I. WHY PRINCIPAL PROGRAM EVALUATION IS IMPORTANT

Truly improving the quality of school leaders requires improving the way principals are prepared and selected (Wallace, 2012). Program evaluation can serve many purposes: supporting program improvement, communicating the value of the program to internal and external stakeholders, and holding the program accountable for results. The ability to measure and report impact on outputs (such as placement rates) and outcomes (such as impact on school practices, teacher effectiveness and student achievement) is particularly relevant for programs that want to hold themselves accountable to reaching higher outcomes, while also supporting those that have strong external expectations (from associations, funders, constituents, or the government) for results and improvement.

Many programs, however, have not adequately invested in program evaluation—and when they have the evaluation plans are often weak. Sanzo (2009), in a review of federally funded School Leadership Program grant evaluation plans, found that most evaluations of leadership programs lacked specificity in evaluation tools, data sources, analyses, and many did not have a robust theory of action to guide the evaluation. She concludes with a call for well-planned program evaluations that can provide evaluative feedback to the program as well as inform the field at large.

This paper is part of a series that describes lessons learned by New Leaders in conducting its own program evaluation. This paper describes the New Leaders approach—which is based on principles of high quality program evaluation—and how evaluation results have informed continuous learning and improvement.

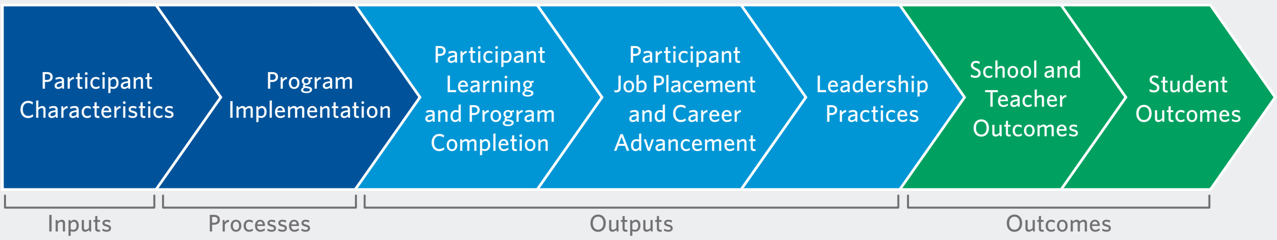
PRINCIPLES OF HIGH QUALITY PROGRAM EVALUATION APPLIED TO PREPARATION PROGRAMS

High quality program evaluation—no matter the subject matter—examines inputs, processes, outputs, and outcomes or impacts (Rossi, Lipsey & Freeman, 2004; Wholey, Hatry & Newcomer, 2010). Both the Education Development Center (EDC) and the University Council of Educational Administration (UCEA) has created a conceptual frameworks for leadership preparation programs that describes the elements that should be measured and possible sources of evidence for measuring them.[1]

1. For more information on EDC's framework and accompanying tools, see http://www.wallacefoundation.org/knowledge-center/school-leadership/principal-training/Pages/Quality-Measures-Principal-Preparation-Program-Assessment.aspx. For more information on UCEA's framework and accompanying tools, see: http://www.edleaderprep.org. All UCEA programs have a robust program evaluation in place, as required by membership in UCEA.

The following diagram is an adaptation of the UCEA Evaluation Pathway framework that New Leaders uses to guide its program evaluation. The accompanying papers in this series are organized by this framework, providing in-depth lessons and recommendations regarding: participant characteristics, program implementation, participant learning and program completion, participant job placement and career advancement, and student outcomes.

Figure 1. Evaluation Pathway for Leadership Preparation Programs (adapted from [UCEA, 2013]):



This paper series does not address leadership practices or school and teacher outcomes. While the New Leaders external evaluation by the RAND Corporation does address these components by leveraging our Memorandums of Understanding with districts to access district school climate survey data and to conduct third-party interviews of district partners, they are not addressed by our internal program evaluation. New Leaders has attempted to gather data on leadership practices in the past. However, we found that alumni and teacher survey response rates were too low to generate useful information and that staff visits to schools were too expensive to sustain. We have not been able to gain access to district principal and teacher effectiveness data and have found that publically available school survey data is often unavailable or inconsistent from year to year.

HOW PROGRAM EVALUATION CAN SUPPORT PROGRAM IMPROVEMENT

By tracking all of the elements in the evaluation pathway (inputs, processes, outputs, and outcomes), a well-designed program evaluation can provide formative feedback that informs single loop learning as well as summative findings that inform double loop learning.

Single loop learning includes improved decision-making about program execution based on fidelity of implementation of the program (Argyris & Schön, 1978). For example, New Leaders faculty members immediately use session survey results to inform their planning for sessions the following month (e.g., whether to reteach a particular topic). Recruiting staff use admissions data from previous years to identify gaps (e.g., not enough secondary applicants) and target recruiting efforts accordingly.

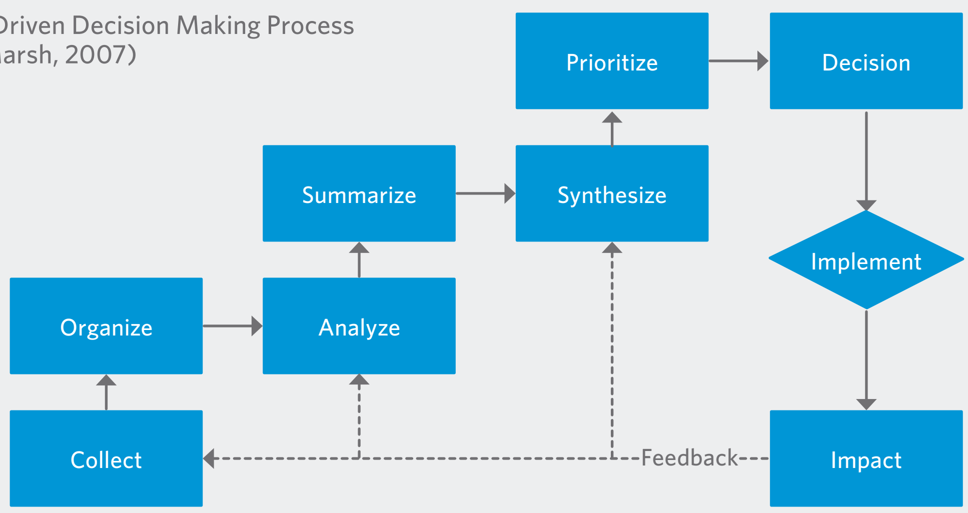
Double loop learning involves improving program design based on using evaluation results to examine the values and assumptions underlying the program design (Argyris & Schön, 1978). For example, New Leaders used results from an external evaluation of impact on student outcomes to question and change its underlying theory of action. While results of the external RAND evaluation have shown statistically significant positive program effects (Gates et al, 2014), New Leaders consistently seeks to improve the consistency and magnitude of our effect. New Leaders therefore launched a series of efforts to gather more information to help understand and explain the student achievement results. It examined the results of a series of RAND case studies of early tenure principals (Burkhauser et al, 2012), which included an analysis of the relationship between the experiences of those principals and the subsequent school-level effects on student achievement. The organization also ran several analyses of existing data to see if there were any correlations between student outcomes and the characteristics of participants, program elements, or the schools led by New Leader principals. The organization also systematically gathered input and insights from coaches working directly with the principals.

New Leaders learned that the principals who were leading the most successful results were principals who could effectively lead adults. While New Leaders principals demonstrated consistent strengths in instructional leadership, the investigation also suggested that the national admissions process was not effectively screening for these adult leadership skills and coaches were not able to adequately remediate weak adult leadership skills once principals were in placed in the principalship. The organization concluded that it needed to redesign its program to better select and train for adult leadership. It has largely replaced its national admissions process with a teacher leadership program—called the Emerging Leaders Program (ELP)—that gives program participants additional authentic practice leading adults and provides them with feedback on those skills. This design not only develops program participant skills, it also generates scores on the authentic practice-based assignments that enable New Leaders to better assess whether someone is ready to be admitted to the Aspiring Principals Program (APP).

USING EVALUATION DATA TO INFORM IMPROVEMENT

Evaluations provide important information, but information and data alone are insufficient for organizational learning. Programs transform data and information into knowledge when they synthesize, interpret, and triangulate it with other information. This knowledge produces improvement when it informs decisions about design and implementation. Continuous improvement occurs when the cycle repeats itself.

Figure 2. Data-Driven Decision Making Process (Ikemoto and Marsh, 2007)



New Leaders ensures that program evaluation informs continuous improvement by dedicating staff resources to analyze and summarize data. New Leaders staff produce analytical reports that are then reviewed by various stakeholders (including program designers and implementers) during meetings (which we call "step back" meetings) at least two to three times per year for each program. During these meetings, program implementers help to identify root causes of patterns observed in the data and the group strategizes on how to make improvements accordingly. For example, during a step back meeting, staff discussed low assignment completion rates. Coaches reported that the assignments were taking participants much longer to complete than the designers had planned. The program therefore reduced the number of assignments for the following year.

RECOMMENDATIONS BASED ON LESSONS LEARNED

Recommendations based on New Leaders' experiences regarding using evaluation to inform improvement are:

- **Clarify your program's theory of action**. The theory of action underlying the program needs to be clear in order to effectively identify the indicators that are important to track.
- Ensure your program evaluation **address the entire evaluation pathway**, including inputs, processes, outputs, and outcomes.
- **Generate hypotheses**. In addition to tracking current beliefs in the theory of action, identify factors that *might* be influencers and track them so that you can empirically test those hypotheses in future years.
- **Address single loop and double loop learning** needs by examining fidelity of implementation and questioning program design assumptions.
- **Create processes to translate evaluation data** into organizational learning and program improvement, such as step back meetings.

Other papers in this series provide more detail on New Leaders lessons learned evaluating inputs, processes, and outcomes. For more information, click here.

REFERENCES

Argyris, C. (1991). Teaching smart people to learn. *Harvard Business Review*, 4(2), 4-15.

Argyris, C., & Schon, D. (1978). *Organizational Learning: A theory of action approach*. Reading, MA: Addision Wesley.

Burkhauser, S., Gates, S. M., Hamilton, L., & Ikemoto, G. S. (2012). First-year principals in urban school districts: How actions and working conditions relate to outcomes. Santa Monica, CA: RAND Corporation. Available on-line: http://www.rand.org/pubs/technical_reports/TR1191.html

Buskey, F., & Karvonen, M. (2012). Evaluating innovative leadership preparation: How what you want drives what (and how) you evaluate. *Journal of Leadership Education*, 11(1).

Darling-Hammond, L., LaPointe, M., Meyerson, D., Orr, M. T., & Cohen, C. (2007). *Preparing school leaders for a changing world: Lessons from exemplary leadership development programs*. Stanford, CA: Stanford University, Stanford Educational Leadership Institute.

Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, DC: Albert Shanker Institute.

Farkas, S., Johnson, J., Duffett, A., & Foleno, T. with Foley, P. (2001). *Trying to stay ahead of the game: Superintendents and principals talk about school leadership*. New York, NY: Public Agenda.

Farkas, S., Johnson, J., & Duffet, A. with Syat, B. & Vine, J. (2003). *Rolling up their sleeves: Superintendents and principals talk about what's needed to fix public schools*. New York, NY: Public Agenda.

Gates, S. M., Hamiilton, L. S., Martorell, P., Burkhuser, S., Heaton, P., Pierson, A., . . . Gu, K. (2014). *Preparing principals to raise student achievement: Implementation and effects of the New Leaders program in ten districts*. Santa Monica, CA: RAND. Available on-line: http://www.rand.org/pubs/research_reports/RR507.html

Ikemoto, G. S., & Marsh, J. A. (2007). Cutting through the "data-driven" mantra: Different conceptions of data-driven decision making. In P. A. Moss (Ed.), *Evidence and Decision Making* (Vol. 106, pp. 105-131). Malden, MA: Blackwell Publishing.

Johnson, J (2007). *A mission of the heart: what does it take to transform a school?* New York, NY: Public Agenda. Available on-line: http://www.publicagenda.org/files/missionheart.pdf

King, C. (2013). *Quality measures: Principal preparation program self-assessment toolkit*. Waltham, MA: Education Development Center, Inc.

Levine, A. (2005). *Educating school leaders*. Washington, DC: The Education Schools Project.

McCarthy, M. (1999). The evolution of educational leadership preparation programs. In J. Murphy & K.S. Louis (Eds.), *Handbook of Research on Educational Administration: A Project of the American Educational Research Association*. San Francisco: Jossey-Bass.

Murphy, J. & Vriesenga, M. (2004). *Research on preparation programs in educational administration: an analysis*. Columbia, MO: University of Missouri-Columbia, University Council for Educational Administration.

Murphy, J., Moorman, H., & McCarthy, M. (2008). A framework for rebuilding initial certification and preparation programs in educational leadership: Lessons from whole-state reform initiatives. *Teachers College Record*, 110(10), 2172-2203.

Orr, M. T., & Barber, M. E. (2005). *Collaborative leadership preparation: A comparative study of innovative programs and practices*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach*, 7th ed. Thousand Oaks, CA: Sage.

Sanzo, K. L. (2009). An analysis of evaluation plans in a federally funded leadership preparation program: Implications for the improvement of practice. *Planning and Changing*, 43(3/4), 243-259.

The Wallace Foundation. (2012). *The making of the principal: Five lessons in leadership training*. New York: Author.

University Council for Educational Administration. (2013). *Developing evaluation evidence: A formative and summative evaluation planner for educational leadership preparation programs*. Charlottesville, VA: Author.

Wholey, J. S., Hatry, H. P., & Newcomer, K. E. (2010). *Handbook of practical program evaluation* (3 ed.). San Francisco, CA: Jossey-Bass Publishers.

ADDITIONAL RESOURCES ON PROGRAM EVALUATION:

Designing Evaluations by the U.S. Government Accountability Office: http://www.gao.gov/assets/590/588146.pdf

W.K. Kellogg Foundation Evaluation Handbook: https://www.wkkf.org/resource-directory/resource/2010/w-k-kellogg-foundation-evaluation-handbook

University of Wisconsin-Extension, Program Development and Evaluation: http://www.uwex.edu/ces/pdande/evaluation/index.html

User-friendly Handbook for Project Evaluation, National Science Foundation: http://www.nsf.gov/pubs/2002/nsf02057/start.htm

United Way Outcome Measurement: http://www.yourunitedway.org/outcome-measurements

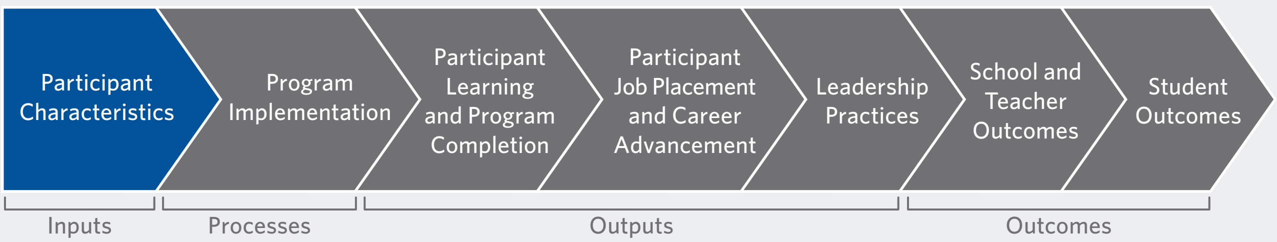
Harvard Family Research Project: http://www.hfrp.org/evaluation

McNamara, C. "Basic Guide to Program Evaluation (Including Outcomes Evaluation) Adapted from the Field Guide to Nonprofit Program Design, Marketing and Evaluation." http://managementhelp.org/evaluation/program-evaluation-guide.htm#anchor1575679

2. TRACKING PARTICIPANT CHARACTERISTICS

High quality evaluation of principal preparation programs examines inputs, processes, outputs, and outcomes. These components are depicted in the evaluation pathway framework below.

Evaluation Pathway for Leadership Preparation Programs (adapted from UCEA, 2013):



This paper is part of a series documenting lessons learned by New Leaders in utilizing this framework to guide internal evaluation of its Aspiring Principals Program (APP). The paper focuses on the first step in the pathway, participant characteristics.

Program participants bring with them a range of demographic characteristics, background experiences, and knowledge, skills, and beliefs—all of which can affect how they perform in training and the principalship (Kottkamp, 2011).

One category of participant characteristics is demographics, such as gender, age, race and ethnicity. Indicators of participant demographics serve multiple purposes in program evaluation. First, demographic variables can be used as controls in predictive analyses. Second, information about demographics can also be used to enable programs to reflect on whether the program design and implementation is meeting the needs of particular demographic groups. For example, programs are likely to be interested in exploring whether women (or older candidates or Hispanic candidates) are more or less likely to complete the program, get placed as a principal, and/or have positive impact on achievement. Based on that information, the program might decide to further target its program towards a particular type of participant, or it might decide that it wants to diversify the program to include a broader diversity of participants. For example, New Leaders serves schools that serve high minority populations—and one of our design assumptions is that minority students benefit when they are led by principals who reflect the student population. As a result we use demographic data to inform recruitment strategies.

A second category of participant characteristics is their background experiences, such as years of teaching experience, years of school leadership experience, type and source of undergraduate degree, and organizational affiliation. Program evaluations should longitudinally track background experience that are used as part of selection criteria because these criteria should reflect design assumptions regarding the prior experiences that are most relevant to success in the program. For example, when a program requires 5 five years of classroom teaching experience, that indicator suggests a program assumption that teaching experience is important. Programs should also consider hypotheses it has about characteristics that *might* matter and track those as well to enable empirical study. For example, New Leaders is examining whether prior experience in the Assistant Principal role predicts success, particularly in secondary schools. If it does, New Leaders may decide to include that indicator in its selection criteria going forward.

A third set of participant characteristics is their prior knowledge, skills, and beliefs. The specific indicators should reflect the program assumptions and selection criteria. Although some programs use standardized tests, such as the GRE, New Leaders conducts its own assessments as part of its admissions process. We have two admissions pathways: 1) National Admissions Process (which includes an application, admissions activities, and a full day of interviews), and 2) the Emerging Leaders Program (which is a program for teacher leaders through which program participants can demonstrate admissions criteria on performance assessments). The National Admissions pathway uses a combination of written application, phone interview, and assessment center activities. In the other pathway, New Leaders uses assignment scores from assignments given to participants in the New Leaders Emerging Leaders Program for teacher leaders. Both pathways are designed to measure participant knowledge, skills, and beliefs related to our selection criteria: personal leadership, instructional leadership, culture leadership, and adult and team leadership. These criteria are based on principal standards as well as internal and external research on effective principal practice.

THE NEW LEADERS APPROACH TO TRACKING PARTICIPANT CHARACTERISTICS

The following chart summarizes the indicators and data sources that New Leaders uses to track the three sets of participant characteristics.

Participant Characteristics	Indicators Used by NL	Data Sources Used by NL
Demographics	• Age • Sex • Race/Ethnicity	Eligibility Form (part of initial written application)
Background experience	• Years of school-based experience (teaching, administrative, other) • Kinds/type of school experience (current and prior school levels, school type, subjects taught); Affiliations (professional organizations, training program completed, etc). • Certification type • Degrees (type, GPA, school)	Eligibility Form (part of initial written application)

Participant Characteristics	Indicators Used by NL	Data Sources Used by NL
Knowledge and skills	NL Selection standards: **Personal Leadership** • Reflective Practice and Continuous Improvement • Communication and Interpersonal Relationships • Vision and Mission **Instructional Leadership:** • Pedagogy/Instructional Strategies • Data Driven Instruction **Culture Leadership** • Urgency and Efficacy **Adult and Team Leadership** • Diagnostic and Strategic Planning • Performance Management • Leadership Development	National Admission scores • Admissions Activity Scores • In Person Interview Emerging Leader Program scores • Assignment scores • Learning Meeting scores • Teacher surveys

RECOMMENDATIONS BASED ON LESSONS LEARNED

Recommendations based on New Leaders' experiences regarding tracking participant characteristics are:

- Make **design assumptions** about participant characteristics explicit. For example, is your program designed to meet the needs of a particular type of participant? What background experiences, knowledge, and skills are necessary prerequisites to successful participation in the program?
- **Generate hypotheses** regarding what characteristics are likely to be correlated with success. For example, if participants with particular work experiences (such as serving as an instructional coach) are expected to be relevant to success, it is important to track that information during the admissions process to be able to test that hypothesis later.
- **Maximize variability in scores** (or scoring options) when developing tools to measure participants knowledge and skills, in order to better differentiate between candidates during selection, and to allow for later analysis of whether particular competencies are more important than others. For example, New Leaders expanded its rubric scores to a seven-point scale to increase variability in the scores.
- **Invest in data systems** with the necessary identifiers that will allow participant characteristics to be linked with other data sources and outcomes. Ensure that qualities and characteristics of interest are consistently tracked over time. Create a unique identifier for each program participant so that their participant characteristics can be linked to other program data, such as assessment scores, program completion/ licensure, and placement data. Collect application data using online forms with drop-down menus that allow information to be easily converted into a database with comparable data points across participants.

REFERENCES

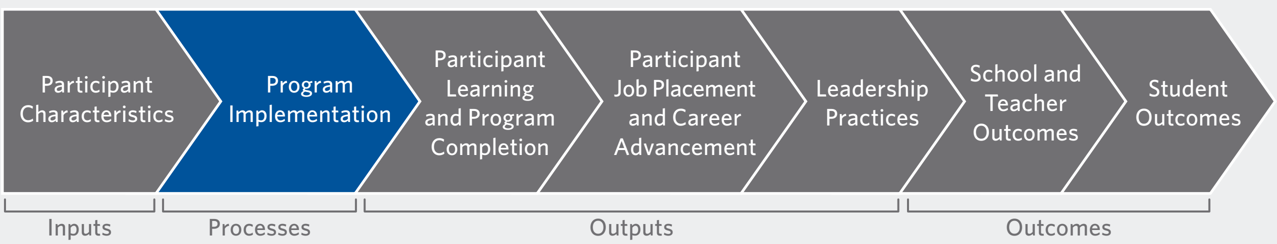
Kottkamp, R. (2011). Introduction: Leadership Preparation in Education. *Educational Administration Quarterly* 47:3.

University Council for Educational Administration. (2013). *Developing evaluation evidence: A formative and summative evaluation planner for educational leadership preparation programs*. Charlottesville, VA: Author.

3. MEASURING PROGRAM IMPLEMENTATION

High quality evaluation of principal preparation programs examines inputs, processes, outputs, and outcomes. These components are depicted in the evaluation pathway framework below.

Evaluation Pathway for Leadership Preparation Programs (adapted from UCEA, 2013):



This paper is part of a series documenting lessons learned by New Leaders in utilizing this framework to guide internal evaluation of its Aspiring Principals Program (APP). The paper focuses on the second step in the pathway, program implementation.

Program implementation studies examine whether the program was implemented as intended, including whether participants received the full dosage of the program "treatment" (i.e., program experiences and support) and with what level of quality. Measuring program implementation can: 1) clarify assumptions embedded in program design; 2) guide program improvements; and 3) help explain and contextualize the impact results (Resnick, et. al, 2005; O'Donnell, 2008; Century, et. al, 2010).

Program implementation evaluation should track components of the program that designers expect will have the most substantial effect on program goals.

Examples of typical principal preparation program activities include:

- Curriculum and coursework content (quality / alignment with standards)
- Content delivery (dosage and quality)
- Assignments (completion and quality)
- Coaching and feedback (dosage and quality)
- Internship/Residency (length and/or quality)
- Instructional strategies and practice (including dosage)
- Knowledge of faculty/instructors, including coaching (quality)
- Technical assistance, such as placement assistance (dosage and quality)

Evaluators often need to engage in conversations with program designers to clarify the program activities, including their intended dosage and characteristics of quality. This clarification process can help to make design assumptions explicit. For example, New Leaders coaches are expected to provide high quality feedback on authentic work products submitted (e.g., video of leader practice or work artifacts) without visiting schools. Although it can be time consuming to surface the theory of action and design assumptions, the process can be helpful to program designers. When program designers make their implicit theories explicit, they are often better able to clearly communicate the critical features to program implementers, which in turn can improve the quality of implementation. For example, New Leaders program design is grounded in socio-cultural learning theory, which suggests that one of the most important drivers of participant learning is high quality feedback on authentic opportunities for practice. In order to track whether this aspect of the program (high quality feedback) was being implemented, the program designers had to define high quality feedback. This work not only informed the program evaluation, it also informed training for the program instructors whose role it was to provide the feedback.

The evaluation should also pay careful attention to program features that are likely to vary in implementation—even if the provider does not fully control that feature—in order to examine the relationship between implementation and results. Implementation might vary due to different levels of staff capacity or due to different contexts. For example, New Leaders historically did not track the quality of the participant's residency site, even though it is a critical program feature, because it had limited control choosing the sites. However, the organization decided to start measuring and tracking resident site quality so that it could study whether and how that aspect of the program was related to program outputs, such as participant learning and program completion. It is important to monitor activities which may have variability so as to understand the program as precisely as possible and document possible patterns across sites, participants, or timeframes.

Program implementation often involves measuring both fidelity (e.g., was the program component implemented?) and quality (e.g., was the program component implemented well?). Measuring fidelity often involves creating dosage tracking systems, such as attendance at courses, assignment completion, and coaching logs. Quality, on the other hand, can be measured via participant perceptions (often captured via survey) and/or program quality review (which often entails rating a component against a rubric). New Leaders uses a set of surveys created by the University Council of Educational Administration (UCEA), called the INSPIRE Leadership Suite of program evaluation surveys, that programs can use to assess common program elements. These annual measures of program implementation can be particularly useful for examining the relationship between implementation and program outputs. The INSPIRE surveys provide the added benefit of being able to benchmark results to national norms. New Leaders also creates and administers its own interim surveys five to ten times throughout the year to monitor quality of implementation.

SUMMARY OF NEW LEADERS APPROACH TO MEASURING PROGRAM IMPLEMENTATION

The following chart summarizes the indicators and data sources that New Leaders uses to track implementation of its key program components.

Program Component	Indicator	Data Sources
Formal Learning Sessions	Session content quality	Participant session surveys; Participant comprehensive surveys (mid-year and end-of-year); Session evaluation by national staff
	Session facilitation quality	Participant session surveys; Participant comprehensive surveys (mid-year and end-of-year); Session evaluation by national staff
	Number of sessions delivered	Program calendar (national and site-specific); Participant participation tracker
	Participant attendance	Participant participation tracker
Residency	Quality of experience	Residency site assessment; Participant comprehensive surveys (mid-year and end-of-year)
Coaching	Quality of coaching	Participant comprehensive surveys (mid-year and end-of-year); observation ratings by managers
	Quantity of coaching	Coaching log (duration & type of support)

New Leaders conducts descriptive quantitative and qualitative analyses immediately upon availability of each new data source and makes results available to relevant program and city staff, often through facilitated data conversations that allow performance feedback to be turned into useful information for adjusting tactics to meet program goals. For example, a facilitated data conversation highlighted evidence that monthly assessments were consuming considerably more time than expected for participants to complete and for program staff to consistently score and provide quality, timely feedback. New Leaders therefore decided to reduce the number of assessments, which in turn yielded more thorough completion from participants and more thoughtful, timely feedback from program staff.

The process of using data for performance feedback—i.e., rich cycles of data collection, analysis, and action—is deeply embedded in the work of all New Leaders staff. For example, local city teams meet at least once per month to assess APP participant and school progress, and to strategize for immediate improvement. Executive Directors from each city meet quarterly with national staff to analyze targeted data streams and to share learning lessons from across program sites, making adjustments for the current year's participants and for future trainees. Finally, each programmatic area holds at least two step backs each year to review data, identify lessons learned, and identify implications for planning for the next year.

RECOMMENDATIONS BASED ON LESSONS LEARNED

Recommendations based on New Leaders' experiences measuring program implementation are:

- Ensure that program implementation evaluation is a **collaborative process** by encouraging program staff to provide input into evaluation tools. For example, New Leaders staff responsible for delivering the Aspiring Principals Program help determine the content of New Leaders-developed surveys. This ensures that participants are responding to items that clearly connect to the program initiatives and hypotheses. Collaboration not only increases the accuracy of the data tools, but also allows program staff to see the value of investing time and energy in collecting the data— -- which in turn increases the likelihood that the data collected may yield information that will inform effective programmatic changes.

- Improve **survey response rates** by administering surveys (of all kinds) during training. We recommend using an online survey platform, but beware that this approach necessitates participants in having internet devices and access during training. It is also a good idea to have back-up paper survey versions available for improved response rates.

- When using **survey administration software**, ensure that the program or service offers suitable functionality to the tasks at hand. For example, New Leaders uses Qualtrics to administer surveys, as it allows for confidential survey responses, as well as the linking of responses to key respondent data (such as demographic data). It also offers a larger variety of survey question types (such as matrix questions and sliding scale questions), and allows for skip or display logic to be used based on previous responses or respondent demographic data. The reporting function also offers several options for filtering and sharing results. The software is also able to translate the data into our core database (Salesforce) for storage and analysis.

- **Examine results regularly** (at least annually and more often if possible). Use results to consider whether and how fidelity of implementation can improve (single loop learning) AND to consider whether and how to adjust the theory of action (double loop learning).

REFERENCES

Blase, K. and Fixsen, D. (2013). Core intervention components: *Identifying and operationalizing what makes programs work, ASPE Research Brief*. Office of Human Services Policy, US Department of Health and Human Services: Washington, DC

Century, J., Rudnick, M., and Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*. 31(2): 199-218.

O'Donnell, C. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in k –12 curriculum intervention research. *Review of Educational Research*. 78(1): 33-84.

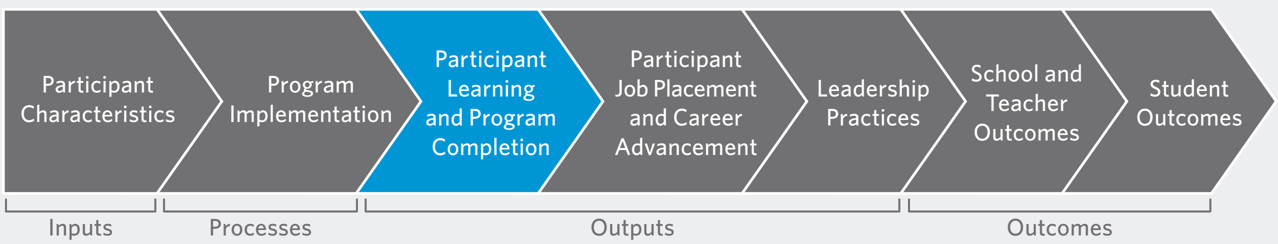
Resnick, B., Bellg, A. J., Borrelli, B., DeFrancesco, C., Breger, R., Hecht, J., et al. (2005). Examples of implementation and evaluation of treatment fidelity in the BCC studies: Where we are and where we need to go. *Annals of Behavioral Medicine*, 29, 46–54.

University Council for Educational Administration. (2013). *Developing evaluation evidence: A formative and summative evaluation planner for educational leadership preparation programs*. Charlottesville, VA: Author.

4. ASSESSING PARTICIPANT LEARNING AND PROGRAM COMPLETION

High quality evaluation of principal preparation programs examines inputs, processes, outputs, and outcomes. These components are depicted in the evaluation pathway framework below.

Evaluation Pathway for Leadership Preparation Programs (adapted from UCEA, 2013):



This paper is part of a series documenting lessons learned by New Leaders in utilizing this framework to guide internal evaluation of its Aspiring Principals Program (APP). The paper focuses on the third step in the pathway, participant learning and program completion.

Assessing candidates to determine whether they have learned the material taught by the program should be an important element of any leadership development program. Although assessments that take the form of a "test" can effectively measure latent knowledge, performance-based assessments that require participants to demonstrate their ability to enact effective leadership practices in authentic setting are preferable for determining whether participants have acquired leadership skills.

All of New Leaders' assessments are performance-based. They consist of three unique assignments that participants carry out two times each in their residency site twice for a total of six assignments. For example, one assignment is to conduct a 10-15 minute mini-observation of one of their primary teachers, collecting literal notes and student actions. Residents then analyze their notes, prepare and hold a coaching session where they provide feedback to the observed teacher, and debrief the coaching session with a faculty member. These assessments serve formative and summative purposes. The local program director provides participants with detailed feedback during learning meetings to enable participants to continuously improve. The assignments and learning meetings are designed to provide evidence related to each of the program standards and competencies (which are listed in the next section of this paper). All of the scores from assignments are factored into summative decisions regarding whether a participant should be endorsed for the principalship, but scores from the second implementation of the assignments are double weighted to encourage and value growth.

In the past, New Leaders has utilized different assessments for formative and summative purposes. We found, however, that this approach overly burdened the program participants and the directors charged with scoring. We also narrowed the assignments from 12 assignments down to three assignments. The three assignments are carried out twice for a total of six assignments, but the smaller number of assignments has enabled more rigorous training and norming for scoring.

SUMMARY OF NEW LEADERS APPROACH TO ASSESSING PARTICIPANTS

The following chart summarizes the indicators and data sources that New Leaders uses to track the three sets of participant characteristics.

Participant Learning and Program Completion	Indicators Used by NL	Data Sources Used by NL
Participant Learning	NL Standards and Competencies • Personal leadership • Instructional leadership • Cultural leadership • Adult and team leadership	• Residency Assignments • Learning Meetings
Program Completion	Minimum requirements • Attend all national and local sessions, program meetings, and virtual sessions • Attend two monthly meetings with Program Director at residency site • Attend all learning meetings • Complete individualized learning plan • Complete all program assignments • Ensure completion of midyear and end-of-year 360 survey	Minimum requirement checklist
Endorsement for Principalship	Endorsement decision (typically 85% of program completers are endorsed as ready for the principalship). Decisions are based on: • Minimum requirements • Participant learning of competencies • School and teacher outcomes	Participant records (stored on Salesforce platform)

RECOMMENDATIONS BASED ON LESSONS LEARNED

Recommendations based on New Leaders' experiences regarding assessing program participants are:

- Make **design assumptions** about participant assessment explicit. For example, is growth or final scores a better measure of participants' potential as a principal?
- For any assessment, it is important to **maximize the variability in scoring** (especially in the case of rubric scoring) by 4 including scales that include at least points. New Leaders currently uses a rubric scoring framework with 7 seven points.
- To **ensure inter-rater reliability and consistency of assessment and scoring**, programs should provide thorough rater training and engage in continuous quality assurance activities. The process to norm scorers should include group norming activities, followed up with individual norming sessions, and accompanied by clear scoring guides. The levels of the rubric should be consistent across each standard and should include very objective, concrete language to facilitate consistent scoring. For purposes of quality assurance and feedback, scoring activities should also include compilation of evidence to support the numerical score. The evidence that connects the participants' behavior, skills, or knowledge to the particular score on the rubrics should be noted.
- Programs need to find a **balance between formative and summative assessment**. As with any assessment, over-reliance on summative assessments places a great deal of weight on end-of-year practices without providing sufficient feedback throughout the year to participants, and without the participants being set up for success by having experienced lower-stakes assessments. Conversely, an over-reliance on formative assessments does not provide a comprehensive snapshot of participants that would be useful for determining success in the program or be useful for communicating to potential employers. A balance between the two formats translates to dispersed workload (for participants and scorers), a consistent schedule of feedback, and multiple opportunities to learn and develop.
- **Invest in data systems** with the necessary identifiers that will allow programs to link participant assessment scores with other data sources and outcomes. New Leaders uses the Salesforce platform to store, manage and retrieve participant data.

REFERENCES

Condon, C., & Clifford, M. (2010). Measuring principal performance: How rigorous are commonly used principal performance assessment instruments? A Quality School Leadership Issue Brief *Learning Point Associates* (pp. 20). Naperville, IL. Available at: http://www.learningpt.org/pdfs/QSLBrief2.pdf

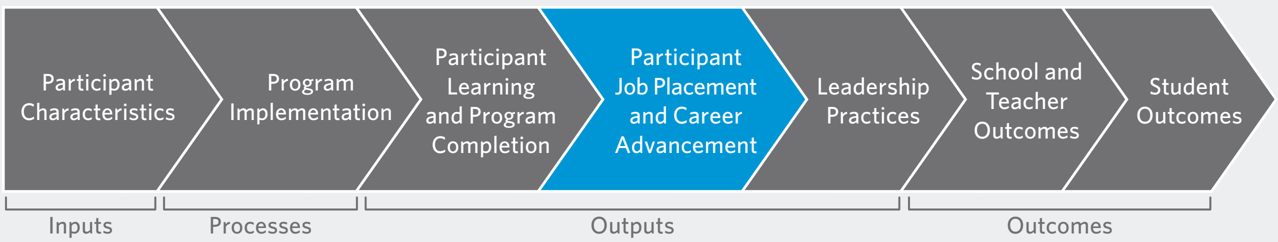
Harlen, W., Gipps, C., Broadfoot, P., & Nuttall, D. (1992). Assessment and the improvement of education. *The Curriculum Journal*, 3, 215–230.

University Council for Educational Administration. (2013). *Developing evaluation evidence: A formative and summative evaluation planner for educational leadership preparation programs*. Charlottesville, VA: Author.

5. TRACKING PARTICIPANT JOB PLACEMENT AND RETENTION

High quality evaluation of principal preparation programs examines inputs, processes, outputs, and outcomes. These components are depicted in the evaluation pathway framework below.

Evaluation Pathway for Leadership Preparation Programs (adapted from [UCEA, 2013](#)):



This paper is part of a [series](#) documenting lessons learned by New Leaders in utilizing this framework to guide internal evaluation of its Aspiring Principals Program (APP). The paper focuses on the fourth step in the pathway, participant job placement and career trajectory.

Tracking principal placements serves two important functions in leadership preparation program evaluations. First, since preparation programs are designed to prepare program participants for leadership roles, the extent to which graduates enter and are retained in those leadership roles is an important output of the program. Second, tracking principal placements is critical to measuring other outcomes like student achievement because it identifies which schools should be included in those analyses. Tracking placements can also help to maintain alumni relations.

Important variables for programs to track on graduates and alumni include:

- Type of role (whether at school or district level)
- Type of school
- Length of tenure
- Retention
- Length of time between program completion and placement in leadership position

Indicators and placement results should be aligned with the program's design assumptions. For example, the New Leaders program is designed and marketed to prepare graduates to immediately enter the principalship. Therefore, we might have higher expectations for immediate placement into the principal role than a program that is designed to prepare graduates to enter the assistant principal role before being promoted to the principalship.

Although we recommend that states track and make placement data available to preparation programs, these data currently are not typically available publicly. Our external evaluator, the RAND Corporation, collects placement information from our partner school districts, but this data acquisition process can be time consuming, the data are often inaccurate, and RAND does not collect the data every year from every partner (Burkhauser et al., 2012). We therefore have to rely on our own data collection. New Leaders delegates the task of tracking placements to local program offices. These local offices use a variety of tactics, but they often begin with district lists of principals and update those lists based on their ongoing communication with participants, surveys to participants, district websites, social media, and/or calling schools to confirm information. Although New Leaders is able to track placements for 92% of its graduates, this work requires a significant investment of staff time across the organization not only for initial updates at the beginning of the school year but also to continuously update over the course of the year.

SUMMARY OF THE NEW LEADERS APPROACH TO TRACKING GRADUATE PLACEMENT

The following chart summarizes the indicators and data sources that New Leaders uses to track graduate job placements. We also use these placements to then calculate retention rates.

Placement Variables	Indicators Used by NL	Data Sources Used by NL
Role	• Assistant Principal • District/CMO Superintendent • Formally Exited Program • Functional Principal • Non-Leader Education field • Other District/CMO Leader • Other field • Other Non-School Based Education Leader • Other School-Based Leader • Principal in Planning • Principal Manager • Principal of Record • Resident • Unemployed • Deceased • Unknown	Local program teams • State and district records • Graduate surveys • Phone calls to schools • Social media • Communication with graduates
Type of placement	• Charter vs. district • Grade levels served • Parent organization (District or CMO)	Publically available school data
Retention / Tenure	Analysis of role data over time	Salesforce

New Leaders tracks graduate and alumni career trajectory information in Salesforce. This electronic database allows the organization to build a comprehensive record of program graduate placements that is used to conduct analyses for various stakeholders. The database also allows New Leaders to merge placement data with other outcomes databases (like the longitudinal student achievement database).

RECOMMENDATIONS BASED ON LESSONS LEARNED

Tracking initial placement is critical for determining initial placement rates, but placements need to be continuously tracked over time to determine retention rates. Recommendations based on New Leaders' experiences regarding tracking participant job placement and career trajectory are:

- Make **design assumptions** about intended graduate job placement explicit. For example, does the program intend for graduates to immediately enter the principalship?
- **Invest in data systems** with the necessary identifiers that will allow programs to link participant placement and career trajectories with other data sources and outcomes.
- **Define the leadership roles that should be tracked**. Consider all the ways that you might want to analyze results (e.g., if you will want to determine how many graduates are in central office roles that manage principals vs. other types of central office roles - then the placement role categories should be defined accordingly). Ensure that any tracking mechanisms allow for capturing the subtleties/complexities of principal placement (e.g., functional principals, principals who are responsible for only a subset of grades).
- **Track the school and school type** in which graduates are placed (e.g., charter versus district, grade levels).
- **Establish consistent policies** regarding determining what leadership roles will be included in various analyses. For example, programs may wish to include program graduates who have assumed the role of principal while not the official principal of record (functional principal) in student achievement analyses, while excluding assistant principals (APs), as APs do not have the same sphere of influence in schools as a functional principal.
- Plan ahead for **strategies to maintain contact with program completers**:
  - Set expectations in a written contract that the program completers will provide updates on placement.
  - Set expectations with district/CMO partners to provide updates.
  - Strategize timing of updates: fall for up-to-date placement results; spring for school/student outcome results.

REFERENCES

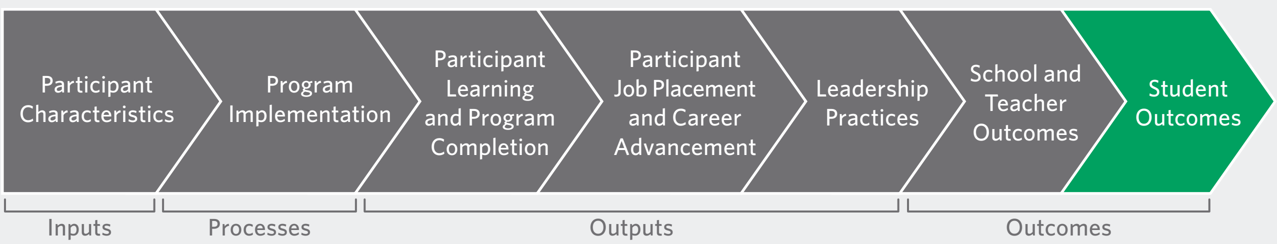
Burkhauser, S., Pierson, A., Gates, S. M., & Hamilton, L. S. (2012). *Addressing challenges in evaluating school principal improvement efforts*. Santa Monica, CA: RAND Corporation.

University Council for Educational Administration. (2013). *Developing evaluation evidence: A formative and summative evaluation planner for educational leadership preparation programs.* Charlottesville, VA: Author.

6. MEASURING STUDENT OUTCOMES

High quality evaluation of principal preparation programs examines inputs, processes, outputs, and outcomes. These components are depicted in the evaluation pathway framework below.

Evaluation Pathway for Leadership Preparation Programs (adapted from UCEA, 2013):



This paper is part of a series documenting lessons learned by New Leaders in utilizing this framework to guide internal evaluation of its Aspiring Principals Program (APP). The paper focuses on the final step in the pathway, student outcomes.

Using student outcomes (achievement test results, graduation, attendance, etc.) for evaluation of principal preparation program evaluation is a critical component of examining program outcomes. However, the use of student outcomes can be somewhat challenging because the data used for these analyses can be flawed and results can be difficult to interpret. In fact, student outcome analyses may not be feasible or appropriate for some programs due to lack of data availability or small sample sizes.

The primary challenge with student outcome results is that they can be an imperfect measure of a program's impact. Despite limitations, New Leaders has found that when results are interpreted with caution and they are used as part of a more comprehensive program evaluation, student outcomes can be an important and useful component of its program evaluation.

VALUE OF USING STUDENT OUTCOMES

As described in the first paper in this series, high-quality program evaluation includes examination of final outcomes. In the case of principal preparation programs, the final outcomes are student outcomes. New Leaders has experienced three major benefits from examining student outcomes:

1. **Internal accountability**. Focusing on student outcomes has helped New Leaders maintain a commitment to rigorous program design and execution. For example, like many programs, New Leaders experiences many different pressures from funders and partner district to endorse as many program participants as possible. Our commitment to assessing and reporting the student outcomes of schools led by endorsed New Leaders principals drives us to only endorse program completers who (according to their assessment scores) are likely to be successful in improving student achievement—even when withholding endorsement disappoints funders (who want to maximize the yield for their investments), district partners (who want as many qualified candidates for open principal positions as possible), and the program participants (who invested significant time and energy into the program).
2. **Program improvement**. While New Leaders has demonstrated statistically significant positive impact on student outcomes for its Aspiring Principals Program, our outcome analyses show substantial variation underneath that overall positive effect. These results became important starting points for further investigation into root causes for the variation in results. Those inquiry processes have generated important insights and learning lessons that have been the basis for major program redesign, such as a completely new approach to program admissions.
3. **Communicating the value of the program**. When student outcomes are positive, they can provide powerful messages to candidates and partners that the program is achieving its goals. Documenting positive student outcomes has helped New Leaders build support to sustain and expand programs.

RESPONSIBLE USE OF STUDENT OUTCOMES DATA

Users of student achievement analysis results should take care to understand the strengths and limitations of the data and always triangulate the results with additional data sources—especially prior to making high stakes decisions. Responsible use includes:

1. **Understanding the limitations of the data and analyses**. Great care is needed because the accuracy and validity of the data (as an indicator for program impact) can be uncertain. Some state tests are widely criticized regarding the extent to which they are a valid measure of whether students have learned the skills they need to be successful in college, careers, and citizenship. Even if tests are valid measures of what students know, school-level results may not be a valid measure of whether the school has improved student learning. For example, in one instance New Leaders' school-level analyses showed that a school had double digit losses in student achievement, but student-level results showed that the school made statistically significant gains. Further investigation revealed that the school-level losses were largely attributable to major shifts in student population caused by boundary line changes. Just like medical tests that have high false-positive or false-negative results can be quite useful when triangulated with other data, imperfect student outcome data can be quite useful to preparation programs when triangulated with other data, appropriately interpreted, and used in the manner described below.[1]

1. Note: this paper is intended for preparation program providers, not states or accreditation organizations. We recommend a much higher bar for using student outcomes data for high-stakes accountability, as described in our report *Change Agents: How States Can Develop Effective Leaders*

2. **Maximizing the validity of results** — to the extent possible—by:

    - **Examining changes in student outcomes**. Analyses that do not examine change (e.g., gains in test scores, improvements in graduation rates) are likely to be more descriptive of the where the program graduate was placed than their impact.
    - **Examining results after 3 year**s. Research conducted by our external partners (the RAND Corporation and Mathematica) and in the broader field suggests that the indirect nature of principal effects mean that they often are not measurable until the principal has three years of tenure in the same school.
    - **Setting minimum thresholds for sample size**. Programs may not be able to make valid interpretations of student outcomes analyses until programs reach a sample size of at least five to ten participants who have at least three years of usable data from the same school.
    - **Using student-level data when possible** because it allows for the most rigorous analyses. Student-level data provides an individual record for each student which includes: actual test scores; personal characteristics, such as free and reduced price lunch (FRPL) eligibility, English language learner (ELL) status, special education services, and ethnicity; it usually can be linked across years to create a longitudinal data set for individual students; and it can be "rolled up" to classroom- or school-level. Student-level data, however, are not publicly available and programs need to arrange for permission to access the data. Access might be costly if the district or state charges a fee for the data or has data safeguarding requirements that are costly to implement.

3. **Using multiple indicators and metrics, including those that are aligned with organizational goals**. While there are some basic measures that might be appropriate for all programs to use, programs may want to select additional measures based on their unique programmatic framework, mission, and goals. Programs should identify research questions and metrics that are aligned with their theory of how they expect the program to impact student achievement. For example, New Leaders mission and goals are very focused on closing the achievement gap. Program design is focused on that goal and therefore it is important for us to identify metrics that help us measure impact on achievement gap.

4. **Triangulating results with other data and information**. In addition to examining multiple indicators of student outcomes, users should also triangulate student outcomes with other data and information, such as principal supervisor perceptions of principal effectiveness and other data collected through the program evaluation regarding program inputs, implementation, and outputs.

These recommendations should improve the validity of student outcome results but the results are still likely to be less than perfect. Yet, they can still be valuable. Nearly all types of data, including survey data and assessment data, have limitations. Student outcome data is no different. The field of medicine routinely uses less than perfect data. For example, during annual exams, doctors monitor indicators of general health such as weight, blood pressure, and visual scans of eyes, ears, and mouth, etc. They seek additional information when an indicator is abnormal or changes substantially. For example, if a patient has high blood pressure, responsible doctors do not immediately assume that the patient has a blocked artery and schedule bypass surgery. Instead, before beginning any kind of medical intervention, doctors gather more information by asking questions about symptoms, checking the patient's' medical history, confirming family history, and possibly running more tests.

Likewise, imperfect student achievement results can be useful to programs. Individual results can be a flag for further investigation. Patterns over large samples and/or over time can be combined with other information to decide when program design changes are warranted. For example, when New Leaders gathered more information about program graduates with poor student outcomes results, we found a theme of weak skills related to adult leadership. This insight led us to design a new teacher leader program focused on adult leadership skills that participants complete prior to the Aspiring Principals APP, which gives them an additional year to develop and practice those skills.

SUMMARY OF NEW LEADERS APPROACH TO STUDENT OUTCOME EVALUATION

The New Leaders Aspiring Principals APP is designed to prepare principals to close the student achievement gap and improve overall student outcomes. The program is also designed to outperform other preparation programs by providing a high-quality preparation experience with a robust residency component.

New Leaders contracted with the RAND Corporation in 2006 to conduct a comprehensive evaluation of the implementation of the New Leaders program and its effects on student achievement and other outcomes, such as graduation rates and attendance rates. The RAND evaluation tracked program effects on student achievement outcomes through a value-add analysis of individual student achievement data provided directly to RAND by school districts.

Since RAND's analyses of student-level data take time and resources, New Leaders also has a robust internal program evaluation to provide quick-turnaround results that can inform program improvements and demonstrate effect. The New Leaders internal program evaluation tracks student achievement outcomes with school-level student achievement data that is publically available on state websites. On an annual basis, the internal research and evaluation department downloads the publicly available student achievement data for each district or state in which there are at last five principals who have been endorsed by New Leaders and who are currently serving as principals.

New Leaders uses these data to build a longitudinal database of student achievement and student learning gains data, including both New Leaders-led schools and other schools in our partner district and charter schools. This database enables New Leaders to map student achievement results with other programmatic data, including placement information, recruitment and selection data, and participant assessment data.

The table below provides examples of indicators and metrics that New Leaders uses to measure progress toward student outcome goals.

Goal	Indicator	Metric
Close Achievement Gap	Percent of students scoring proficient or greater in Math and Reading	Percent of K-8 schools on track to statewide averages in math, reading, and in both subjects
Close Achievement Gap	High school graduation rate	Percent of schools on track to statewide average graduation rates
Improve Overall Achievement	Change in percent of students scoring proficient or greater in Math and Reading	Percent of K-8 schools making math, reading, and combined proficiency gains

Goal	Indicator	Metric
Outperform	Performance on Local Accountability Metrics	The percent of New Leaders schools outperforming partner district performance on local accountability metrics
Outperform	Change in percent of students scoring proficient or greater in Math and Reading	The average change in percent of students scoring proficient or greater in K-8 New Leader schools compared to the district and a matched sample of schools. The percent of K-8 New Leader schools making proficiency gains greater than the average proficiency gain of the partner district and a matched sample of schools.
Outperform	Value Added Measures	Statistically significant effect sizes calculated by external RAND evaluation

New Leaders tracks graduate and alumni career trajectory information in Salesforce. This electronic database allows the organization to build a comprehensive record of program graduate placements that is used to conduct analyses for various stakeholders. The database also allows New Leaders to merge placement data with other outcomes databases (like the longitudinal student achievement database).

RECOMMENDATIONS BASED ON LESSONS LEARNED

Many programs have funders that require them to examine student outcomes. Government agencies and accreditation processes might also require reporting of student outcome results. Even if this expectation is not the case, we recommend that principal preparation programs use student outcomes for:

- Internal accountability
- Program improvement
- Communicating the value of the program

Programs should use student outcomes data responsibly by:

- Understanding its limitations
- Maximizing validity of results (to the extent possible within resource, time, and data availability constraints)
- Using multiple indicators and metrics, including those that are aligned with organizational goals
- Triangulating results with other data and information

We also recommend the following strategies for addressing challenges we have encountered in collecting and reporting student outcome data:

- **Principal tenure data.** Having updated information on where program graduates are placed is critical because it defines which schools to include in student achievement analyses. Unfortunately these data are often imperfect or absent because principal tenure data is usually not publicly available, district data files are often outdated, and programs often lose touch with program graduates. When tenure data is missing or its accuracy is unknown, it can severely hinder the usefulness of examining student achievement results.

- Invest in meticulous tracking of program participant placement and tenure, including maintaining relationships with alumni. See the paper in this series on tracking graduate job placement for more information.

- **Data accessibility**. Often, it is difficult to either obtain or use the desired data. Preparation programs may not have necessary permissions to handle individual student- level data and processes for obtaining the data from districts may be slow, cumbersome, and sometimes costly. Publicly accessible data may be in formats that are difficult to use, have missing elements, or lack longitudinal consistency.
  - Build relationships with districts to aid in accessing student-level data (preferably with memorandums of understanding (MOUs)). New Leaders builds into all of its agreements with districts provisions for gathering student- level data, usually at minimal or no cost. New Leaders, whenever possible, builds personal relationships with the head of the district research and evaluation offices as well.

- **Cost to collect and analyze data**. Collecting and analyzing student outcomes can be time consuming and complicated given the limitations above, and organizations have limited human and financial resources to devote to the process.
  - Seek funding for rigorous, individual-level analyses.
  - Embed other program evaluation costs into the program cost model.

- **Time lapse between program participation, placement, and results availability**. The time between when our program participants complete our program and when their student outcomes results are available is typically two to four 2-4 years. While some program completers enter the principalship immediately, others might take a couple of years.
  - Try to set accountability metrics for three or more years in the principalship, while tracking interim data.

- **Stakeholder timelines**. In spite of logistical challenges, various stakeholders want results as soon as possible, leaving little time to address the lack of availability, the complexity of the data, or the time lapses in results. For example, many funders require updated results on an annual basis.
  - Make the investment in student-level data when possible, but use publicly available data as well to meet timeline requirements. New Leaders draws on both kinds of data to report on multiple metrics.

- **Comparability across states**. Since states typically have different state assessments, many metrics and indicators cannot be appropriately aggregated across states.
  - Develop metrics that do not refer to specific tests or performance indices (e.g., New Leaders uses percent of schools on track to meeting statewide proficiency averages).

- **Test changes over time**. This issue is particularly challenging when states and districts change tests, which makes gains or losses attributable in part to changes in the test rather than to the school or students.
  - Examine the school achievement relative to district averages from one year to the next.

REFERENCES

University Council for Educational Administration. (2013). *Developing evaluation evidence: A formative and summative evaluation planner for educational leadership preparation programs.* Charlottesville, VA: Author.