Potential Issues and Pitfalls in Outcomes Assessment in Leadership Education

David M. Rosch
Assistant Director, Illinois Leadership® Center
University of Illinois at Urbana-Champaign

Leslie M. Schwartz
Program Manager, Illinois Leadership® Center
University of Illinois at Urbana-Champaign

Abstract

As more institutions of higher education engage in the practice of leadership education, the effective assessment of these efforts lags behind due to a variety of factors. Without an intentional assessment plan, leadership educators are liable to make one or more of several common errors in assessing their programs and activities. This article reviews some of the most common errors in assessing leadership programs, including failing to consider what the authors describe as the Honeymoon, Horizon, Hollywood, Halo, and Hallmark effects (otherwise known as recency, response-shift, socially desirable, and respondent biases). Suggestions for minimizing these potential issues in assessment are also included along with suggestions for practitioners and areas for future related research.

Introduction

Developing students' capacity to effectively lead in a global society is currently at the forefront of our higher education outcomes (Day, 2001; Roberts, 2003; Rost & Barker, 2000; W. K. Kellogg Foundation, 1999; Zimmerman-Oster & Burkhardt, 1999). Astin and Astin (2000) assert that "virtually all of our social institutions are hungry for people who are self-aware, authentic, innovative, empathic, committed, comfortable working collaboratively, and able to lead constructive change efforts" (p. 31). Additionally, multi-institutional data upholds that participation in leadership programs positively impacts students' educational and personal development, producing increases in civic responsibility, multicultural awareness, understanding personal values, and other outcomes related to leadership development (Cress, Astin, Zimmerman-Oster, & Burkhardt, 2001). Based on these outcomes, higher education has experienced an increasing focus on promoting student leadership development through curricular and co-curricular initiatives that have surfaced in our institutions over the past decade (Astin &

Astin, 2000; Riggio, Cicilla, & Sorenson, 2003; Astin & Cress, 1998; Schwartz, Axtman, & Freeman, 1998). Moreover, the post-industrial paradigm shift from viewing leadership as positional to embracing a relational model has made leadership education more accessible to a diverse group of students, instead of reserving the development of leadership to a privileged elite or a select group (Cress et al., 2001; Komives, Lucas, & McMahon, 2007; Rost & Barker, 2000). Viewing leadership as a collaborative, participatory process requires that one understands leadership within the context of a much larger system of relationships and connections, transforming how leadership is understood and taught within our higher education system by involving multiple stakeholders within our campus communities. This has become manifest on the academic front where the shift away from studying leadership through a single discipline to taking a multidisciplinary approach has shaped how the field of Leadership Studies is conceptualized. Rost (1991) upholds that this "leadership studies approach allows scholars and practitioners to think radically new thoughts about leadership that are not possible from a unidisciplinary approach" (p. 2).

The various connections among disciplines and departments have opened up new and exciting avenues to explore best practices for leadership education. We are experiencing a time in higher education where student leadership development is branching out beyond student affairs as the academic field of Leadership Studies continues to grow with the creation of interdisciplinary undergraduate minors, baccalaureate degrees, and graduate programs (National Clearinghouse for Leadership Programs, n.d.). While all of the new developments in the field of Leadership Studies provide opportunities for growth, it is critical to consider related challenges and issues related to measuring the impact of these interventions.

Defining the Problem

With these changing dynamics in higher education and an increasing focus on promoting student leadership development within our institutions, it is essential to accurately gauge the outcomes and impact of our leadership development efforts to ensure that they are effectively reaching their stated goals. Recently, Eich (2008) argued that "there has been little empirical research on student leadership program quality and program activities that contribute significantly to leadership development and learning" (p. 176). Effectively studying and understanding students' leadership development is connected to the complex challenge of accurately gauging the impact of such experiences, which is often confounded because the construct of leadership is so difficult to isolate in contemporary multi-disciplinary approaches (Grove, Kibel, & Haas, 2005).

Many researchers and scholars have developed means to assess student leadership ability and development in college students (Astin, 1996; Chambers, 1992; Cress et. al, 2001; Dugan & Komives, 2007; Pascarella & Terenzini, 2005; Posner & Brodsky, 1992). Yet even with all of these instruments and measures, the fact that there still exists the challenge of how to accurately define leadership (Burns, 1978; Bennis & Nanus, 1985; Barker, 1997). Rost (1991) upholds the point that "neither the scholars nor the practitioners have been able to define leadership with precision, accuracy, and conciseness so that people are able to label it correctly when they see it happening or when they engage in it" (p. 6).

Moreover, Day (2001) highlights the differences between leader and leadership development, claiming that leader development focuses on developing individual skills, knowledge, and abilities to enhance human capital, while leadership development requires relationships and interpersonal exchange to build social capital. Based on this distinction, leadership development in higher education should be conceptualized as an "integration strategy" where students learn the skills required for their own individual success in areas such as project management and acting ethically, as well as how to better relate to and connect with others in order to develop organizations and promote transformative change.

Kirkpatrick (1996) devised four increasingly impactful "levels" in evaluating training programs: (a) gauging participant reactions, (b) gauging participant learning, (c) evaluating changes in participant behavior, and (d) measuring organizational impact as a result of participant behavior change. Educators must be aware that when assessing student leadership development, it is critical to differentiate between measuring students' satisfaction and immediate reaction to the program *and* assessing learned leadership skills that are applied and practiced by students once a program is over or a course is complete. While having students complete program evaluation forms may tell educators about satisfaction levels or perceived benefits of a specific leadership experience based on students' immediate self-report data, accurately measuring behavioral and developmental changes in students' leadership development because of a program or experience is more complex.

For the purpose of this article, we incorporate Upcraft and Schuh's (1996) definition of assessment: "any effort to gather, analyze, and interpret evidence which describes institutional, departmental, divisional or agency effectiveness" (p. 18), but they narrow it to include evidence related specifically to the gain in leadership outcomes by the students involved with the institution. We also differentiate assessment from research, which Upcraft and Schuh describe as evidence collected to test hypotheses. While increasing research in the field of leadership studies is important and necessary, this article will focus on the current issues and potential pitfalls that exist for leadership educators regarding the

measurement of learned leadership skill in students participating in educational courses and programs.

Assessing Learned Leadership Skill

While scholars have examined how to effectively measure leadership experiences (Zimmerman-Oster & Burkhardt, 1999; Cress et al., 2001; Owen, 2001), it is also essential to consider various issues that may potentially confound the accuracy of such assessments and the related understanding of the impact of leadership interventions. The following section introduces various issues to consider when assessing leadership interventions and discusses five effects that we have identified as problematic in the measurement of students' learned leadership skill. Once an appropriate assessment tool has been chosen or constructed, several difficulties remain in the accurate measurement of learned leadership skill. While some of the difficulties previously discussed are comprehensive to the field of leadership education, others can be minimized by those seeking to expand their ability to measure student change in leadership skill through programs, courses, and other means in the field of education. To this end, we have identified five issues common to the assessment of leadership skill in the field of education that we describe below along with a discussion about strategies to minimize these issues; the issues are ranked from easier to minimize to more difficult to minimize. A summary of these issues can be found in Table 1.

Table 1. Summary of Assessment Issues

Assessment Issue	Description	To Minimize Bias
Honeymoon Effect	Participants may overstate effects of intervention immediately upon its completion.	Assess effects after participants have had a chance to integrate curriculum into their everyday experiences.
Horizon Effect	Participant responses on a pre-test may become less valid as their perceptions of their learning change throughout the intervention.	Avoid pre-post tests and utilize then-now methodology, assessing learning only after the intervention.
Hollywood Effect	Participants often rank themselves higher in areas that are socially desirable, such as the effective practice of leadership.	Utilize a 360-degree assessment that includes the use of peers and other observers to assess whether learning has been integrated into practice.
Halo Effect	Observers who have a positive conception of someone in one area (e.g., public speaking) may rank them higher in other areas as well (e.g., conflict resolution).	Explicitly define the specific leadership behaviors participants are expected to practice, especially when utilizing observers familiar with participants in the assessment process.
Hallmark Effect	Participants not confident of their skills or who do not identify themselves as "leaders" depress their self-report of the impact of their learning.	Avoid "lumping" specific behaviors into a broad definition of leadership. Assess these behaviors in addition to, or instead of, a general assessment of "leadership."

The Honeymoon Effect

Many educators attempt to assess student learning immediately upon completion of a program. Past research has supported the efficacy of such an approach as a means to avoid inaccurate reporting based on faulty recall (Uleman, 1991). However, within the field of leadership education, the lack of delay between intervention and assessment can also prove to be problematic. The Honeymoon Effect, called the recency effect in the field of evaluation studies (Evertson & Green, 1986), refers to bias in participant response in overstating the effect of the intervention immediately upon the completion of the program. Many leadership retreats, classes, and institutes focus on teaching value and goal clarification,

emotional intelligence, relationship building, and other "soft" technical and adaptive skills. At the completion of the intervention, participants may feel that they have made significant progress in a particular area within their own practice of leadership, but have had no opportunity yet to test their assumptions in their everyday environment. Therefore, asking participants to accurately rate their learning and, in essence, the effectiveness of the program in teaching these skills, is close to impossible and opens any assessment to significant criterion validity issues.

Educators who wish to minimize the Honeymoon Effect should seek to have participants assess their learning and practice after they have had a chance to effectively integrate their newly developed abilities in the crucible of their everyday experiences. While waiting an appropriate period of time to implement such assessment practices may reduce response rate, it will also increase the accuracy of the data gained from such efforts. Moreover, many educators practice a combination of assessment efforts where some or all of the participant population is assessed immediately upon program completion, and then another sample of the population is chosen to complete a follow-up assessment weeks later.

The Horizon Effect

Another oversight that many professionals responsible for assessing leadership programs make is failing to consider the Horizon Effect. When utilizing repeated within-subjects assessment measures (e.g. participants complete an assessment more than once, often before or at the beginning of an intervention and then at the end), a fundamental error can be made leading to a control response shift. This "response shift bias," as it is referred to the field of evaluation, occurs when the internal standards of the participant changes during the assessment process – in effect, the horizons of the participants shift (Taylor, Russ-Eft & Taylor, 2009). Understanding the impact of the Horizon Effect is particularly important in leadership education, as these educators often seek not just to change the behavior of participants, but the perspectives of the participant as well. For example, some students entering a program that teaches emotional and social intelligence skills may feel that they are relatively competent in this area of leadership before they begin the program, but have some technical aspects to their skills that they are seeking to augment; some students may even enter believing they have nothing to learn.

However, through their participation in the program, students may come to discover that they are not as skillful as they initially thought they were, that the effective management of one's emotions is more difficult than they initially thought, and that there are foundational aspects within this area of leadership in which they have much to learn. If these students were asked to complete a pre-test

before the program and a post-test afterwards, educators may be surprised to learn that these students may rank themselves below their initial level at the completion of the program, especially if the program took place over a long period (such as in a class) where students do not remember how they ranked themselves initially. Without taking the Horizon Effect into account, the impact of this program may be misunderstood because the outcomes would appear to show the program ineffective in meeting its stated goals.

One relatively simple method for minimizing the Horizon Effect is to eschew pre-program assessments and structure post-program assessments using a then-now structure (Rohs, 2002). Here, educators ask participants to rate their abilities after the completion of the intervention at the same time as rating their abilities before participating. While there are numerous issues involved in asking participants to accurately remember the past (Uleman, 1991), utilizing a then-now assessment structure allows participants to rank their abilities both before and after the intervention through the same prism of their present experience. This practice provides a more accurate method for effectively measuring students' learned leadership skill, while minimizing the impact of the Horizon Effect.

The Hollywood Effect

Another fundamental issue in the accurate assessment of learned leadership skill is that the practice of leadership is considered socially desirable in many parts of contemporary society, introducing an issue we call the Hollywood Effect. Research has shown that the practice of leadership is often conflated with positions of authority and power (Heifetz, 1994). Society lionizes those it considers good leaders through a variety of popular media, not dissimilar to the way society glorifies celebrities. This effect can lead to some leadership program participants consciously or unconsciously inflating the rating of their leadership skills, regardless of when they are asked to assess them, due to the attractiveness of considering oneself an effective leader. This effect is often referred to as a social desirability bias (Hill & Betz, 2005). Moreover, when asking about complex constructs such as the conscious possession and acting upon of values or the ability to manage interdependent authentic relationships, the accurate assessment of the practice of such skills and attributes is already difficult. A primary method of minimizing the Hollywood Effect is through the utilization of 360-degree assessment procedures, which have been popular in evaluation research for many years (Atkins & Wood, 2002) and have been shown to be more effective when evaluating leadership practices than self-ratings (Atwater & Yammarino, 1992). Not only do the participants themselves rate their practice of the skills involved in the program, others who know participants rate their practice as well. The most accurate way to utilize this assessment is by instructing participants to invite observers from a broad cross-section of their life (e.g., classmates, supervisors, friends, family) for inclusion in the assessment. A 360-

degree structure allows for assessment to include quantifiable and observable behavior, not just self-perceptions of ability. This minimizes the ability of the individual to inflate his or her potential to practice leadership based on a conscious or subconscious desire to be considered an effective leader.

The Halo Effect

While many educators may be familiar with the term "Halo Effect." We have found that some misunderstand the issue it is meant to explain. The Halo Effect describes a phenomenon where the perception of an individual's skill in an area of expertise is inflated based on his or her possession of a high degree of skills or a desirable attribute in an unrelated area (Thorndike, 1920). The desirable attribute places a halo around the individual, and causes others to overestimate his or her abilities in unrelated areas. This may explain, in part, why we see famous sports figures in commercials advocating non-sports-related items and why the average height of Fortune 500 CEOs is above six feet (Engemann & Owyang, 2005). The existence of the Halo Effect makes the effective assessment of leadership education more difficult, particularly when educators utilize a 360-degree structure. Students who display a high degree of skill in one area of leadership (e.g., charisma) may receive high scores in other areas of the practice of leadership (e.g., conflict management, organizational effectiveness) from their peers who perceive the halo of charisma around them, even when those students do not display these other skills.

Minimizing the Halo Effect, or even regulating it, when the definition of leadership is vague is difficult. The elimination of 360-degree assessments as a tool to minimize the Halo Effect only hampers effective assessment in other ways and opens evaluators to the Hollywood Effect. Therefore, educators should operationalize their definition of leadership as much as possible and be explicit with program participants and those completing assessments for others regarding the skills required for good leadership, and how those skills may relate or are independent to each other. Teaching students as part of a program that the skill of creating and maintaining good relationships with team members is only marginally related to the skill of managing a complex project helps them understand that someone might be effective in one area, but not another.

The Hallmark Effect

To many students, the word leadership is loaded with dense meaning. To some, it implies a sense of structured responsibility, is attached to prestigious titles such as president and chair, and is tightly coupled with other aspects of "officialdom" that are either consciously avoided or thought to be undeserved. Past research has shown that many students feel they do not possess the skills required for effective leadership (Schertzer & Schuh, 2005). For students of color and women

particularly, the title leader is not something they aspire to receive. This is because for them, leading is often associated with "acting White" especially in mixed-gender and mixed-race groups (Arminio et al., 2000; Kezar & Moriarty, 2000). Regardless of their reasoning, these students, in effect, avoid the hallmarks associated with the words leader and leadership.

The Hallmark Effect can create a subtle bias in responses to leadership assessments, especially when the students who complete them have not personally chosen to be part of a leadership program – such as when students complete assessments as part of a control group compared to students who do participate. This could lead to inflated perceptions of program impact or effectiveness due to the lower scores of the control group. The evidence for differences in self-ratings of leadership skill may be related to the differences in attractiveness students have for describing themselves as leaders and the skills they possess as leadership. These issues can persist even when assessing for skills that are only marginally related to leadership (e.g., ethical behavior, self-awareness, organizational development) when the overall assessment is labeled as a measurement of leadership. Therefore, the Hallmark Effect is, in some ways, directly relates to the Halo Effect.

Minimizing the Hallmark Effect in the assessment of learned leadership skill can be difficult; however, a number of methods may reduce the impact of this effect. Assessment efforts should include explicitly defined types of behaviors that make up effective leadership as well as examples of these practices across a wide range of contexts both inside and outside a university setting. For example, helping students to see that learning to resolve group conflict is a leadership act regardless of assigned position within the group may reduce differences in response rates based on the Hallmark Effect. Assessment staff should also avoid unintentionally labeling behavior as practicing leadership unnecessarily. There may be a difference in responses between the questions "how often, as a leader, are you able to help group members see both sides of conflict" and "how often are you able to help group members see both sides of a conflict." These efforts are particularly important when including uninvolved students as a control in assessment efforts because they have not had the benefit of learning the same vocabulary included in the program or course.

Suggestions for Practitioners

This article highlights several issues that may exist when trying to accurately measure learned leadership outcomes within the context of the educational environment. As previously discussed, each of the issues can be potentially minimized by specific actions taken on the part of the leadership educator. Additionally, some broader themes that encompass these issues exist that should

also be highlighted including focus clarity, participant vs. observer reporting, and measurement timing.

Focus Clarity

As mentioned in the introduction, a central weakness to the study of leadership is the lack of clarity over what is meant by the word "leadership" (Rost, 1991). Therefore, educators must identify an explicit focus when assessing their leadership programs. Common sense dictates that simply asking participants "has this program aided you in becoming a more effective leader" is, in many ways, excessively broad if the goal is to accurately answer it through realistic assessment efforts. Instead, program or course assessments should have a specific focus based on the intervention's learning outcomes. This not only helps participants focus on particular aspects of their behavior as a result of the program or course, but it also aids in minimizing the existing confusion and student ambivalence around what is meant by leadership.

Leadership educators who wish to build a foundation of accurate assessment of learned leadership skill must utilize an assessment tool that explicitly reflects the learning outcomes embedded in the intervention. While many popular leadership assessments instruments exist (Owen, 2001), many educators and practitioners often choose an assessment tool without first considering how this instrument relates to their desired program goals and outcomes. For example, requiring students to complete the Student Leadership Practices Inventory (SLPI) based on Kouzes and Posner's (1987) five exemplary leadership practices, when the learning being assessed only marginally relates to these practices is not an effective means to assess program effectiveness even if the SLPI is proven as a statistically valid and reliable tool for assessing these five practices (Posner, 2004). Therefore, leadership educators must be intentional in choosing or constructing an assessment tool that appropriately addresses their intervention to uphold the accuracy of their measurements of impact.

At the University of Illinois students who participate in programs sponsored by the Illinois Leadership® Center are asked to rate the change in their behavior only in areas the program covered. For example, IGNITE, a program designed to teach project management and effectiveness skills includes specific questions regarding participant behavior in the areas of systems thinking and leading change in groups. For these reasons, educators and practitioners need to be intentional about clearly identifying the focus of the intervention and related learning outcomes in order to align assessment measures to accurately gauge impact.

Participant versus Observer Reporting

It is important to also consider who will be participating in assessing each student's learning. Many leadership interventions collect self-report data from participants to learn more about their individual impressions of what they learned and retained and how they plan to implement it in their lives. While having participants complete self-report evaluations in this way is relatively straightforward and can be done without expending many resources, the results from these self-report assessments yield information more related to how students perceive the program, rather than the program's effectiveness. This is valuable data, to be sure, but self-report data is limited in its ability to measure how much change the program creates in participants.

In addition to having participants complete self-report assessments, one added alternative is to invite observers to share their impressions of participants' learned leadership skill through 360-degree feedback assessments. Using this assessment structure, participants compare their self-report results against the ratings of other observers (e.g., classmates, supervisors, friends, family) that they invite to complete an identical assessment about their leadership practices; all of the observer results are anonymous, providing participants with a more accurate reflection of how they are viewed as they put these leadership skills into practice. Through our work at the Illinois Leadership® Center using 360-degree feedback, we have found that this assessment method provides students and professional staff with a more effective and accurate way to interpret how they apply leadership skills and attributes in their daily lives compared to simple self-evaluations of their learning.

Measurement Timing

Another factor previously discussed with regard to the Honeymoon and Horizon Effects is measurement timing. When structuring an assessment plan, it is critical to consider *when* participants evaluate the program and their learning in addition to *how* the evaluation will be carried out. A popular practice within leadership education is having participants complete a survey or other assessment measure at the conclusion of the intervention when educators have a captive audience; this is generally done not only for convenience sake, but also because it tends to increase the response rate compared to waiting a short time after the intervention is complete. However, as touched upon when discussing the Honeymoon Effect, these immediate assessment measures may create biased and inaccurate results because participants have not yet had the opportunity to apply these leadership skills in their everyday environment to gauge how much they actually learned through the intervention.

One consideration to combat this issue is to include a follow-up assessment in addition to one immediately after the intervention in order to determine how well participants feel they have been able to apply the specific leadership skills taught in the program. At the University of Illinois at Urbana-Champaign, for example, students who participate in any of the university's leadership programs are asked to complete an online program assessment within hours of each program's completion. Three months later, all participants are asked to complete a much shorter assessment that includes questions about their ability to implement the learning goals of the program into their typical practice of leadership. This assessment structure of combining immediate feedback with a follow-up evaluation helps to increase the accuracy of the overall measurement of students' learned leadership skill. Educators may also want to critically consider the use of pre-tests in their assessment of program effectiveness, owing to the Horizon Effect, and use a then-post structure discussed earlier.

Areas for Future Research

Future related research could serve to inform several areas of leadership outcomes assessment discussed in this article. We will examine two areas here, as they have been relatively ignored in the field of higher education research in the past: achieving a better understanding of the impact and biases that exist within 360-degree leadership assessments and how one's leadership perspectives impact one's overall responses on leadership assessments.

As previously discussed, many leadership assessment professionals suggest that a 360-degree assessment can be superior to a simple individual measure (Owens, 2001; Posner, 2004). However, owing to both the Halo and potentially the Hallmark Effects, achieving accuracy might be an issue in some contexts. Conducting further research in how 360-degree requirements are completed by differing populations might help to better inform leadership educators in how to best achieve their assessment goals. Are there differences by gender, race, or age in how individuals rate the leadership efficacy and skills of others? How do family members differ from friends and co-workers in responding to assessments with regard to how they rate the individuals they are assessing? Such research might help educators create accepted protocols or suggested procedures when instructing students in who they should request a 360-degree from.

In addition, little is known about how one's own sense of leadership efficacy and belief in the "hallmark" of leadership affects one's responses on assessments of learned leadership skills. Clearly, if students feel they do not have the skill to lead, they will rate themselves lower than students who feel they are more effective, even if their observed actions are similar across groups. Therefore, more research should be conducted in studying how self-perception of leadership

efficacy affects a skill-based or behavioral assessment score. What factors might minimize the difference in scores? How prominent should leadership be as a descriptor of the assessment? How explicitly described should behaviors be on these assessments? Little is currently known in attempting to answer these questions.

Achieving assessment validity in the field of leadership education is both a necessary and attainable goal. Moreover, as the assessment of learned leadership skill becomes increasingly popular and necessary to demonstrate student learning, leadership educators should possess an understanding of the potential issues and pitfalls as they try to reach this goal. To this end, this article is intended as a review of some of the common challenges that leadership educators should consider and become familiar with when assessing students' learned leadership skill to ensure increased accuracy in such measurements.

References

Arminio, J. L., Carter, S., Jones, S. E., Kruger, K., Lucas, N., & Washington, J. (2000). Leadership experiences of students of color. *NASPA Journal, 37*(3), 496-510.

Astin, H. S. (1996). Leadership for social change. *About Campus, 1*(4), 4-10.

Astin, A. W., & Astin, H. S. (Eds.). (2000). *Leadership reconsidered: Engaging higher education in social change*. Battle Creek, MI: W. K. Kellogg Foundation.

Astin, H. S., & Cress, C. (1998). *The impact of leadership programs on student development*. UCLA-HERI Technical Report to the W. K. Kellogg Foundation. Battle Creek, MI: W. K. Kellogg Foundation.

Atkins, P. W., & Wood, R. E. (2002). Self- versus others' ratings as predictors of assessment center ratings: Validation evidence for 360-degree feedback programs. *Personnel Psychology, 55,* 871-904.

Atwater, L E., & Yammarino, F. J. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Personnel Psychology, 45,* 141-164.

Barker, R. A. (1997). How can we train leaders if we do not know what leadership is? *Human Relations, 50*(4), 343-362.

Bennis, W. G., & Nanus, B. (1985) *Leaders: The strategies for taking charge*. New York: Harper & Row.

Burns, J. M. (1978). *Leadership*. New York: Harper & Row.

Chambers, T. (1992). The development of criteria to evaluate college student leadership programs: A Delphi approach. *Journal of College Student Development, 22*(4), 339-347.

Cress, C., Astin, H. S., Zimmerman-Oster, K., & Burkhardt, J. C. (2001). Developmental outcomes of college students' involvement in leadership activities. *Journal of College Student Development, 42*(1), 15-26.

Day, D. V. (1991). Leadership development: A review in context. *Leadership Quarterly*, *11*(4), 581-613.

Dugan, J. P., & Komives, S. R. (2007). *Developing leadership capacity in college students: Findings from a national study.* A Report from the Multi-Institutional Study of Leadership. College Park, MD: National Clearinghouse for Leadership Programs.

Eich, D. (2008). A grounded theory of high-quality leadership programs: Perspectives from student leadership development programs in higher education. *Journal of Leadership & Organizational Studies*, *15*(2), 176-187.

Engemann, K. M., & Owyang, M. T. (2005). So much for that merit raise: The link between wages and appearance. *The Regional Economist, April 2005*. Retrieved January 3, 2009 from http://www.stlouisfed.org/publications/re/2005/b/pages/appearances.html

Evertson, C. M., & Green, J. L. (1986). Observation as inquiry and method. In M. C. Wittrock (Ed.), *Handbook of research on teaching*. New York Macmillan.

Grove, J. T., Kibel, B. M., & Haas, T. (2005). EVALULEAD. *A guide for Shaping and evaluating leadership development programs*. Oakland, CA: Sustainable Leadership Initiative. Public Health Institute.

Heifetz, R. A. (1994). *Leadership without easy answers*. Cambridge, MA: The Belknap Press of Harvard University Press.

Hill, L. G., & Betz, D. L. (2005). Revisiting the Retrospective Pretest. *American Journal of Evaluation, 26(4)*, 501-517.

Kezar, A., & Moriarty, D. (2000). Expanding our understanding of student leadership Development: A Study Exploring Gender and Ethnic Identity. *Journal of College Student Development, 41(1),* 55-69.

Kirkpatrick, D. L (1996). *Evaluating Training Programs: The Four Levels.* San Francisco: Berrett-Koehler.

Komives, S. R., Lucas, N., & McMahon, T. R. (2007). *Exploring leadership: For college students who want to make a difference (2nd ed.)*. San Francisco: Jossey-Bass.

Kouzes, J. M., & Posner, B. Z. (1987). *The leadership challenge*. San Francisco: Jossey-Bass.

National Clearinghouse for Leadership Programs. Retrieved December 15, 2008 from http://www.nclp.umd.edu/resources/curricular_programs.asp

Owen, J. E. (2001). *An examination of leadership assessment*. Leadership Insights and Applications, Series #11. College Park, MD: National Clearinghouse for Leadership Programs.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: Volume 2, a third decade of research.* San Francisco: Jossey-Bass.

Posner, B. Z., & Brodsky, B. (1992). A leadership development instrument for college students. *Journal of College Student Development, 33*(3), 231-237.

Posner, B. Z. (2004). A leadership development instrument for students: Updated. *Journal of College Student Development, 45(4),* 443-456.

Riggio, R. E., Cicilla, J., & Sorenson, G. (2003). Leadership education at the undergraduate level: A liberal arts approach to leadership development. In S. E. Murphy & R. E. Riggio (Eds.). *The future of leadership development* (pp. 223-236). Mahwah, NJ: Lawrence Erlbaum.

Roberts, D. C. (2003). Crossing the boundaries in leadership program design. In C. Cherrey, J. J. Gardiner, & N. Huber (Eds.). *Building leadership bridges 2003* (pp. 137-149). College Park, MD: International Leadership Association.

Rohs, F. R. (2002). Improving the evaluation of leadership programs: Control response shift. *Journal of Leadership Education, 1*(2), 50-61.

Rost, J. C. (1991). *Leadership for the twenty-first century*. New York: Praeger.

Rost, J. C., & Barker, R. A. (2000). Leadership education in colleges: Toward a 21st century paradigm. *Journal of Leadership and Organizational Studies, 7*, 3-12.

Schertzer, J. E., & Schuh, J. H. (2004). College student perceptions of leadership: Empowering and constraining beliefs. *NASPA Journal, 42*(1), 111-131.

Schwartz, M. K., Axtman, K. M., & Freeman, F. H. (1998). *Leadership education source book (7th ed.)*. Greensboro, NC: Center for Creative Leadership.

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pretests. *American Journal of Evaluation, 30(1)*, 31-43.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Uleman, J. S. (1991). Leadership ratings: Toward focusing more on specific behaviors. *Leadership Quarterly, 2*(3), 175-187.

Upcraft, M. L., & Schuh, J. H. (1996). *Assessment in student affairs: A guide for practitioners.* San Francisco: Jossey-Bass.

W. K. Kellogg Foundation (1999). *Building leadership capacity for the 21*st *century: A report from global leadership scans.* Battle Creek, MI: W. K. Kellogg Foundation.

Zimmerman-Oster, K., & Burkhardt, J. C. (1999). *Leadership in the making: Impacts and insights from leadership development programs in U.S. colleges and universities.* Battle Creek, MI: W. K. Kellogg Foundation.

Biographies

David M. Rosch, Ph.D.**,** is the Assistant Director of the Illinois Leadership® Center at the University of Illinois at Urbana-Champaign, where his responsibilities include coordinating the research, assessment, and evaluation efforts of leadership programs. He also teaches leadership classes and serves on the university-wide Illinois Leadership Coordinating Committee. Rosch received his bachelor degree from Binghamton University, his masters from Colorado State University, and his doctorate from Syracuse University.

Leslie M. Schwartz, M.Ed.**,** is the Program Manager of the Illinois Leadership® Center at the University of Illinois at Urbana-Champaign, where she coordinates the campus-wide Leadership Certificate and Consulting Programs. She also manages the Center's i-programs that focus on teaching specific leadership skills and attributes to undergraduate and graduate students. Schwartz has a bachelor degree in Communications as well as a masters degree in Student Affairs from UCLA.