Tower.

Wilkinson, L. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594-604.

Authors

ROBERT G. BRINGLE is Professor of Psychology and Director of the Center for Public Service and Leadership at Indiana University-Purdue University Indianapolis (IUPUI). His interests in service-learning include developing ways to implement and institutionalize service-learning, and conducting research on student and faculty attitudes, educational outcomes, and institutionalizing service-learning. He is also interested in the role of service in the academy and institutional change to support community engagement. He is editor of *With Service in Mind: Concepts and Models for Service-Learning in Psychology* (with D. Duffy), and *Colleges and Universities as Citizens* (with R. Games & E. Malloy).

JULIE A. HATCHER is Associate Director of the IUPUI Center for Public Service and Leadership. She is an instructor in Philanthropic Studies and conducts the Community Service Seminar. She has offered a variety of professional workshops. Her special interests include faculty motivations in service-learning, institutionalization of service-learning, and the relationship of service-learning to goals of general education. She is co-editor (with C. L. Foos) of "Service-Learning Workshop Curriculum Guide" and she edited "Service-learning Tip Sheets: A Faculty Resource Guide," both published by Indiana Campus Compact.