CENTER *for*
URBAN EDUCATION
LEADERSHIP

Continuous Improvement Brief

Selection of School Leadership Candidates for UIC's EdD Urban Education Leadership Program (Part II)

Lisa Walker
Kathleen K. Parkinson
Steve Tozer
Katonja Webb
Samuel P. Whalen

THE UNIVERSITY OF ILLINOIS AT CHICAGO EDUCATION

CENTER *for* URBAN EDUCATION LEADERSHIP

June 2018

Selection of School Leadership Candidates for UIC's EdD Urban Education Leadership Program (Part II)

Center researchers and EdD program administrators describe how they build program capacity to assess & measure the characteristics they associate with school leader success.

Lisa Walker
Senior Researcher
lwalker@uic.edu

Kathleen K. Parkinson
Research Specialist
kparki2@uic.edu

Steve Tozer
Director
stozer@uic.edu

Katonja Webb
Associate Director
katonja@uic.edu

Samuel P. Whalen
Director of Research
spwhalen@uic.edu

In the first part of this brief (April 2017), we consider selection from a development perspective while describing the characteristics we select for and how and why we select for these characteristics. In Part II, we describe how we built capacity for our theory of action for selection, as stated below:

If we know the qualities that we are looking for in candidates and we have a selection process that: (a) enables us to assess and measure evidence of these qualities and (b) we implement with fidelity, THEN we will make better choices about who is a promising candidate for our program. If our choices do not turn out well, we can revisit the selection process to see where we went wrong and how to improve.

We use concepts from improvement science (Bryk, Gomez, Grunow, & LeMahieu, 2015) to highlight lessons learned from our efforts to improve selection practices over a three-year period from 2014 to 2016. The brief illustrates that selection at UIC is a highly structured but flexible program practice supported by a set of tools and social processes that will continue to evolve with the UIC EdD program over time. The program's long-standing commitment to equity and diversity has been a strong shaping force.

One indicator of the effectiveness of UIC's selection process over the past 13 years is a 97% rate of placement in administrative roles in schools or school systems for those who successfully complete the pre-service portion of the program. Additionally, UIC has had a declining rate of attrition from the program due to academic and/or leadership performance problems—approximately 10% in recent years, as compared to up to 20% in earlier years. Another indicator of effectiveness is the diversity of our student population; overall, 60% of students have been minority and 40% have been white. Finally, for the past decade, Chicago has been decisively outperforming Illinois and national averages on measures of growth in student learning outcomes (Reardon, 2017; Zavitkovsky, Roarty, & Swanson, 2016). The two-decade Chicago Public School commitment to school leadership development, with residency-based programs such as UIC and New Leaders leading the way, has contributed to this trend toward higher performance (Emanuel, 2016; The Chicago Public Education Fund, 2015).

Our intent in writing this brief is to encourage discussion in the field of principal preparation about selection. It is also to encourage individual programs to use our experiences and approaches to reflect on their own selection processes. We caution that we do not regard our process as one for others to copy or emulate, because there is no one-size-fits-all process. As we discussed in Part I, what a program selects for depends on what a program prepares for and this can vary depending on program mission and local context. Additionally, we want to encourage other programs to engage in continuous improvement processes as we have. Smylie (2009) emphasizes the concept of "equifinality" in continuous school improvement, or the idea that organizations can begin at different starting points and pursue different paths to similar goals. In our case, the goal is to produce principals who can continuously improve schools as learning organizations for students, teachers, and staff.

LEARNING FROM THE NFL PLAYER DRAFT

To set the stage for the significance of selection in learning about leadership development, we first discuss the work of Richard Thaler, who examined the selection of football players for the National Football League (NFL). Thaler, who won the Nobel prize in economics in 2017, observes that accurately picking players who go on to perform at high levels is an extremely difficult task. Thaler demonstrated that although teams in the NFL stake millions of dollars on choosing star talent, it turns out that they often make the wrong choices during the NFL draft (Massey & Thaler, 2006; Thaler, 2015). By drawing parallels below between the NFL draft and selection for principal preparation, we begin to see the bigger picture of why it is important for the field of leadership development to learn to get better at the practice of selection.

Performance: In both the case of the NFL and selective principal preparation programs, the intent is to choose candidates who will perform well in the future. In our case, as discussed in Part I of this brief, we care most immediately about a candidate's suitability for development to achieve long-term performance as a leader in urban schools.

Value: In both cases, value is also a consideration, although it may receive less explicit attention than performance and may be more complicated to measure. In the case of an NFL team, the value of a player to the team is primarily economic—how much money he generates for the team considering the cost of drafting and compensating him. In our case, value is in sustained school performance and quality for the system and as experienced by students and families.

Decision-making: In both cases, making choices among candidates involves decision-making fed by information and influenced by biases. The NFL accumulates and generates abundant information on players' past performances and characteristics, and decisions are highly visible and high stakes. In our case, the decision-making process is the topic of this continuous improvement brief, more specifically, how we have learned to generate and use information to make decisions while at the same time mitigating biases.

Measurement: Thaler devises a way to evaluate team decision-making during the drafting of players against the performance and value of players once in the NFL. Key for him is the placement of players in the draft, both their draft round and their position within a draft round. This is a form of measurement. The improvements we have made in our selection process yield a rank order of candidates' levels of development on our selection domains that we will be able to use in ways similar to Thaler. We were motivated to improve information about candidates at the selection point for just this purpose.

Human behavior: Common underlying human behaviors influence selection in both cases as well. Thaler observes that unless a team attends to the fact that there is significant uncertainty about the future performance and value of any player, information about a player can lead to overconfidence in one's ability to choose strong performers. Overconfidence shows up in our selection process as reviewer attitudes of "I can pick them" and faith in a "gut" sense of a candidate's fitness for the principalship based on idiosyncratic rules of thumb and selective attention to the available data. We have replaced this with a focus on evidence of candidate characteristics, and selection tools and processes for generating, examining, and scoring this evidence to support reviewer accountability to a systematic approach.

Thaler ends his piece on the note that there are opportunities for the NFL to learn to make better choices through systematic data collection and analysis. We share some of Thaler's findings below as examples of lessons for the NFL that may be relevant to principal selection.

• Draft rounds reliably predict player differences in later performance, but predictions are less reliable within a draft round, particularly when players are close to each other in placement.

• The top-placed players in the draft are indeed the strongest performers. However, because these players are expensive to draft, they actually yield less value than lower-placed (also strong) players in the 1st draft round.

• Teams tend to succumb to pressures to exercise their opportunities to choose star players in the draft even though it can cost them dearly and they would do better choosing less highly ranked players.

Thaler developed insights around the trade-offs between performance and value of football players relative to draft star power. Similar trade-offs may occur in school leadership. Perhaps school leadership candidates with certain measurable characteristics, but not top candidates at entry, tend to remain in the principalship long-term and in doing so develop leadership expertise that yields value for students, their families, and the system. Issues of diversity and variability in career trajectories are key considerations here. This value might encourage us to select for these characteristics and influence how to accelerate candidates' leadership development. Only by doing the kind of work we describe in this brief can the field of principal preparation begin to explore selection data for these kinds of insights.

OVERVIEW OF SELECTION IMPROVEMENT PROCESS

A data infrastructure in the form of a powerful relational database system, FileMaker, provides a foundation for our improvement science work. UIC considers its primary and ultimate "clients" to be the students who attend Chicago's public schools and their parents. This sense of accountability requires an infrastructure to track the individuals we train in our program from selection, to placement and performance as novice leaders, to retention and performance as school leaders over time. One database in our Filemaker system contains considerable student record information, including work experience prior to program entry, progress in the program, assessments against program standards, and employment in administrative roles. Another complementary database contains records of CPS schools led by our graduates with measures of school organizational capacity and leading and lagging indicators of student learning and school culture and climate. The selection process developments we describe next, and the data they yield, will make it possible to leverage this data infrastructure to learn about the effectiveness of our selection process and, perhaps more significantly, about different developmental trajectories of our candidates with regard to important questions of performance and value.

Our selection improvement work began with our question of the student entry characteristics that are most related to performance as a successful school leader. This prompted us to examine and analyze existing evidence of these characteristics in student admissions' files for the first ten cohorts of students admitted to the UIC program. Data analysis revealed clear challenges for the program in selecting candidates in the middle range of ratings, that is, those who were neither clearly outstanding nor clearly unacceptable. It also revealed a selection process that was too variable and uncertain in its implementation to provide confidence in the reliability of the data generated. Working with program administration and with ongoing feedback from selection panel reviewers, research staff members led multiple inquiry cycles around the selection of three cohorts over three years to drive selection process improvements. Below we describe the context for this work and then we describe the work itself using improvement science.

BACKGROUND: UIC'S SELECTION PROCESS IN THREE STAGES

UIC's selection process has been through three stages of improvement over the program's 15-year history. Insights from the first two stages were generated more frequently through trial and error processes rather than systematic efforts disciplined by data. This changed when UIC received two federal grants to develop its program further, including it recruitment and selection practices, and hired several researchers as part of the program team to help meet grant objectives. One of the researchers led the cycles of inquiry described in this brief while another provided analytic expertise for measurement development. The recent stage 3 work is built on earlier lessons learned and existing program tools, processes, and data. We were not starting from scratch nor were we pursuing wholesale change. Indeed, the foundational elements for improvement were solidly in place.

The researchers' ready access to all aspects of UIC's selection work practices—from administration and coordination, to interviews, to final decisions—facilitated their cycle of inquiry activities. However, more significant to their ability to do this work was program leadership, who invited the researchers to participate as both learners and critical friends in the improvement process. Faculty members and leadership coaches were highly receptive to researchers' role as members of the team and to their contributions to practice improvements when data was brought to the table to inform discussions.

Key developments in each of the three stages of selection development work follow. Stages one and two created the foundation for stage 3 work and influenced the logic of how we proceeded with improvements.

Stage 1: Learning about the importance of dispositions

Errors in selection during the period from 2003 to 2007 made it critical for the program to get better at selection. Program leadership counseled out individuals at a higher rate than at any time since—in some years, at a rate of 15% of the matriculating cohort. In general, those counseled out tended to exhibit weak performance in both academic coursework and fieldwork practice. The program encountered common problems in these candidates, including a lack of commitment to their own learning and the work it required, and a disinclination or lack of ability to plan and to manage time and tasks. Candidates who did not respond to coaching to address these challenges continued to struggle until they chose or were advised to leave the program. Drawing from the literature, we speculate that these candidates lacked sufficient self-regulatory strength to manage and learn from the significant challenges they experienced in the leader development process (Day, Harrison, & Halpin, 2009). As a result, when they became overwhelmed, their performance deteriorated overall. In some cases, though less commonly, individuals were unable to make the transition from the identity of a classroom teacher to the identity of a school leader and administrative authority. Program staff came to understand these as problems of disposition rather than of knowledge or skill. This led to an emphasis on dispositions in our current selection process.

Stage 2: Assessing leader qualities

In the second stage of improvements from 2007 to 2013, stable application requirements and scoring sheets reflected the program's greater confidence in and satisfaction with its selection process. A particularly noteworthy change that occurred from the first to the second stage was the shift from scoring performance on application elements to using application elements to score leader qualities. In the first stage, for example, review panels scored a candidate's presentation on a scale of low to high. In the second stage, they scored the substantive qualities of an applicant as a leader and educator, for example, "Demonstrates a deep knowledge of the instructional practice needed to achieve high academic success."

To appreciate this change, it is helpful to consider that an application and interview for admission is a performance. Scoring *evidence of characteristics* generated during a performance differs from scoring the performance itself. The former is a harder task because reviewers must focus on the strength of the evidence separately from the confidence and skill of the performance. From our perspective as developers of leadership, performance certainly matters and indeed, one of our selection domains is *presence and attitude as a leader.* However, we also know candidates can develop skills in this domain when there are significant strengths in the other domains. We also know that a strong performance can mislead if it masks dispositions that can compromise leadership development, a lesson we learned in Stage 1.

Analysis of selection data conducted in stage 3 supports this emphasis on evidence of characteristics over evidence of performance in selection. Equity and diversity are central concerns around this distinction. We have found that biases become more influential when there are signs of weakness in a candidate during a selection interview. In particular, reviewers may be more forgiving of weaknesses when a candidate is of their own gender or race/ethnicity. While a focus on performance *plays into implicit biases,* a focus on characteristics valued by the program *enables discussion of the evidence* and its strength.

We believe this distinction between evidence of characteristics and evidence of performance helps even the playing field and allows us to maintain diversity in and across our student cohorts. In a field where white men have tended to dominate, we have an admissions record of 60% minority candidates and 40% white candidates, and 60% female and 40% male candidates. Our focus on seeking the strongest candidates for the development of leadership capabilities maintains our openness to candidates of diverse ages (from 25 years of age to over 50), races/ethnicities, and educational backgrounds, including CPS graduates who may have attended low performing schools, and it contributes to the equity of our process. At the same time, it creates challenges for us in differentiating our training to meet the developmental needs of a highly diverse student body. Yet, IF we use continuous improvement approaches to meet these challenges, THEN we will better serve our primary clients of students and their parents or caregivers.

Stage 3: Ensuring equitable selection practices

In the third stage starting in 2014, UIC's selection process has undergone significant development to better structure, systematize, and ground it in theory. Making our selection process more systematic has meant:

- clarifying and elaborating our selection criteria to name more specifically the abilities, behaviors, attitudes, and values we seek evidence for,
- mapping our application elements to the criteria and developing rating protocols and tools to assess evidence on the criteria,

• establishing policies around the use of educational credentials (GREs, GPAs, and institutions attended) in the selection process to prevent bias,
• developing protocols to structure and guide interview panels,
• developing implementation standards, including for selection panel membership,
• developing modules for training interview panel members and calibrating them to the rating system, and
• creating measures from the data and developing a decision-making protocol using these measures.

In doing this work, the literature has informed us to:

• keep in focus that selection leads to a developmental process involving professional training and personal transformation. First and foremost, we are looking for people with the characteristics we believe are the basis for strong development in a program such as ours (Browne-Ferrigno & Muth, 2012). This is the central theme of Part I of this brief (April 2017).
• come to terms with the requirements and procedures for admission to traditional programs in educational administration, which emphasize criteria such as GRE scores, grade point averages, and letters of recommendation focused on the likelihood of academic success (Mountford, Ehlert, Machell, & Cockrell, 2007).
• bring our program mission and values into focus in our selection process (Murphy, Moorman, & McCarthy, 2008).

The seven domain descriptions in Part I of this brief, and represented in Figure 1 on the next page, were key outcomes of stage 3 work, and in particular, were developed based on: lessons learned in stages 1 and 2 described above, the criteria used in stage 2 for scoring applicants, and cycles of inquiry work described in the next section.

ACTIVATING IMPROVEMENT SCIENCE

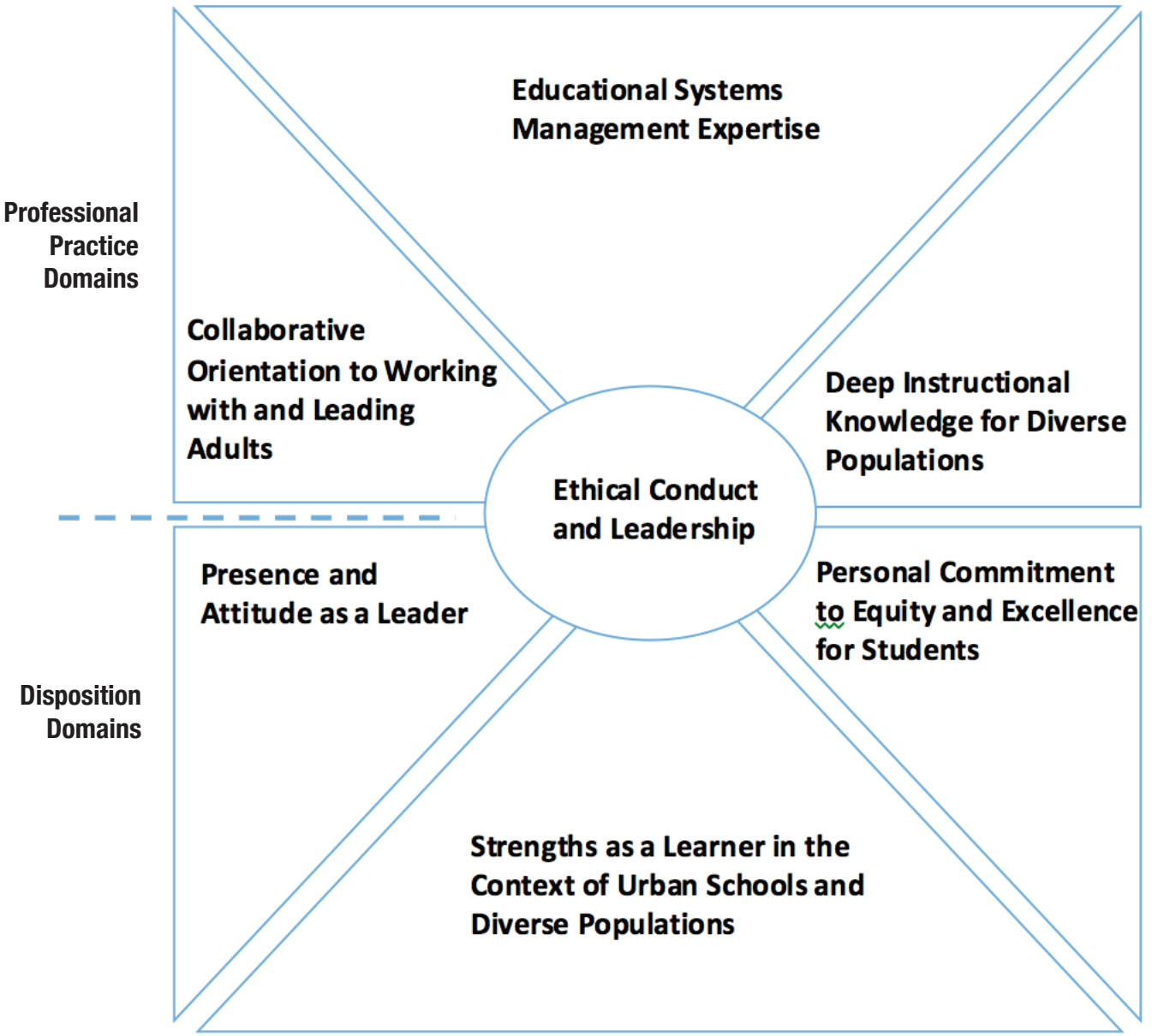
Working with program administration and with ongoing feedback from candidate reviewers, a Center researcher led multiple inquiry cycles around the selection of three cohorts over three years. Table 1 below and Table 4 in the Appendix summarize the inquiry cycles using a PDSA format (Plan-Do-Study-Act). Essential to each cycle was an inquiry stance, hence each PDSA cycle in Table 4 is preceded by a "focus of inquiry" question.

Table 1: Investigation Stages, 2014-2016

INVESTIGATION STAGE	FOCUS	PDSA CYCLES (see Table 4 for detail)
Exploratory 2014	Principal candidate characteristics and principal performance	1a – 1b
Year 1 Selection 2014	Selection processes and procedures	2a – 2c
Year 2 Selection 2015	Domain development for selection; Selection tool revision	3a – 3b
Year 3 Selection 2016	Reviewer manual; Training and calibration; Use of selection data in admission decisions	4a – 4b

Inquiry cycle 1a (see Table 4) into what student characteristics most correlate with performance as a successful school leader initiated work on our selection process. We used existing program data available in student admissions' files, scoring sheets used by the program over time, Excel spreadsheets of student progress in the program, and reflective accounts of program administrators and leadership coaches about why some students struggle and leave the program. This developed our understanding of the program's history of selection (stages one and two above) and led us to make changes in and further study our selection practices as we selected three cohorts of students over three years, as documented in cycles 2 thru 4 in Table 4.

Figure 1



Below we use five "structuring agents" of improvement science, as described by Bryk, Gomez, Grunow, and LeMahieu in ***Learning to Improve,*** to highlight lessons learned from this work and its outcomes. These agents are in shorthand form: work practices, measurement, system, inquiry and variability.

WORK PRACTICES: Make the work problem-specific and user-centered. Inquiry cycle 1a raised the question of the measures of student characteristics available to us and focused our attention on student admissions files. This is turn led us to examine the EdD program's selection practices. The work to identify student characteristics took shape as we began to address the problems of practice in the program's selection process and data. Data generated directly from applications and reviewer assessments, that is, from the work practices themselves, informed most of the changes we made to selection processes, tools, and measurement. The only exception was in the first year when we augmented the data with observations of candidate interviews to record qualitative accounts of candidate characteristics and to match them to reviewer ratings (cycle 2c).

Our users are coaches, faculty members, and other staff members who participate in the selection process. They experienced significant changes in selection work practices due to the improvement efforts. Users were also the candidates themselves, though their experience of the process changed relatively little. Three ways in which we were user-centered in our approach follow:

• We valued the expertise in the program's practice of selection and sought to clarify it so faculty, coaches, and other staff could examine and agree upon it, newcomers could have access to it, and it could be codified in tools and processes.

• We attended to the cognitive and task demands on reviewers when the number of items they rated increased from 10 to 34 during an intensive two-hour interaction. Prior to stage 3 work, reviewers used a single one-page scoring sheet at the end of the review process. Now they work through a 28-page scoring packet throughout the interaction with the candidate and application materials. To keep demands on them to produce data reasonable, we focused on producing practical rather than scientific measures (see page 100, 102 in Learning to Improve).

• We attended to how to reduce application and reviewer data effectively and to turn it around quickly to enable decision making in a time-constrained admissions committee meeting. Our first inclination was to provide too much information.

MEASUREMENT: We cannot improve at scale what we cannot measure. The important operating word in this structuring agent is "what." Most obviously for selection, "what" is about getting the right people into leadership positions. This could lead us to focus on developing measures to choose the strongest candidates, not unlike NFL teams focusing their bids on star players during the draft.

However, "what" can also mean development, in the sense of, "we cannot improve leadership development at scale without measurement." The measures we have developed are in service of this latter meaning consistent with the technical core of our work.

Primary Measures. In Table 2 below, we describe the three primary measures we have developed for candidates applying to our program. These measures, and how we use them, represent overall outcomes of our stage 3 improvement work.

Table 2: Primary Measures

Overall Development measure	This score informs us about the overall level of development of the candidate on our domain characteristics. As a single score, it represents the scoring of three interview panel members on twelve items during the application review and interview process. It summarizes 36 data points on a six-point scale: exemplary, advanced, developing, emerging, undeveloped, and red flag. This score is adjusted for differences in interviewer rating patterns, that is, whether interviewers tend to be easy or hard in their scoring practices. Prior to our stage 3 work, we lacked the technical capacity to produce scores such as this. Most significantly, this score allows us to differentiate our candidates through rank ordering and to categorize them as strong, middle, or low relative to the categories used in our scoring tools: exemplary, advanced, developing, emerging, and undeveloped. We do not set cut-points in the use of the measure, though we use "developing" (4.0) as a benchmark signaling sufficient strength for entry, depending on the other measures. Validation of these scores is evident in the decisions of our partner school district in awarding principal internships. (We should note however that we have concerns these decisions tend to be focused on performance prior to fieldwork placement and not development.) Our research analyst derives the scores statistically using multi-facet Rasch analysis. She is trained in measurement here at UIC and has access to faculty expertise in this approach. While this measure is sufficiently informative for admissions' decision-making, it does not isolate specific domain characteristics. Also of note is that scores are not comparable across cohorts. Hence, this is truly a "practical" improvement measure (see *Learning to Improve*, p. 100, 102); it is informative for our selection practices and it is potentially of use in analysis of later performance and value, but it lacks characteristics that would make it scientific.
Leadership Roles measure	This score informs us about the leadership experience of the candidate in schools, in particular, team leadership. It is an average of a single item score by the three interviewers. It provides additional information to consider along with the overall development score.
Recommendation measure	This score informs us about the how strongly the panel members recommend the candidate for admission to our program after scoring evidence of domain characteristics. It is an average of the final item in the panel review process. It is valuable in cases where a review panel does not find a candidate to be suitable for the program. In the past, this score received the most weight in making final admissions decisions. It encourages focus on those who are highly recommended by all three reviewers. However, it is not a measure of development and it does not serve to differentiate individual candidates well.

These three measures make it possible for us to consider the strength of recommendation and level of development, supplemented by a measure of leadership roles. Using multiple measures in this way helps to mitigate an "I/we can pick them" attitude. They allow us to identify and discuss cases in which the recommendation is strong, but the overall development score is low, and vice versa. The measures are a significant improvement over relying on recommendations alone and when used together, they help keep us honest about the fact that selection is an uncertain process. We can make the best judgment possible given the data we have, but we can never be certain it is the right judgment nor that we have all the most relevant data.

Secondary Measures. Although secondary measures used in making admissions' decisions focus on educational credentials (see Table 3) might led us to reject a candidate, they would not lead us to accept a candidate for whom the primary measures are not sufficiently strong. As noted in Part I of this brief, and resulting from our inquiry work (cycles 1b and 2a in Table 4), information about educational credentials is now redacted from application materials to prevent bias on the part of interview panel members. Only the committee that makes final admissions decisions sees candidate GRE scores, GPAs, and higher education institutions attended.

Table 3: Secondary Measures

GRE scores	These scores are known to correlate with performance in graduate school, but not the workplace. In general, we look for GRE scores that are in the middle 50% of the distribution. We tend not to credit high GRE scores when making decisions because there is no evidence of their link to leadership performance. We regard GRE scores that are low across the board (verbal, math, and analytic writing) as a warning sign that an applicant may not be sufficiently strong for doctoral work. In general, linking GRE scores to performance is complicated by the diversity of our candidates and their language and educational backgrounds. The Graduate Record Exam's Guide to the Use of Scores (2016-2017) advises caution in interpreting the GRE scores of individuals from minority groups in part because they are underrepresented in GRE validation studies. The math scores of some candidates we accept can be quite low, but we know little about how this relates to performance in our program or as a principal. We have some evidence of a link between GRE verbal scores and principal performance, though we have also observed complications of race/ethnicity and language backgrounds in this association.
Writing measure	This measure is the product of reviewer data and supplements the information from GRE scores, particularly analytic writing. It can confirm a low analytic writing score or raise questions about it. It is derived using Rasch analysis (as is the *overall development score*) of reviewers' assessment of writing skill on three application elements.
Grade Point Averages	Low undergraduate and graduate GPAs may raise questions or concerns. Similar to GRE scores, high GPAs are not predictive of leadership potential. The average GPA for our program is 3.0 and unlike GRE scores does not differ according by race/ethnicity.

Through our systematic use of the primary and secondary measures above, supplemented by demographic and employment data, we make final judgments about a candidate's suitability for development in our program. However, for middle candidates, we continue to question whether these measures are sufficient for making decisions. We know from previous experience that some of these candidates will do well in the program, though may take longer to develop into principals, and some will not do well. To make final decisions in these cases, the admissions committee may review and discuss qualitative notes on the interview panel's discussion of the candidate to identify the strongest candidates in the middle group. They are guided by the questions: Can we develop this person as a leader? Is there evidence of weaknesses across multiple data points that may suggest the candidate would have difficulty performing, in our program and as a leader?

SYSTEM: See the system that produced the current outcomes. Our work to improve data and measures was driven by two findings from our exploratory work (cycles 1a & 1b in Table 4). One was that principals who were identified by our leadership coaches as strong performers tended to receive strong ratings and recommendations during the selection process. This provided validation of our selection process. At the same time, this exploratory work and subsequent inquiry cycles drew attention to the uneven qualities and limited information value of the data produced by the process in the first ten years of the program. We knew we were doing something right, but the process and tools were too flawed and variable in their implementation for us to be sure what it was. They also did not result in usable measures.

The second finding was that there were challenges for the program in selecting candidates in the middle range of ratings and recommendations. Reviewers tended to diverge in their assessments of these candidates and it was unclear how the program decided to select one candidate over another. "Seeing the system" of selection was important for us to develop the existing strengths in the selection process and tools while also addressing the weaknesses. As documented in Table 4, our PDSA cycles targeted the following system drivers:

- Criteria for selection: We developed the Selection Domains (cycles 2b & 2c).
- Evidence of characteristics: Application/performance requirements mostly remained unchanged from previous years, except to bring them into alignment with the other drivers. One exception was the addition of an interview protocol (cycle 2a).
- Data generation: We did major work here by developing and/or revising reviewer processes and tools to assess evidence (cycles 2a, 3a, 4a).
- Measurement: We developed measures from scoring sheets, most notably the overall development measure described above (cycles 2b, 3b).
- Data use in admissions: We developed decision-making routines and processes (3b, 4b).

INQUIRY: Use Disciplined Inquiry to Drive Improvement. Here we illustrate more specifically the workings of a discrete inquiry within cycles 2a & 2b. The lead researcher raised a concern in the first year of the improvement work, based on review of student admission files, that both exceptionally strong and relatively weak GRE scores could result in reviewer

bias unrelated to leadership potential and may disadvantage minority candidates in particular. Because this rang true for program administrators, they made a decision to redact all educational credentials during the interview/performance stage of selection and for the admissions committee alone to weigh them when making final admissions decisions. This focused interview panel members on candidate development uninfluenced by perceptions of educational background. Later assessment of educational credentials against reviewer ratings affirmed the equity and measurement benefits of this practice. It yields scores on the primary measures that are uninfluenced by educational credentials, which will strengthen findings based on any later analysis of the data. We found, and continue to find, that where we observe overall weak educational credentials, interview panel assessments also tend to be weak. Additionally, the administrators' decision met a standard of "no harm done" due to the fact that educational credentials were still examined in the process. Essentially, we created a clear division of labor around the roles of interview panels and of the admissions committee, where the former assessed leadership and the later reviewed primary and secondary measures to make admissions decisions.

VARIATION: Focus on Variation in Performance. Discussion of this "structuring agent" requires us to observe that the program theory of action for leadership development shared in Part I of this brief has a flaw. While it shows variable development of leadership competence among candidates, it also shows that all candidates enter our program at the same level of development when selected. Our ability to see this flaw is an outcome of this improvement work. We have sought to create greater consistency in the process of selection in order to increase our confidence in measures that capture candidate variability. In particular, our use of multi-facet Rasch analysis has allowed us to disentangle the ways reviewers, candidates, and rubric items contribute to variability (or lack of) in scores at the selection point. This in turn has informed our improvements in selection criteria, processes, and tools. Although consistency in reviewer rating patterns remain a challenge due to differences in perceptions of different role groups (coaches, faculty, researchers) and/or experience levels (novice, expert), we have achieved a level of "practical measurement" of applicant quality far superior to what existed previously.

Our increasing confidence in the resulting measures allows us to look afresh at our program theory of action and to bring critical questions related to variability into the foreground. How do patterns of progression through our program and into the principalship vary depending on levels of development upon entering the program? How do these patterns relate, if at all, to retention in the role of the principal? How do patterns differ by race/ethnicity, age, gender, and educational backgrounds? If the field is to shift its understanding of selection from being about the strongest performers to being about the strongest people to develop, these questions will become central. The improvements we have made in our selection process will begin to enable answers to questions such as these.

CONCLUSION

Although we have drawn general parallels between the NFL and principal preparation, these sectors have very different selection processes and are selecting for different candidate qualities. Yet, there are similarities in what Thaler found and what we have found. In particular, Thaler recommends that teams invest their resources in selecting players who fall short of being the most desirable top performers. Teams will find strong performers and better value in this second or third tier group. We have learned that selecting strong candidates is relatively easy for us. Panel reviewers agree on their strengths, and while we cannot be sure any particular candidate will be a strong performer, our experience, theory of action, and data tell us strong candidates at the selection point tend to develop into effective leaders. However, to go to scale, the field of principal preparation needs to learn more about the developmental possibilities of different types of candidates rated in the middle third of candidates. These candidates are a potential source of diversity and strength for the field of leadership and for principal preparation programs using cohort models.

We will continue to interrogate our selection decisions against the variable progress candidates make in our program and in schools and as effective leaders, including the potential trade-off between candidate performance at the selection point and long-term value as a practicing principal in the school district. This will help us better understand how to select and develop the people who have the qualities to become the high-performing leaders we envision.

Selection is a resource-intensive practice, but in addition to ensuring principal quality, it pays off in developing the collective understanding of program staff in diverse roles. Reviewing thirty applicants requires 90 hours of time total from panel reviewers, not including training time. The EdD program's associate director spends many hours preparing materials and coordinating the review process for candidates and panel members. Producing data reports for the admissions committee meeting requires focused efforts from research staff. Yet selection is also an annual program event for faculty, coaches, and research staff when they come together to apply the values of the program collectively in the form of our selection criteria and learn from each other's perspectives as a new cohort of school leaders is selected into the field.

APPENDIX

Table 4: PDSA Cycles, 2014 to 2016

PDSA Cycle	Focus of Inquiry	Plan	Do	Study	Act
Cycle 1a: Retrospective Study on 10 Cohorts Summer 2014	What characteristics predict high performance as an urban school principal?	Explore correlations between characteristics available in student files and high performance as identified by leadership coaches.	Identify readily available data on student/principal characteristics; collect coach identification of high performers; enter data to create dataset; describe data and run correlations.	Suggestive findings around verbal skills based on GRE scores. Informative, preliminary descriptive work on UIC students. Limitations apparent in the available data on principal characteristics	Awareness of limitations of data led to examination of admissions files for the information/data they might contain on assessment of candidates at the point of selection into the program. See
Cycle 1b: Retrospective Study on 10 Cohorts Summer 2014	How predictive are admissions' data of high performance as a principal?	Develop coding scheme to apply across five different scoring sheets to evaluate strength of assessment at the point of selection.	Code available selection scoring sheets for all candidates admitted to program (data for 112 out of 158 students were available) and examine against coach identification of high performers, as well as progress in the program.	• Selection codes were predictive of high performance. • Selection data did not enable reliable identification of specific characteristics. • Panel assessments of "middle" candidates tended to be divergent, raising questions about the decision making process to admit. • Assessments of leadership potential could be biased by GRE scores and other educational credentials.	Findings from Cycle 1a and 1b were shared with coaches and program faculty. This led to recommendations to tighten up and improve selection process procedures to better support panel members in review of candidates and to strengthen data quality. See Cycle 2a.

APPENDIX

PDSA Cycle	Focus of Inquiry	Plan	Do	Study	Act
Cycle 2a: Selection of Cohort 13 Summer/Fall 2014	How well do the improvements made in selection processes and procedures work, both for the interview panel members and in strengthening the overall process?	Research team member to observe process, solicit/respond to feedback, and work with program coordinator to make/support changes as needed.	Multiple changes introduced into the process to better structure it for interview panel members. Key substantive changes were: • introduction of an interview protocol, • redaction of educational credentials from materials reviewed by panel members, • establishment of a formal admissions committee.	• Interview panel members found process improvements to be supportive/helpful overall. • Redaction of educational credentials prevented reviewer bias in interviews. • Establishment of admissions committee enabled examination of educational credentials prior to making final decisions.	Discrete changes made to the scoring sheet during this cycle had not addressed whether candidate characteristics were uniquely defined. We lacked data to inform such changes and took an, "It if ain't broke…" attitude. Questions about the scoring tool that had been in the background became more prominent once process changes were successful: What characteristics are interview panel members looking for and can we measure them? See Cycles 2b and 2c.
Cycle 2b: Selection of Cohort 13 Academic Year 2014	What measures can we develop from the selection data?	Use multi-facet Rasch analysis to analyze data collected under known conditions	Analysis of Cohort 13 selection data using Rasch methods	Scoring sheet contributes to measurement problems: • Ideas in items overlap/ not sufficiently distinct • Scoring levels not defined for interviewers who mostly use top two categories. • Reliability is a concern due to limited number of items (10), as well as their application/interpretation.	• Demonstrated potential for developing Rasch measures • Identified scoring problems based on Rasch analysis See Cycle 3b.
Cycle 2c: Study of Cohort 13 data Academic Year 2014	What characteristics are interview panel members selecting on?	Researcher to observe and take notes on interviews	Recorded notes of characteristics presented by candidates and probed/remarked on by interviewers along with interview member scores.	Domains of characteristics developed with aid of practice-based literature and EdD program selection scoring sheet.	Domains to guide revision, development, and alignment of application requirements and assessment tools and processes. See Cycle 3a and 3b.

APPENDIX

PDSA Cycle	Focus of Inquiry	Plan	Do	Study	Act
Cycle 3a: Selection of Cohort 14 Summer/Fall 2015	How do the significantly revised assessment tools work?	Pilot of new assessment tools	New tools used in Cohort 14 selection process with support from and observation by research staff member	Assessment tools must be aligned to process flow to support interviewers and prevent errors. Use of assessment tools are more demanding for interview panel members and require training and practice.	Tools passed the usability test after in-process revisions, though training of interview members is imperative. See Cycle 4a.
Cycle 3b: Selection of Cohort 14 Summer/Fall 2015	Can we develop measures for our selection domains? What does our analysis of selection data tell us about how the assessment process is working? Can we develop the measures in a timely enough fashion to use scores in admissions decisions?	Use multi-facet Rasch analysis to analyze selection data; Share results with admissions committee	Multiple analyses conducted of data; Scores reported to admissions committee in Excel spreadsheets	Initial finding that it is possible to develop: • broad measures of "disposition" and "professional practice," but not measures of individual domains, • overall "leadership potential" measure, • writing score measure Interview panel members differed too widely in their scoring practices. Timely analysis of selection data for use by admissions committee is possible. Admissions committee tends to be guided by recommendations, not development scores.	Revised assessment tools are incorporated into the selection process with the expectation that measures can be created from the selection data. Using measures and interview process data in the admissions decision-making process requires clearer procedures.

APPENDIX

PDSA Cycle	Focus of Inquiry	Plan	Do	Study	Act
Cycle 4a: Selection of Cohort 15 Summer/Fall 2016	Can we narrow the gap in scoring patterns through training and stronger panel reference/support materials?	Develop calibration training; Develop handbook for reviewers	Conduct 3 hour training, including use of handbook	We see improvements in attention to the tools and evidence, but panel members still differ too much in their scoring practices. Differences seem to correspond to their role group and/or expert/novice status as interview panel members. We see greater consensus on recommendations. We see few of the highly divergent assessments that were common in the first 10 cohorts.	Training will continue to be important, but differences seems to be related to role perspectives, which can be difficult to shift. Diverse panel membership can ensure equity.
Cycle 4b: Selection of Cohort 15 Summer/Fall 2016	How can scores on measures inform admissions decisions? What process data are helpful in admissions decisions? What procedures efficiently support use of measures and process data? What is the role of qualitative comments in the admissions decision-making process?	Develop process reports to flag implementation concerns, particularly those that may have affected assessments Develop data reports for admissions committee Develop protocol for use of the data	Prepared process reports including: * Interview panel composition * Panel member participation in training * Missing data report by interview panel member and candidate * Inconsistencies in items, candidate, and raters based on Rasch analysis Developed reports with final data for admissions committee Provided qualitative comments for “middle” candidates	Panel members continue to tend to favor (be led by) final recommendations rather than measures. Admissions committee members had the essential information they needed to make decisions. Qualitative comments helped panel members know where to “draw the line” and provided reassurance about their “picks” in the middle of the pack.	Continue to emphasize the different information value of the measures.

REFERENCES

Browne-Ferrigno, T., & Muth, R. (2012). Call for research on candidates in leadership preparation programs. Planning and Changing, 43(1/2), 10.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). Learning to improve: How America's schools can get better at getting better. Cambridge: Harvard Education Press.

Day, D. V., Harrison, M. M., & Halpin, S. M. (2009). Identity processes in leader development. An integrative approach to leader development: Connecting adult development, identity, and expertise (pp. 193). New York: Routledge.

Leonhardt, D. (2017, March 17) Want to Fix Schools? Go to the Principal's Office. The New York Times.

Massey, C., & Thaler, R. H. (2006). The loser's curse: Overconfidence vs. market efficiency in the National Football League draft. NBER Working Paper, (W11270).

Mountford, M., Ehlert, M., Machell, J., & Cockrell, D. (2007). Traditional and personal admissions criteria: Predicting candidate performance in US educational leadership programmes. International Journal of Leadership in Education, 10(2), 191-210.

Murphy, J., Moorman, H. N., & 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.

Reardon, S. F., & Hinze-Pifer, R. (2017). Test score growth among public school students in Chicago, 2009-2014. Stanford Center for Education Policy Analysis: Palo Alto, CA.

Thaler, R. H. (2015). Football. In Misbehaving: The making of behavioral economics (pp. 277-294). Allen Lane.