Compendium of Existing Measures for Understanding Leadership in Early Care and Education

ExCELS

OPRE Report #2021-220

December 2021

ACKNOWLEDGMENTS:

The authors would like to express their appreciation to Nina Philipsen, Bonnie Mackintosh, and Amy Madigan in the Office of Planning, Research, and Evaluation and Rachel McKinnon of the Office of Child Care for their input into the conceptualization of this work and their guidance throughout its development. We thank the Mathematica team, including Mark Ezzo, Robert Nutt, Louisa Tarullo, Emily Moiduddin, Colleen Fitts, Nicole Schatten, Jazmine Faherty, Jennifer Brown, and Mike Donaldson who contributed to the development of this product. We are also grateful to Sally Atkins-Burnett at Mathematica for her expert review.

Compendium of Existing Measures for Understanding Leadership in Early Care and Education

OPRE Report 2021-220

December 2021

Submitted to:
Nina Philipsen and Bonnie Mackintosh
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Lizabeth Malone, Mathematica
Scilla Albanese, Mathematica
Christopher Jones, Mathematica
Yange Xue, Mathematica
Gretchen Kirby, Mathematica
Anne Douglass, University of Massachusetts-Boston

Project Director:
Gretchen Kirby, Mathematica
1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221

Contract Number: HHSP233201500035I /HHSP23337038T

Mathematica Reference Number: 50708.C1.T05.000.000

Suggested citation:

L. Malone, S. Albanese, C. Jones, Y. Xue, G. Kirby, and A. Douglass. (2021). *Compendium of Existing Measures for Understanding Leadership in Early Care and Education.* OPRE Report 2021-220. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

DISCLAIMER:

The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research, and Evaluation are available at www.acf.hhs.gov/opre.

This page has been left blank for double-sided copying.

Overview

Leadership is widely recognized as an essential driver of organizational performance and improvement, but little is known about its role in improving the quality of early care and education (ECE) settings, or outcomes for staff and children. Additionally, information on how to measure the key constructs associated with leadership and the activities that demonstrate leadership is lacking. The Early Care and Education Leadership Study (ExCELS) focuses on leadership within an ECE center-based setting, at the building or center level. The ExCELS project approaches leadership as a construct that defines the range of people who participate in leadership in ECE centers—who leaders are—as well as what they bring to leadership, and what they do as leaders. *Leadership*, defined in this way, is broader than one *leader*, even while a strong center leader can be an essential ingredient to effective leadership.

Introduction

The goals of the ExCELS project are to: (1) fill the definitional and measurement gaps to understand what leadership looks like as defined by who participates in leadership in center-based ECE settings and the ways in which leaders can improve quality experiences for children in ECE settings, (2) develop a measure of ECE leadership, and (3) identify actionable leadership quality improvement (QI) initiatives and methods of evaluating them. The initial work of ExCELS focused on two foundational products that will guide the rest of the work: a literature review to inform a theory of change of ECE leadership for quality improvement, and this compendium of existing measures. The information from these products will inform the design of a descriptive study to develop and test a new measure of ECE leadership.

Topics

We document the following topics for 24 measures that focus on aspects of leadership that are relevant to early care and education center-based settings. The measures come from the early care and education, K-12 education, management, and health fields.

1. Purpose, context, and content measured
2. Administration characteristics and technical information
3. Availability and developer and/or publisher contacts

Purpose

The purpose of the ExCELS compendium is to describe information on measures of leadership relevant to early childhood settings. Further, the compendium identifies what information exists and what information is needed to better understand leadership and its influence in ECE settings. For the ExCELS project, the compendium will inform the development and testing of a new measure of ECE leadership. More broadly, the compendium will provide the ECE field —including researchers, program evaluators, and leadership program or quality initiative developers—with an overview of how various related disciplines are conceptualizing and measuring leadership in ways relevant to ECE.

Key findings and highlights

Based on the 24 measures profiled in the compendium, we have an initial picture of the landscape of leadership measurement.

- Measures tap aspects of leadership from the perspective of a particular field, generally management or ECE
- Content commonly taps aspects of what leaders do (the practices they engage in and promote)

- Leadership structure of who participates in decision-making—who leaders are—is captured less commonly in measures
- Primary purpose of all measures is research and evaluation
- Measures often aggregate staff reports about leaders to produce site-level scores
- Measures demonstrate acceptable reliability
- Validity information that demonstrates the measure captures what it intends to is generally available

Methods

The compendium involved:

- Identifying potential measures and screening them based on a set of criteria
- Review of the measure source documentation

Implications for next steps

Based on this compendium, the landscape of existing measures of leadership demonstrates breadth in the content areas available. However, the depth of and connection between content areas is still incomplete. We propose five considerations for future directions for ECE leadership measurement.

1. Increase measurement of who leaders are—the leadership structure within a center of who participates in decision-making
2. Expand the depth of information about what leaders do
3. Distinguish what center leaders and teacher leaders do
4. Differentiate the constructs of who leaders are and what they bring and do within a single measure
5. Connect who leaders are and what they do to relational coordination processes and distributed leadership approaches

Glossary

ECE: Early care and education

Center leader(s): Can be one or more persons who hold formal responsibility for overseeing administrative, operational, and instructional activities within an ECE center.

Distributed leadership: Leadership that recognizes behaviors or actions rather than job title or formal position alone, and that involves the primary center leader along with a range of staff—including teaching staff—in learning, decision-making, and planning and implementing change for improvement.

Leader: Who participates in leadership by contributing to decision-making and influencing change and quality improvement; leadership can include center leaders and teacher leaders.

Leadership: the combination of center and teacher leaders that exist within an ECE center

Relational coordination: Shared goals, shared knowledge, mutual respect, and high-quality communication among center leaders, teacher leaders, other center staff, and families.

Teacher leader(s): Teaching staff (lead, head, or co-teachers and assistant teachers) who carry responsibilities in the classroom and hold formal responsibilities to supervise and support other teaching staff or informally contribute to decision making and improvement

Contents

I. Introduction

Leadership is widely recognized as an essential driver of organizational performance and improvement, but little is known about its role in improving the quality of early care and education (ECE) settings, or outcomes for staff and children. The goal of the Early Care and Education Leadership Study (ExCELS) is to fill the definitional and measurement gaps to help the early childhood field better understand who participates in leadership in center-based early care and education (ECE) settings and how leadership can improve the quality of experiences for young children.

Leadership has not been well defined or measured in ECE (Dunlop 2008; Douglass 2017). ECE leadership is generally defined as "influencing or motivating groups of people to work together toward change, to accomplish a goal or solve a problem" (Douglass 2018; Nicholson et al. 2018). One review from the health field notes four common elements of leadership: (1) it is a process, (2) it entails influence, (3) it occurs within a group setting or context, and (4) it involves achieving goals that reflect a common vision (Cummings et al. 2010). Other studies in the education, health care, and management fields note similar concepts in defining leadership (Gumus et al. 2018; Hitt and Tucker 2016; Wong et al. 2013; Montano et al. 2017; Dunst et al. 2018).

The Early Care and Education Leadership Study (ExCELS) focuses on leadership within the center-based ECE context, at the building or site level. The ExCELS project approaches leadership as a construct that reflects the range of people who participate in leadership in ECE centers—who leaders are—as well as what they bring to leadership, and what they do as leaders. The study team will specify the elements of a leadership construct, map out and test the associations between leadership and staff and center outcomes, and develop a new measure of ECE leadership in center-based settings.

ExCELS will examine:

Leadership in ECE center-based settings at a *physical site or building level*, referenced throughout this report as the **center**.

Who leaders are in ECE centers defined by who participates in leadership by contributing to decision-making and quality improvement.

What leaders bring, defined by education, knowledge, skills, attributes, and values about ECE.

What leaders do, defined by taking action or pursuing practices that can affect positive outcomes.

The foundational work of the ExCELS project—a literature review and this compendium of existing measures of leadership—will help us identify indicators of ECE leadership in center-based settings that are important to measure and test because they are likely to contribute to positive outcomes for staff, center quality, and families and children. In this chapter, we present the purpose of the compendium and an overview of the ExCELS theory of change as the foundation for guiding the measurement content selected for the compendium.

Purpose of the compendium

The purpose of the compendium is to gather information on measures of leadership relevant to early childhood settings to identify what information exists and what information is needed to better understand leadership and its influence in ECE settings. For the ExCELS project, the compendium will inform the development of a new measure of ECE leadership. It is also intended to identify measures that might be useful in testing the reliability and validity of the measure in the ExCELS descriptive study. More broadly, the compendium will provide the ECE field —including researchers, program evaluators, and leadership program or quality initiative developers—with an overview of how various related disciplines are conceptualizing and measuring leadership in ways relevant to ECE.

Foundation for the compendium—Early Care and Education Leadership Study (ExCELS) theory of change of ECE leadership for quality improvement

The work of the ExCELS project is situated within a shifting policy and practice context that is placing a focus on how leadership in ECE settings can affect change for quality improvement. The K–12 literature has established the importance of having the principal and teachers work collaboratively to affect change and improve student outcomes (Bryk et al. 2010; May et al. 2016). The ExCELS project builds on the premise that leadership in ECE centers, as a construct, includes both center leaders and teaching staff in facilitating quality improvement in ECE settings.

We are interested in learning about the primary center leader—the individual who holds responsibility over the core administrative functions of the center, including operations and the educational program. In addition, we are interested in exploring who, beyond the primary center leader may also be involved in leadership. In particular, we want to examine the role that teaching staff play as leaders in ECE centers and the extent to which these roles are formally designated leadership positions or informal contributions to leadership based on their level of participation in decision-making and quality improvement for their classroom or the center as a whole.

Key definitions

Center leaders: can be one or more persons who hold formal responsibility for overseeing administrative, operational, and instructional activities within an ECE center

Teacher leaders: teaching staff (lead, head, or co-teachers and assistant teachers) who carry responsibilities in the classroom and hold formal responsibilities to supervise and support other teaching staff or informally contribute to decision making and improvement

Leadership: the combination of center and teacher leaders that exist within an ECE center

The findings from a literature review informed a theory of change that shows how ECE leaders can act as change agents for quality improvement and that reflects the unique elements of ECE settings. In this section, we present an overview of the theory of change as grounding for constructs used to identify measures for the compendium. For more information on the literature review of leadership within ECE, please see Kirby et al. 2021.[1]

Effective leadership is a driver of quality improvement in the literature we reviewed from the fields of K–12 education, management, and health. The ECE leadership literature is limited but emerging, and it identifies essential elements of leadership that align with aspects of effective leadership demonstrated in other fields. Little research or rigorous evidence exists in the ECE field about how ECE leadership may be effective in promoting quality and providing positive experiences for children that can lead to good outcomes. This highlights the need for a measure to define leadership in center-based ECE settings and test the pathways to outcomes as presented in the ExCELS theory of change.

A draft theory of change of ECE leadership for quality improvement. Using the research base available, we developed a draft theory of change of ECE leadership for quality improvement (Exhibit I.1). The ECE leadership literature lacks causal evidence to confirm the relationships or pathways we depict in the theory of change. However, each element in the theory of change—and the relationships between them—are constructed from the base of theoretical and descriptive empirical findings. Starting at the left of the theory of change, we list the external factors that might influence how ECE leadership is structured and what leaders are able to do and achieve in center-based settings, including the broad national, state,

[1] Kirby, G., A. Douglass, J. Lyskawa, C. Jones, and L. Malone. "Understanding Leadership in Early Care and Education: A Literature Review." OPRE Report 2021-02. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2021.

and local contexts. From there, we move on to center characteristics: the policy, regulatory, and fiscal infrastructure. Finally, we list the professional development and workforce supports that exist as part of systems external to the center. On the far right are the potential outcomes that ECE leadership might influence, including staff, center quality, and family and child outcomes. Each box of influences and outcomes lists elements that the research suggests might be important to better understand because of their influence on, contribution to, or outcome of leadership.

In the middle of Exhibit I.1 is our proposed construct of ECE leadership that is influenced by external factors and is hypothesized to contribute to the outcomes of interest. The ECE leadership construct comprises who leaders are, what leaders bring, and what leaders do.

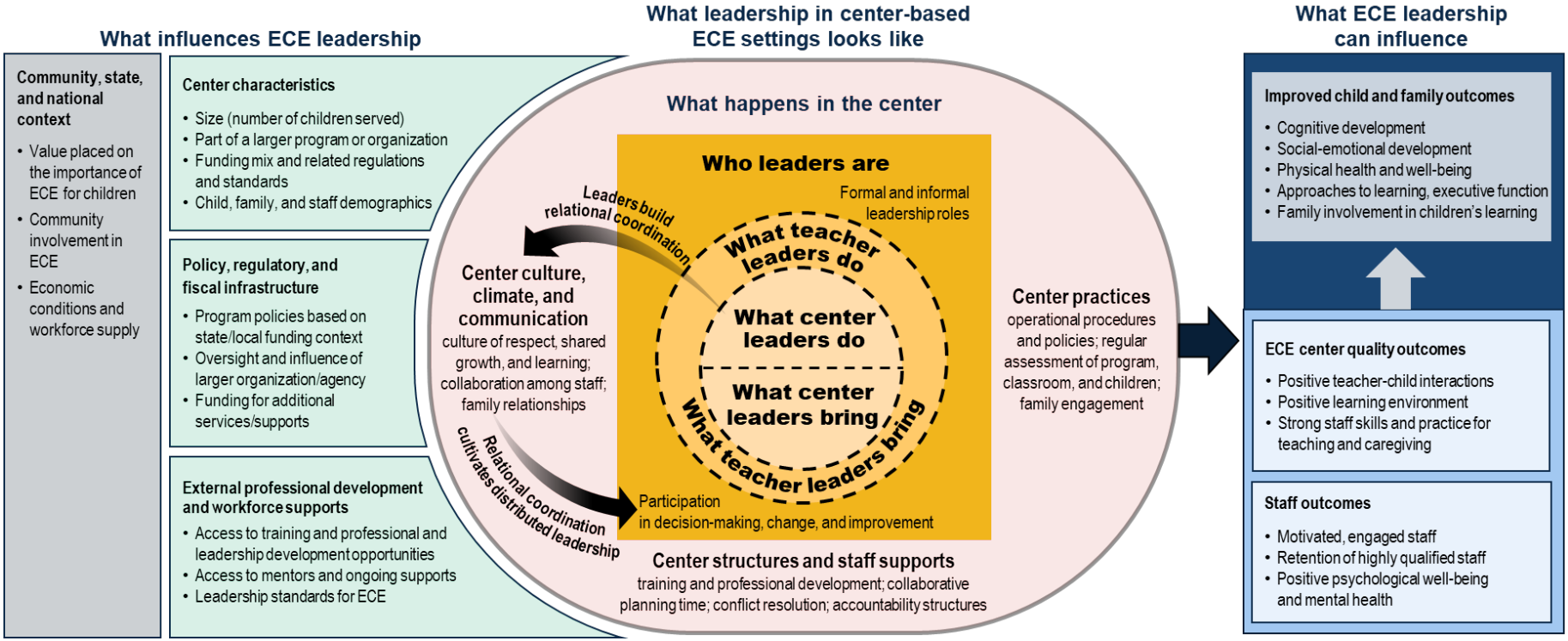
- "Who leaders are" captures the leadership structure that exists within a center, including staff involved in formal and informal leadership roles (notably teaching staff roles in leadership) based on who participates in decision making.
- "What leaders bring" includes the range of backgrounds, experiences, and characteristics staff bring to the task of leadership and can develop as leaders. Leaders bring (and can develop or advance) education, training, and experience; values, beliefs, and attributes; and knowledge, skills, and abilities. A range of knowledge and skills might contribute to successful ECE leadership, such as those concerning (1) personal development and critical thinking, (2) interpersonal and team building, (3) pedagogy and instruction, (4) advocacy and community building, and (5) administrative, business, and management.
- "What leaders do" describes the actions leaders take (or should take) to influence positive outcomes. Based on the practices identified in the ECE literature, we organized leadership practices into five categories: those that (1) promote, facilitate, and enable high-quality teaching and classroom quality; (2) create and sustain a culture of respect, collaboration, and continuous learning; (3) establish and implement a shared strategic vision; (4) promote family and community partnerships; and (5) establish and manage consistent and efficient organizational structures, operations, and performance management.

These interacting features together define leadership in a center-based setting. The leadership construct makes a distinction between center leaders and teacher leaders, given the need to better understand what each brings, develops, and does to contribute to leadership. However, the dotted lines between center leaders, teacher leaders, and who leaders are represent the permeability in who participates in leadership—particularly among teacher leaders—based on how individuals develop skills, knowledge, and abilities (what they bring and develop) and what leaders do to build or contribute to what happens in a center. Within the circles of these two types of leaders, center leaders are depicted as the inner circle because the pathway for center leaders to effect change in quality is through teaching staff. Center leaders influence how much teaching staff participate in leadership and, thereby, enable teacher leadership to occur and grow. Additionally, the center leaders may be one person or a group of different people that may include a site leader, along with others serving different roles as organizational leaders (such as an educational manager).

Because leadership is carried out within the center environment, the leadership constructs are situated within the outer circle, representing what happens in a center. The literature in ECE and other fields suggests an interdependent relationship between leadership and what happens in a center, and suggests that the latter might mediate the influence leadership has on intended outcomes. From the literature, we identified essential elements that create the symbiotic relationship between leadership and what happens in a center that can lead to positive staff and child outcomes, depicted by the arrows that loop between the leadership and center environment constructs. Through what they do, leaders can build relational

coordination, or having shared goals, shared knowledge, mutual respect, and high-quality communication among all members. This approach creates a supportive and collaborative environment between leaders, staff, and families. This type of environment, in turn, allows for broader participation in leadership through decision making and quality improvement, and creates distributed leadership structures that include teaching staff as leaders. The loop depicts relational coordination (as something leaders build through what they do) and distributed leadership (as an approach that is cultivated by the climate) to illustrate the importance of the following in promoting quality improvement: continual action and responsiveness among center leaders, teacher leaders, and what happens in a center.

Exhibit I.1. Early Care and Education Leadership Study (ExCELS): Theory of change of ECE leadership for quality improvement



Contents of the compendium

In the chapters that follow, we describe the process we used to identify and review measures (Chapter II) and then provide a summary of key elements across the measures (Chapter III). Chapter IV features the individual measure profiles. Appendix A contains a glossary of key terms on measurement and measure performance.

Compendium users can refer to Chapter III to identify potential measures that match their interests and needs in examining or understanding leadership. The specific measure profiles detail what a measure includes; how the measure is administered, scored, and interpreted; and whether the measure demonstrates acceptable evidence of reliability and validity.

II. Process for identifying and reviewing measures

In this chapter, we describe our process in developing the compendium. We outline the steps for identifying measures, list the measures screened for inclusion in the compendium, and our approach to reviewing those measures. The measure identification occurred in summer 2019. Reviews focus on documentation available as of 2019.

Identification of measures

First, we defined a measure as a set of items with technical documentation on its performance to provide a summary score (or set of scores). We did not include single items from surveys or competency frameworks with lists of behaviors. Second, we determined which measures to include in the compendium based on their relevance to the ExCELS theory of change. Our focus in ExCELS is on leadership that occurs within center-based settings serving young children (birth to age 5). The project focuses on leadership within ECE center-based settings, within the building (as compared to a larger organization or program structure), to understand and measure what leadership looks like within an ECE center—who participates in leadership, what they bring to leadership, and what they do as leaders. We approach centers that are part of a larger program as a center characteristic that might influence the structure of ECE leadership (see Chapter I). Measuring how leadership is structured and functions between the center and its parent organization falls outside the scope of the measures for this compendium. The measures in this review are not exclusively center-level measures; many focus on the primary leader (in a range of settings or organizations) but inform elements of the leadership constructs as defined for ExCELS. We describe the steps to identify and screen measures that align with this focus and scope.

Steps for identifying measures

We first identified measures by using the search conducted for the ExCELS literature review on leadership in ECE and other fields (Kirby et al. 2021). We also gathered written recommendations on measures that capture elements of the theory of change from experts specializing in ECE research or leadership research and measurement in the fields of management, K–12 education, and health.

In considering the range of measures applicable to the field, we included measures that assess styles of leadership (for example, the Multifactor Leadership Questionnaire) and provide information on behaviors, values, beliefs, or actions that may inform what leaders bring or do. Such measures can provide a foundation for understanding the range of leadership behaviors or practices to consider for measurement. However, we excluded measures that focus on a particular program or leadership development initiative (such as needs assessments or change scales).

More specifically, we included measures developed for the ECE field to capture the unique features and context of those center-based settings. For constructs for which few or no ECE measure exists, we identified existing leadership measures used in other fields—such as K–12 education, management, and health—that could be adapted for the ECE population and setting. Our intent in searching other fields is not to be exhaustive but rather to identify measures that may address gaps in ECE for assessing leadership. Therefore, we specified that non-ECE measures had to be broad enough to be adapted to the ECE setting. For example, we excluded measures targeted to duties or context unique to the other disciplines (for example, nursing work indices or patient safety climate). Also, given that several measures exist on the ECE work environment, we did not search for measures that focus on the work environment in other fields.

Screening measures for final selection

We used a set of criteria to screen for relevance of measures obtained from the literature search and suggestions from experts. The screening criteria were as follows:

Summary of screening criteria

- Construct found in core of ExCELS theory of change
- Published in past 25 years
- Technical documentation available on reliability and validity
- Developed or tested in the United States

- The initial list of measures was further screened to confirm that the content (Exhibit II.1) and scales assess leadership components in the core of the ExCELS theory of change, as opposed to broader contextual factors or outcomes. We identified whether only specific subscales align with our constructs and if so, focused our review on them.
- Measures needed to have been developed or tested within the last 25 years but used within a study and referenced in the literature within the past 10 years. For the purpose of the compendium, the measure year is defined as the year of the publication or technical documentation corresponding to the most current version of the measure. However, we want a measure to be recent enough to reflect the current policy, practice, and research context on leadership.
- Measures needed to have technical documentation available on their reliability or validity. We prioritized fully developed measures with demonstrated properties. Technical documentation was limited to that prepared by or identified by the measure developer or publisher.
- We limited the measures to those developed in the United States (unless using a U.S. sample). The measure and its documentation needed to be available in English. These criteria help focus on measures reflective of the national context and policies.

Exhibit II.1. Measure content reviewed: element and related constructs that align with the ExCELS theory of change

Element	Related construct
Who leaders are	Role-based constructs examined (such as formal and informal leaders or supervisory structure, teaching staff role in leadership, types of leadership responsibilities/duties, participation in decision making)
What leaders do	Behavioral constructs examined (such as ways to promote or facilitate teaching or quality practices, leader behavior to support respect and continuous learning, whether leader shares a strategic vision or mission, practices to promote family and community engagement, and management and business operations)
What leaders bring	Knowledge and skills that the literature suggests may contribute to ECE leadership (such as pedagogical knowledge; personal development or critical thinking knowledge and skills; interpersonal and team-building knowledge and skills; advocacy and community-building skills; administrative, business, and management knowledge and skills) Relevant values, beliefs, or attributes (such as beliefs about quality, collaboration; purpose-driven vision; committed, inspiring, or charismatic; authentic) Education, experience, and qualifications when the items are part of a scale is within a measure (not single item indicators for background purposes)
Center culture, climate, and communication	Specific relational constructs examined (such as culture of respect, shared growth, and learning; collaboration among staff; family relationships for process for interactions and building trust; relational coordination as a process in support of center communication)
Center practices	Administrative procedures, policies, and practices in place typically as a results of leadership activities (such as operational procedures and policies; regular assessment of program, classroom, and children; family engagement on practices or policies in place such as modes of communication, frequency of conferences)
Center structures and staff supports	Personnel and accountability constructs examined (such as training and professional development; planning time exists as a structure to allow collaboration; conflict resolution; accountability structures)

Based on this identification process, we included 24 measures in the compendium (Exhibit II.2).

Exhibit II.2. Measures profiled in the ExCELS compendium

Measure	Field
Administrator Role Perception Survey, Revised (ARPS)	ECE
Attributes of Leader Behavior Questionnaire (ALBQ)	Management
Authentic Leadership Inventory (ALI)	Management
Authentic Leadership Questionnaire (ALQ)	Management
Collective Leadership Survey	Management
Conger-Kanungo Scale of Charismatic Leadership (C-K Scale)	Management
Distributed Leadership Inventory (DLI)	K–12
Early Childhood Work Environment Survey, Third Edition (ECWES)	ECE
Essential 0-5 Survey (previously Early Education Essentials)	ECE
Implementation Leadership Scale (ILS)	Health
Leadership Practices Inventory (LPI)	Management
Multifactor Leadership Questionnaire, Third Edition (MLQ [5X-Short])	Management
Organizational Climate Descriptive Questionnaire (OCDQ-RE)	K–12
Preschool Instructional Leadership Survey, Version 2 (PILS)	ECE
Principal Instructional Management Rating Scale (PIMRS)	K–12
Program Administration Scale, Second Edition (PAS)	ECE
Program Quality Assessment Form B – Agency Items for Infant-Toddler and Preschool Programs (PQA Form B)	ECE
Program Sustainability Index (PSI)	Health
Relational Coordination Survey (RC Survey)	Cross-sector
Shared and Vertical Leadership Questionnaire (SVLQ)	Management
Supportive Environmental Quality Underlying Adult Learning (SEQUAL)	ECE
Survey of Transformational Leadership (STL)	Health
Tripod Teacher Survey	K–12
Vanderbilt Assessment of Leadership in Education (VAL-ED)	K–12

Field represents the discipline or setting the measure was developed or identified in the Early Care and Education Leadership Study literature review (Kirby et al. 2021)

ECE = early care and education.

Approach to measures review

We developed a profile for each measure, containing an overview followed by a detailed narrative. The same information is covered in the overview and narrative but these two profile components differ in the level of detail. For example, the overview indicates whether reliability and validity information is available, whereas the narrative describes the type of reliability and validity examined and its properties.

We documented information on measures along key dimensions in the profiles, including purpose and context, content (constructs measured), administration characteristics, technical information, and availability and costs (see Exhibit II.3 and Exhibit II.4 for the key dimensions within the overview and narrative sections of the profile, respectively). We used the primary sources on the measures based on those cited in the literature review articles and follow-up searches for documentation cited. As needed, we obtained technical documents from publishers and measure developers.

Exhibit II.3. Key dimensions within the measure profile overview page

Measure Name (Measure Acronym), Measure Development Year

Purpose and context

Purpose: Development/improvement (e.g., training and technical assistance, continuous quality improvement); Monitoring (to include accountability and reporting); Research/evaluation (e.g., studying analytic or descriptive questions or evaluating a particular program or initiative)

Field (relevant setting or discipline measure developed for): ECE (any age range for birth to age 5); K–12 education; Management (leadership, human resources, industrial or organizational psychology, organizational development and change, etc.); Health (nursing, mental health, etc.)

Content

Who leaders are

What leaders do

What leaders bring

Center culture, climate, and communication

Center practices

Center structures and staff supports

See Exhibit II.1 for definitions

Administration characteristics

Respondent: Manager/director; Staff (teaching staff roles for ECE when available; employee type for non-ECE field); Other (specify type of role, e.g., education manager)

Level of measure: Site (e.g., center for ECE; school for K–12; hospital for health); Group (within a site); Individual

Data sources: Survey (mode—self-administered, interview, or computer assisted; and report level—self-report versus report about leader); Direct observation; Document review

Usability Requirements for administering and analyzing measure (technology or application to complete survey, special statistical software needed to score, training requirements to use measure, and qualifications of staff or guidelines established based on who can purchase and interpret the measure)

Time/length: Number of minutes (or items)

Administration interval: How frequently can measure be administered without negating the validity: annual, semiannual, as frequently as desired

Languages available: English, Spanish, and list of any other languages

Technical information

Development sample: Information on the location[a], setting type, demographic characteristics of sample participants or settings, and year of development when data collected to assess measure performance

Measure performance

Reliability: Overview rating: 1 (none described); 2 (all or mostly under minimum acceptability ratings—0.70); 3 (meets minimum acceptability ratings—0.70). Rating prioritized documentation of internal consistency for surveys and inter-rater reliability for observations

Validity: Overview rating for whether information is available for construct/concurrent validity and for predictive validity

[a]Though all selected measures had to include United States leaders in development of the measures, we also reviewed measures that had non-U.S. respondents as part of overall sample (and were not analyzed separately) or when the developers use a non-U.S. sample only as support for key information on measure performance.

Availability

Permission needed for use Indicate whether in the public domain, a published source, and/or permissions required

Whether measure has costs List of costs associated with materials, training, or scoring (*as of time of review in 2019*)

Developer(s)/publisher contacts

Full names of those listed on the primary technical source of the measure; Publisher name, phone, and web address

Exhibit II.4. Key dimensions within the measure profile narrative page

Narrative

Description: Overarching summary of what the measure was designed to measure, for whom, and how that information is collected (type of data source, content measured, fields and type of settings the measure was developed for and used in; number of items; average administration time, and information about subscales)

Uses of Information: Summary of the measure purpose how the information that comes from the measure may be used (development/improvement; monitoring; research/evaluation)

Methods of Scoring: Overview of scoring process to include type of responses (and values/anchors if detail available in technical documentation) and how scores are calculated. If available, for measures based on multiple reporters (e.g., site level when collected from multiple individuals), estimates on an index of agreement that summarizes the appropriateness of aggregating individual reports

Interpretability: Information provided by developers/publishers on how to interpret a score or range of scores

Reliability: Available information on internal consistency, test-retest reliability, alternate form reliability, and/or inter-rater reliability (as applicable)

Validity: Available information on indicators of validity (evidence about whether the measure assesses what it is supposed to and the use/purpose examined) including content, construct, concurrent, and/or predictive validity). For example, construct validity includes information on factor analysis work to establish scales (exploratory or confirmatory factor analysis). For concurrent validity, there are varying ways to look at relationships or correlations with other measures (known as convergent or divergent validity) or with other criteria or outcomes. Direction, significance level, and sample size for the correlations are provided when available. See the glossary for guidance on magnitude of correlations as indication of strength of the relationship. Any concurrent validity analyses between measures included in the profile are cross-referenced and hyperlinked. For predictive validity, noting evidence examined (between the measure and another measure or criterion administered later in time) and including the coefficient in the narrative.

Bias Analysis: Description of work to determine whether items are fair for different groups. That is, do they function in the same way across different cultural, linguistic samples or settings. Analysis may include Differential Item Functioning (DIF) analysis.

Training Support: Information on training provided by or recommended by developers/publishers

Key Considerations for Early Care and Education (ECE): Discussion of aspects of measure to consider for use in ECE for understanding leadership *within* a center-based setting. For ECE measures, noted if developed for birth to age 5 or a specific age range, such as only preschool settings. For non-ECE measures, noted any need for revisions or adaptations in terminology on setting, roles, and duties. Note. Readers are encouraged to also review the sample description when considering use in an ECE setting.

Previous Version: Description of differences between the latest version reviewed in the compendium and a previous version

References: Citations for the measure, manuals, and other sources of information used to complete measure profile (if developers/publishers cite other work within these sources, those are noted for the interested reader to identify in the measure documentation)

This page has been left blank for double-sided copying.

III. Summary of Measures

Our review of the existing measures on leadership provides some directions for measuring ECE leadership in terms of what is measured and how it is measured (including how well a measure performs).

Overview of profiled measures

- Measures tap aspects of leadership from the perspective of a particular field, generally management or ECE
- Content commonly taps aspects of what leaders do
- Leadership structure of who participates in decision-making—who leaders are—is captured less commonly in measures
- Primary purpose of all measures is research and evaluation
- Measures often aggregate staff reports about leaders to produce site-level scores
- Measures demonstrate acceptable reliability
- Validity information that demonstrates the measure captures what it intends to is generally available

What is measured—field of study, content, and purpose

To identify a measure for a particular use, one needs to know what is measured. The field of study for a measure provides context on how questions may be framed (and the potential adaptability for use in ECE settings). Content focuses on what leadership looks like in a site (a center, school, or business setting, for example, depending on the field) as aligned with the ExCELS theory of change (see Chapter I). The content can be collected for multiple purposes.

Among the 24 measures profiled in this compendium, we find that measures draw on particular fields of study, which differ in the focus of content (Exhibit III.1). For example:

- Nine measures apply to the ECE field, 9 to management, 6 to K–12 education, and 5 to health care settings. Three of the measures have been used in more than one field.
- Collectively, the 9 ECE measures capture information across all six content areas that align to the leadership elements depicted in the ExCELS theory of change—(1) who leaders are; (2) what leaders bring; (3) what leaders do; (4) center culture, climate, and communication; (5) center practices; and (6) center structures and staff supports. However, only 3 measures provide information on who the leaders are in a center. The measures more commonly assess the center culture, climate, and communication (n = 8). Many of the measures also capture elements of what leaders do (n = 7).
- For the 15 measures in non-ECE fields, a similar pattern emerges that few provide information on who the leaders are, or leadership structure, (n = 2), but instead focus on what leaders do (n = 15) and what leaders bring from their background, experience, and characteristics (n = 11).

We see that measures do not usually provide information beyond the formal site leader.

- Across all 24 measures, 6 provide insight on the separate efforts of the site leader versus the staff (teaching staff or employees, depending on the field) in what leaders do.

- Two of these 6 measures highlighting the role of teacher leaders are specific to ECE settings (that is, Essential 0-5 Survey and Supportive Environmental Quality Underlying Adult Learning). One measure is from the K–12 education field, and the remaining three measures are broader across management and health roles.

All 24 measures serve to provide information primarily for research or evaluation (Exhibit III.2). Most measures can also be used for development and improvement (*n* = 21). Five measures specify use for monitoring purposes—2 from ECE, 2 from K–12 education, and 1 from management.

How is leadership measured—administration and performance

Based on key administration characteristics, we see that the measures use similar types of data sources and respondents and provide information to assess the measures' performance (Exhibit III.3).

- All measures rely on surveys or interviews as the data source.[2]
- Respondents to the measures represent varied roles and relationships to the leader.
  - 13 measures require self-report by the leader (for example, a manager, a director in an ECE setting, or a principal in a K–12 school), generally in concert with other reporters (10 of the 13)
  - 16 measures include a report on leaders by the staff, whereas another 3 focus on a report across all members of the team
  - 3 measures have supervisors report on the leader they oversee
  - 1 measure includes a report by those being served (parents)
- By aggregating the reports from multiple respondents, the measures create a site-level picture of what leadership looks like (n = 17). Having individual reports by either the leader or staff is still common practice for some leadership measures (*n* = 14).

The measures profiled demonstrate acceptable reliability, with one exception: the Program Quality Assessment Form B Agency Items. Overall these measures of leadership appear reliable, especially for how well items in the measure "hang together" (represented by internal consistency) and tell a coherent story about the center or individual's leadership and environment.

The measures have validity information available to determine whether they measure what is intended. Few measures have predictive validity indicators on the extent to which the measure results are related to longer-term or more distal staff or quality outcomes. Before using a measure, researchers or program evaluators or developers should consider the setting, group characteristics (like education or job position), and the measure developer's goal for the measure. For example, the developer of the Program Sustainability Index (PSI) examined the program elements of collaboration and leadership competence using a meeting of community-based programs and evaluators. Such prior use by a developer or publisher of a measure may indicate how broadly it may be used. Further, for measures that include international work and multiple settings, researchers or program evaluators or developers should use information provided in the profile to determine if the culture or use of the measure is generalizable to the population or group they are currently working with.

[2] Information not presented in exhibit (see individual profiles). One measure also used data from observations and/or document review (Program Administration Scale), and another measure could use observations if not completed by the leader (Program Quality Assessment Form B).

Exhibit III.1. Overview of field and content for measures included in the ExCELS compendium

Measure	Field				Leadership content					
	ECE	K–12 education	Manage-ment	Health	Who leaders are	What leaders do	What leaders bring	Center culture, climate, and comm-unication	Center practices	Center structures and staff supports
Administrator Role Perception Survey, Revised (ARPS)	X ♦				X	X	X	X		
Attributes of Leader Behavior Questionnaire (ALBQ)			X			X	X			
Authentic Leadership Inventory (ALI)			X			X	X			
Authentic Leadership Questionnaire (ALQ)			X			X	X			
Collective Leadership Survey			X			X †		X		
Conger-Kanungo Scale of Charismatic Leadership (C-K Scale)			X			X	X			
Distributed Leadership Inventory (DLI)		X			X	X †				
Early Childhood Work Environment Survey, Third Edition (ECWES)	X ♦					X		X	X	X
Essential 0-5 Survey (previously Early Education Essentials)	X ♦				X	X †		X	X	X
Implementation Leadership Scale (ILS)				X		X	X			
Leadership Practices Inventory (LPI)		X	X	X		X	X			
Multifactor Leadership Questionnaire, Third Edition (MLQ [5X-Short])			X			X	X			
Organizational Climate Descriptive Questionnaire (OCDQ-RE)	X	X				X		X		
Preschool Instructional Leadership Survey, Version 2 (PILS)	X ♦					X				X
Principal Instructional Management Rating Scale (PIMRS)		X				X		X		

Exhibit III.1 (continued)

Measure	Field				Leadership content					
	ECE	K–12 education	Manage-ment	Health	Who leaders are	What leaders do	What leaders bring	Center culture, climate, and comm-unication	Center practices	Center structures and staff supports
Program Administration Scale, Second Edition (PAS)	X ♦					X	X	X	X	X
Program Quality Assessment Form B – Agency Items for Infant-Toddler and Preschool Programs (PQA Form B)	X ♦						X	X	X	X
Program Sustainability Index (PSI)				X	X	X †	X	X	X	
Relational Coordination Survey (RC Survey)	X		X	X				X		
Shared and Vertical Leadership Questionnaire (SVLQ)			X			X †	X	X		
Supportive Environmental Quality Underlying Adult Learning (SEQUAL)	X ♦				X	X †	X	X	X	X
Survey of Transformational Leadership (STL)				X		X	X			
Tripod Teacher Survey		X				X	X	X	X	X
Vanderbilt Assessment of Leadership in Education (VAL-ED)		X				X				

An "X" indicates that the field and content applies to the measure.

ECE = early care and education

♦ Developed for/with ECE sample (serving any age group birth to age 5 but not necessarily the full age range)

† Includes what teacher/staff do as leaders

Exhibit III.2. Overview of measure purpose for measures included in the ExCELS compendium

Measure	Purpose		
	Development/ improvement	Monitoring	Research/ evaluation
Administrator Role Perception Survey, Revised (ARPS)	X		X
Attributes of Leader Behavior Questionnaire (ALBQ)	X		X
Authentic Leadership Inventory (ALI)			X
Authentic Leadership Questionnaire (ALQ)	X		X
Collective Leadership Survey			X
Conger-Kanungo Scale of Charismatic Leadership (C-K Scale)	X		X
Distributed Leadership Inventory (DLI)			X
Early Childhood Work Environment Survey, Third Edition (ECWES)	X		X
Essential 0-5 Survey (previously Early Education Essentials)	X		X
Implementation Leadership Scale (ILS)	X		X
Leadership Practices Inventory (LPI)	X		X
Multifactor Leadership Questionnaire, Third Edition (MLQ [5X-Short])	X		X
Organizational Climate Descriptive Questionnaire (OCDQ-RE)	X		X
Preschool Instructional Leadership Survey, Version 2 (PILS)	X		X
Principal Instructional Management Rating Scale (PIMRS)	X		X
Program Administration Scale, Second Edition (PAS)	X	X	X
Program Quality Assessment Form B – Agency Items for Infant-Toddler and Preschool Programs (PQA Form B)	X	X	X
Program Sustainability Index (PSI)	X		X
Relational Coordination Survey (RC Survey)	X		X
Shared and Vertical Leadership Questionnaire (SVLQ)	X		X
Supportive Environmental Quality Underlying Adult Learning (SEQUAL)	X		X
Survey of Transformational Leadership (STL)	X	X	X
Tripod Teacher Survey	X	X	X
Vanderbilt Assessment of Leadership in Education (VAL-ED)	X	X	X

An “X” indicates that the purpose applies to the measure.

Note. Purpose is defined as follows: development/improvement includes training and technical assistance and other efforts to support continuous quality improvement of the center or staff; monitoring includes accountability and reporting; research/evaluation includes work to answer analytic or descriptive questions or evaluate a particular program or initiative.

Exhibit III.3. Overview of measure administration characteristics and properties for measures included in the ExCELS compendium

Measure	Respondent				Level of information			Measure performance	
	Manager/ director/ principal/ leader	Staff (teachers, employees)	Parents	Other	Site	Group within site	Individual	Reliability	Validity
Administrator Role Perception Survey, Revised (ARPS)	X						X	+	Available
Attributes of Leader Behavior Questionnaire (ALBQ)	X	X					X	+	Available
Authentic Leadership Inventory (ALI)		X					X	+	Available
Authentic Leadership Questionnaire (ALQ)	X	X			X		X	+	Available
Collective Leadership Survey				Road team	X			+	Available
Conger-Kanungo Scale of Charismatic Leadership (C-K Scale)		X					X	+	Available
Distributed Leadership Inventory (DLI)	X	X			X	X	X	+	Available[a]
Early Childhood Work Environment Survey, Third Edition (ECWES)	X	X			X			+	Available
Essential 0-5 Survey (previously Early Education Essentials)		X	√		X			+	Available
Implementation Leadership Scale (ILS)	X	X			X			+	Available
Leadership Practices Inventory (LPI)	X	X		Manger/ director's supervisor	X		X	+	Available
Multifactor Leadership Questionnaire, Third Edition (MLQ [5X-Short])	X	X			X		X	+	Available▲
Organizational Climate Descriptive Questionnaire (OCDQ-RE)		X			X			+	Available
Preschool Instructional Leadership Survey, Version 2 (PILS)				Instructional leader			X	+	Available
Principal Instructional Management Rating Scale (PIMRS)	X	X		Principal's supervisor	X		X	+	Available

Exhibit III.3 (continued)

Measure	Respondent				Level of information			Measure performance	
	Manager/ director/ principal/ leader	Staff (teachers, employees)	Parents	Other	Site	Group within site	Individual	Reliability	Validity
Program Administration Scale, Second Edition (PAS)	X				X			+	Available
Program Quality Assessment Form B – Agency Items for Infant-Toddler and Preschool Programs (PQA Form B)	X				X			–	Available
Program Sustainability Index (PSI)				Community-based program professional			X	+	Available
Relational Coordination Survey (RC Survey)		X			X	X	X	+	Available
Shared and Vertical Leadership Questionnaire (SVLQ)	X	X			X			+	Available▲
Supportive Environmental Quality Underlying Adult Learning (SEQUAL)		X			X			+	Available
Survey of Transformational Leadership (STL)		X					X	+	Available
Tripod Teacher Survey		X			X			+	Available
Vanderbilt Assessment of Leadership in Education (VAL-ED)	X	X		Principal's supervisor	X		X	+	Available

An "X" indicates that the administration characteristic and properties of measures applies to the measure.

[a] Validity information primarily based on a non-U.S. sample.

Reliability key:

+ indicates scores meet minimum acceptability ratings—0.70

– indicates all or mostly all scores under minimum acceptability ratings

Reliability ratings prioritize internal consistency for surveys; inter-rater reliability for observations or document review. Ratings may also reflect availability based on a previous version, as noted in the specific profile.

Validity key:

"Available" indicates information on construct/concurrent validity is available.

▲ indicates that predictive validity is also available for measure.

Ratings may also reflect availability based on a previous version, as noted in the specific profile.

Future directions for ECE leadership measurement

Based on this compendium, the landscape of existing measures of leadership demonstrates breadth in the content areas available. However, the depth of and connection between content areas is still incomplete. We propose five considerations for future directions for ECE leadership measurement.

Considerations for measuring ECE leadership

1. Increase measurement of who leaders are—the leadership structure within a center
2. Expand the depth of information on what leaders do
3. Distinguish what center leaders do from what teacher leaders do
4. Differentiate the constructs of who leaders are and what they bring and do within a single measure
5. Connect who leaders are and what they do to relational coordination processes and distributed leadership approaches

1. We need to develop more measurement of "who leaders are" to better understand the structure of formal and informal leadership roles and who participates in decision making across settings.

2. We should expand the depth assessed about what ECE leaders do. Although "what leaders do" appears prominent in the existing measures, the details in the profiles demonstrate that measures vary on the number and specificity of items to fully address "what leaders do" across the key activities identified in the literature as drivers of quality—(1) promote or facilitate teaching quality, (2) support respect and continuous learning, (3) share a strategic vision or mission, (4) promote family and community engagement, and (5) demonstrate efficient management and business operations.

3. Related to the area of "what leaders do," we need to clearly distinguish what center leaders do and what teaching staff do as leaders. Some measures include information on specific activities undertaken by staff. For ECE, this distinction needs further development, particularly to measure what teaching staff do beyond instructional practices to participate in leadership activities.

4. The ECE field would benefit from a single measure that addresses "who leaders are," "what leaders bring," and "what leaders do" but differentiates them to maintain the three constructs as distinct subscales for scoring and interpretation. Based on our review, few measures include all three constructs or when measures do include all three they do so in a way that does not measure each area deeply on its own or separately from others. Differentiating the three constructs, as defined by the ExCELS theory of change, will improve the ability to disentangle the aspects of leadership that are most important for specific outcomes for staff, center quality, families, and children.

5. Based on the ExCELS theory of change, the intertwining influences of relational coordination and distributed leadership are important (Kirby et al. 2021). Separate measures are available to assess relational coordination processes. However, in the future, measures that can link or promote the connection between these constructs would strengthen the ECE field's ability to understand how leadership operates within a center among all staff.

Taken together, these considerations and future directions will help ECE researchers, program administrators, and practitioners to fully assess the various factors that can develop and support ECE center-based leaders to be effective agents of quality improvement.

ExCELS Measure Profiles

This page has been left blank for double-sided copying.

ExCELS Measure Profiles

This page has been left blank for double-sided copying.

Administrator Role Perception Survey, Revised (ARPS), 2019

Purpose and context

Purpose: Development/improvement,

Field: Early care and education (ECE)

Content

Who leaders are (Leadership roles)

What leaders do (Promote quality practices, foster respect and learning, establish vision, promote family/community partnerships, manage efficient operations)

What leaders bring (Pedagogical knowledge; personal development or critical-thinking knowledge and skills; interpersonal and team-building knowledge and skills; advocacy and community-building skills; administrative, business, and management knowledge and skills)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; family relationships)

Administration characteristics

Respondent: Director

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–self-report)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills with some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 25 minutes

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States (49 states and the District of Columbia)

Setting: ECE center

Sample: 1,530 early childhood program administrators in McCormick Center's contact database; *position*: 67% director, 16% owner/director, 6% supervisor, 4% manager, and 6% other; worked in ECE field for average of 22.5 years; average age 48 years; 96% female; *race/ethnicity*: 79% White, 12% Black, 3% two or more races, 1% American Indian, and 8% Hispanic; *education/credentials*: 34% master's degree, 37% bachelor's degree, 17% associate's degree, and 7% high school; 67% degree major in child development or ECE; 17% Child Development Associate credential

Year of development: 2018

Measure performance*

Reliability: 3 (meets minimum acceptability rating—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

*Version examined based on measure adjustments that do not appear to have been revalidated with final set of items.

Availability

2-Published source, contact developer(s)/publisher about permission requirements
Material, training, and scoring costs: Not available

Developer(s)/publisher contacts

Developer(s): Michael Abel, Jill Bella, Teri Talan, and Marina Magid
Publisher: McCormick Center for Early Childhood Leadership
6200 Capital Drive
Wheeling, IL 60090
www.McCormickCenter.nl.edu
(847) 947-5312
Michael.abel@nl.edu

Narrative

Description: The Administrator Role Perception Survey (ARPS) is a self-report measure completed by early childhood program administrators to measure administrators' perceptions about their roles, perceived leadership competency, and professional development needs aligned with the McCormick Center for Early Childhood Leadership's Whole Leadership Framework. It can be self-administered on the web or on paper. The survey takes about 25 minutes to complete. The survey has multiple sections covering various topics, including 25 items on leadership self-efficacy used to assess an administrator's confidence in their ability to perform various leadership functions, 24 items to determine time spent on leadership functions, 8 items that provide context about the administrator's perception of their job (for example, whether they feel uncertainty in their authority, whether they feel respected by staff or families), questions to characterize one's administrative role, and questions to determine the respondent's developmental stage as an administrator. The ARPS also includes items about the administrators' commitment to their job and job satisfaction, demographic and background questions (for example, types of positions and years of experience), and questions about their program characteristics (for example, location and funding). Please review Abel et al. (2019) for more information on these measures, as they are not the focus of this compendium. The developers used the Whole Leadership Framework to build on earlier work by Rafanello and Bloom (1997), which used the Directors' Role Perception Survey (DRPS). The 25 leadership self-efficacy items on the ARPS (referred to as a scale) are divided into three subscales (referred to as whole leadership domains) including Leadership Essentials (6 items), Pedagogical Leadership (8 items), and Administrative Leadership (11 items) to study leadership as an overall construct and by its individual subscales—see Abel et al. (2019), p.12. The three self-efficacy subscales have content related to "What leaders do" and "What leaders bring." The job perception items relate to "What leaders bring" and "Center culture, climate, and communication." The remaining items on leadership functions, roles, and developmental stage include content across the "Who leaders are" and "What leaders do" content areas.

Uses of Information: The ARPS can be used for development or improvement purposes. The developers stated that it could be used to identify areas where administrators are most effective and areas where they could benefit from additional professional development. The developers also intend the survey to be used for research or evaluation purposes to study how role perceptions and professional development needs may vary by developmental stage of the administrator.

Methods of Scoring: The ARPS items use various answer choices and scales. The leadership self-efficacy items are scored on a 4-point scale, with respondents indicating their level of confidence from not confident (1), sometimes confident (2), confident (3), and very confident (4). A mean score for the three self-efficacy subscales and a total self-efficacy score can be calculated by adding the item scores together and dividing the value by the number of items in each subscale and the total number of items, respectively. For the ARPS leadership functions items, respondents are first asked how much time they spend on a particular function on a scale from 1 to 5: no time (1), a little time (2), some time (3), quite a bit of time (4), and a great deal of time (5). Respondents are then instructed to identify if another staff member is responsible for the functions from a list of positions. Respondents are asked about their current perceptions of their job on a negative to positive perception scale with 4-points from often (1), sometimes (2), most (3), and always (4), where "often" and "sometimes" indicate the presence of a negative perception and "most" and "always" indicate the presence of a positive perception. Respondents are asked to describe their role as an administrator by selecting 3 words or phrases from a list of 15. The ARPS also includes a section that asks respondents to read three descriptions and to check the one that best describes

them. That section can be used to categorize respondents into the three developmental stages related to their job and leadership skills (novice, capable, and master). For these items, the perception items, and the leadership function items, additional scoring information is not provided by the developers. See the Interpretability section for additional details.

Interpretability: Higher scores on the leadership self-efficacy subscales indicate higher levels of confidence in perceived leadership competence. Responses to the ARPS scales can be interpreted across the three developmental stages. Please review Abel et al. (2019) for item-level and subscale descriptive statistics, including means and standard deviations. McCormick Center for Early Childhood Leadership (2018) provides additional information on the developmental stages and how ARPS responses can be interpreted across them.

Reliability:

(1) Internal consistency reliability: The Cronbach's alpha coefficient for the leadership self-efficacy total score was 0.94. Cronbach's alpha for the three self-efficacy subscales were 0.84 (Pedagogical Leadership), 0.87 (Administrative Leadership), and 0.85 (Leadership Essentials).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers used the Whole Leadership Framework to revise the Rafanello and Bloom (1997) DRPS and renamed it the ARPS. The ARPS consisted of 48 items regarding role perceptions and self-efficacy, 14 demographic items, and 7 items about program characteristics. The developers conducted a pilot study in Texas with 34 expert reviewers that included 29 leadership self-efficacy items and involved cognitive interviews with the expert reviewers. As a result of the pilot, the developers made minor revisions to the ARPS and refined protocols for administering the ARPS in the validation study.

(2) Construct/concurrent validity:

Construct validity: The developers conducted an exploratory factor analysis (EFA) on the self-efficacy items, which resulted in a four-factor model: operational management (corresponding to the Administrative Leadership subscale), communication/advocacy (or the Leadership Essentials subscale), pedagogy (the Pedagogical Leadership subscale), and technology. The factor loadings ranged from 0.59 to 0.68 for Administrative Leadership, 0.51 to 0.74 for Leadership Essentials, and 0.50 to 0.72 for Pedagogy. The fourth factor comprised the technology items, with factor loadings of 0.83 and 0.84. As a result of the EFA, the two technology items and three items that did not load to any self-efficacy factors were reassigned to one of the other three factors (as derived from the Whole Leadership framework). The factor loadings for those items after the reassignment are not available.

Concurrent validity: The developers found that responses to the leadership self-efficacy items varied by developmental stage of administrators; individuals who identified as being in a higher developmental stage were more likely to have higher confidence scores for the self-efficacy items. This was also the case across the three self-efficacy subscales based on these items. The developers also discovered that participating in a leadership academy, a form of leadership professional

development, increased respondents' total and subscale scores for leadership self-efficacy compared to those scores for administrators who did not participate in an academy. The analysis suggests evidence for concurrent validity.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Designed for use in ECE settings. The ARPS was validated within a sample of 1,530 early childhood program administrators in McCormick Center's contact database. A majority of the sample was female (96 percent), white, (79 percent), and holding a director position (67 percent). The average experience of sample members was 48 years. A majority of respondents indicated holding a master's degree (34 percent) or bachelor's degree (37 percent), with 67 percent holding a degree in child development or ECE.

The psychometric information noted above is based on the national study of the ARPS that was conducted after the pilot. After the national study, the developers revised the measure and renamed it to ARPS. The developers changed the language throughout the measure to encompass the various types of program leaders, changing from "director" to "administrators"; made minor tweaks to item wording, removed items, added new items, and revised the names of the developmental stages. The self-efficacy items dropped from 29 items to 25 items, and the developers added 24 items to the survey to measure time spent by administrators on leadership functions. The developers did not conduct psychometric analyses on the measure after these revisions.

Previous Version: The ARPS is a revision of the Directors' Role Perception Survey (DRPS; Rafanello and Bloom) developed in 1997. The developers revised the original DRPS measure and expanded it to include items on leadership self-efficacy and time spent by administrators on leadership functions as described in the previous section.

References:

Abel, M., J. Bella, T. Talen, and M. Magid. "Administrators Role Perceptions Survey Validation Study Technical Report." Wheeling, IL: McCormick Center for Early Childhood Leadership, August 2019.

McCormick Center for Early Childhood Leadership. "Administrator Role Perception Survey—Center Based." Wheeling, IL: McCormick Center for Early Childhood Leadership, April 2019.

McCormick Center for Early Childhood Leadership. "Director's Professional Development Needs Differ by Developmental Stage." Wheeling, IL: McCormick Center for Early Childhood Leadership, July 2018.

Rafanello, D., and P.J. Bloom. "The 1997 Illinois Directors' Study: A Report to the Robert R. McCormick Foundation." Concord, CA: Chapman University, August 1997.

Attributes of Leader Behavior Questionnaire (ALBQ), 1996 (also known as the attributes of leader behavior scales, the Attributes of Leader Behavior Instrument, and the Leader Behavior Scale or LBS)

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Management

Content

What leaders do (Foster respect and learning, establish vision)

What leaders bring (Values, beliefs, and attributes)

Administration characteristics

Respondent: Leader, employees (see Narrative)

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–self-report or report of others) (see Narrative)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 18 items

Administration interval: None described

Languages available: English, other (French, German, Japanese, Korean)

Technical information

Development sample

Locale: United States

Setting: Other (undergraduate and graduate business students)

Sample: (1) 205 undergraduate business students, (2) 94 graduate business students with substantial full-time work experience.

Year of development: 1996

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

2-Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Orlando Behling and James M. McFillen

Narrative[3]

Description: The Attributes of Leader Behavior Questionnaire (ALBQ)[4] is a self-administered survey that measures attributes of leaders' behaviors that relate to concepts of charismatic and transformational leadership. The developers state that the reporters may be a leader reporting on oneself (as a self-report), or subordinates or an independent observer. To date, the published information about the ALBQ is based on employee reports on their leaders (in some cases these were graduate students with full-time work experience who rated their current or most recent supervisor) and university undergraduate students watching short videos of actors portraying leaders. No data have been presented about use as self-report. The current version of the ALBQ is an 18-item survey. The current version of the ALBQ has six subscales: Displays Empathy, Dramatizes Mission, Projects Self-Assurance, Enhances (the Leader's) Image, Assures Followers of (the Followers') Competency, and Provides Opportunities to Experience Success. Each subscale has three items. The ALBQ is not specific to a particular field and is intended for use in a variety of fields of business.

The ALBQ reflects the "What leaders do" and "What leaders bring" components of the ExCELS theory of change. The Displays Empathy, Dramatizes Mission, Assures Followers of Competency, and Provides Opportunities to Experience Success subscales mainly align with the ExCELS component of "What leaders do." The Projects Self-Assurance and Enhances Image subscales, although to some extent also about leader actions, are more focused on the leader's general abilities and attributes and better align with the ExCELS component of "What leaders bring."

Uses of Information: The ALBQ was created primarily for leadership research. The developers' goals were to present a model that reflected key constructs in common across previous models of charismatic and transformational leadership. By operationalizing the model through measures, researchers could test relationships involving the model, such as leader behaviors, follower beliefs, and follower behaviors. In discussing the results of their study, the developers also list several ways managers in organizations can use the measure for development and improvement purposes, such as assessing training needs and evaluating training results.

Methods of Scoring: Each item is scored on a 5-point agreement scale with the options strongly agree (1), agree (2), neither agree nor disagree (3), disagree (4), and strongly disagree (5).[5] Three items describe undesirable behaviors and are reverse coded. Further details on calculating total and subscale scores are not given, although one study calculated subscale scores as the sum of the scores of the items in the subscale and defined the scale so "strongly agree" responses correspond to the highest number.

Interpretability: Assuming the developers' 5-point scale, lower scores on the ALBQ correspond to higher levels of charismatic and transformational leadership.

[3] The analysis described in this profile was done on the current, 18-item ALBQ, with the exception of the content validity research reported by the developers, which was done on a draft version of the measure.

[4] The developers do not provide a definitive name for this measure. It is also referred to by the developers or in the literature as the attributes of leader behavior scales, the Attributes of Leader Behavior Instrument, and the Leader Behavior Scale or LBS.

[5] The developers used this set of response options in their study (Behling and McFillen 1996). The subsequent study by a different group of researchers (McCann et al. 2006) also mentions a Likert scale ranging from strongly disagree to strongly agree, but said it had six response options, so the middle options appear to have been different.

Reliability:

(1) Internal consistency reliability: While creating the ALBQ, the developers tested it with multiple samples in the United States (Behling and McFillen 1996). Sample 1 consisted of undergraduate business students who completed the ALBQ after viewing videos of actors portraying leaders. Cronbach's alphas for four of six subscales ranged from 0.77 (Projects Self-Assurance) to 0.84 (Provides Followers with Opportunities to Experience Success).[6] Sample 2, graduate business students who rated their current or most recent supervisor, had Cronbach's alphas ranging from 0.71 (Enhances Leader's Image) to 0.86 (Assures Followers of Competency) for the six subscales.

McCann et al. (2006) conducted a further test of the ALBQ by asking 178 employees in 17 Australian organizations in a variety of business fields to rate their managers (of which there were 29). Cronbach's alphas ranged from 0.68 (Displays Empathy) to 0.85 (Assures Followers of Competency). The only alpha below 0.70 was for the Displays Empathy subscale.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers selected the six attributes of leader behavior to define the subscales based on general agreement of their importance from experts in the field. The developers drafted 66 potential items. Each member of a small group of expert reviewers (graduate business students in an MBA program and another similar field's program) individually assigned each item to one of the six subscales based on developer-provided subscale definitions and noted any problems they encountered in assigning items. As a result of this process, 19 items were dropped and others were reworded. After further testing these 47-items with a sample of undergraduate students, the developers dropped many items, leaving the 18 items in the current measure, which is the version described throughout this profile. (Samples 1 and 2 from Behling and McFillen (1996) and the Australian sample from McCann et al. (2006) all use the current, 18-item measure.)

(2) Construct/concurrent validity:

Construct validity: The developers conducted exploratory factor analyses with the 18-item ALBQ (Behling and McFillen 1996). For Sample 1 (undergraduate students who watched videos of actors portraying leaders), a five-factor model had the strongest fit. Three factors aligned perfectly with the Dramatizes Mission, Projects Self-Assurance, and Enhances Image subscales; the fourth factor partly aligned with the Provides Opportunities to Experience Success subscale; and both the Displays Empathy and Assures Followers of Competency subscales loaded onto the final factor. For Sample 2 (graduate students rating a supervisor from work), all 18 items loaded perfectly onto six factors that aligned with the six subscales, with all primary factor loadings greater than 0.55.

McCann et al. (2006) also conducted a factor analysis with their Australian sample and found that all but one of the 18 items had the highest loading on the factor corresponding to its subscale. The exception, an item from the Displays Empathy subscale, had a 0.16 loading for that subscale but a

[6] The other two subscales had items loading on one factor. However, a five-factor model does not completely align with the current six ALBQ subscales. The construct validity section describes how the five factors in this model aligned with the six ALBQ subscales.

loading of 0.49 under the Assures Followers of Competency subscale. Primary factor loadings ranged from 0.16 to 0.62 for the Displays Empathy subscale, 0.42 to 0.54 for Dramatizes Mission, 0.33 to 0.80 for Projects Self-Assurance, 0.59 to 0.79 for Enhances Image, 0.57 to 0.90 for Assures Followers of Competency, and 0.54 to 0.84 for the Provides Opportunities to Experience Success subscale.

Concurrent validity: With Sample 2, Behling and McFillen (1996) tested convergent validity by administering five constructs from other measures, each of which aligned with one of the six ALBQ subscales; the developers were unable to find a parallel construct for the Dramatizes Mission subscale. The five statistically significant ($p \leq .01$) correlation coefficients ranged between 0.50 (Projects Self-Assurance and individual prominence) and 0.71 (Provides Opportunities and supervisor support and participation) (evidence of convergent validity). The developers also analyzed correlations between each pair of ALBQ subscales and other constructs that do not cover similar content. In only one case was the correlation with a similar construct weaker than the correlations with a dissimilar construct. In many cases the correlations with dissimilar constructs were substantially weaker (evidence of divergent validity). However, there were several cases where these divergent correlations were only slightly weaker than for the convergent pair. Overall, the median significant correlation with dissimilar constructs was 0.30 with a range from −0.32 (Displays Empathy and recognition orientation) to 0.69 (the Provides Opportunities subscale with consideration and positive reward behavior). The developers concluded that these results demonstrated a "reasonable, though far from perfect," level of divergent validity (p. 181).

McCann et al. (2006) examined criterion-related validity with their Australian sample by analyzing relationships between leader behavior, as measured by the ALBQ, follower beliefs, as measured by three subscales of Inspiration, Awe, and Empowerment created by the ALBQ developers[7], and a measure of organizational commitment with three subscales for affective, continuance, and normative commitment. The analysis generally found significant associations between the ALBQ subscales and affective commitment, but not continuance or normative commitment. These associations were mediated by follower beliefs of Awe (most strongly) and Inspiration, but not Empowerment. Of the six ALBQ subscales, Projects Self-Assurance and Enhances Image had weaker associations with the other measures, compared to the other four ALBQ subscales.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): The ALBQ uses general terminology and is not setting specific, so it likely can be used in ECE settings without substantive adaptation. However, the ALBQ was tested with samples of business students in bachelor's and master's degree programs and its findings might not be generalizable to ECE settings where respondents are likely to be more diverse and have lower levels of education. Because it asks respondents to assess their manager or supervisor, it is most applicable to those serving in formal leadership roles, such as program or center directors.

Previous Version: None.

[7] The three follower belief subscales form the Follower Belief Questionnaire (also known as the Follower Belief Scale) created by the developers in the same study (Behling and McFillen 1996) as the ALBQ.

References:

Behling, O., and J.M. McFillen. "A Syncretical Model of Charismatic/Transformational Leadership." *Group & Organization Management*, vol. 21, no. 2, June 1996, pp. 163–191.

McCann, J.A.J., P.H. Langford, and R.M. Rawlings. "Testing Behling and McFillen's Syncretical Model of Charismatic Transformational Leadership." *Group & Organization Management*, vol. 31, no. 2, April 2006, pp. 237–263.

Authentic Leadership Inventory (ALI), 2011

Purpose and context

Purpose: Research/evaluation

Field: Management

Content

What leaders do (Foster respect and learning, establish vision)

What leaders bring (Interpersonal and team-building knowledge and skills; values, beliefs, and attributes)

Administration characteristics

Respondent: Employees

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 14 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States

Setting: Business office and other (undergraduate and graduate students)

Sample:

Sample 1: 499 undergraduates in management classes; average age 20 years; 59% male; 35% currently employed; 62% White, 20% Hispanic, 7% Black

Sample 2: 38 executive MBA students (employed full-time) and 190 mid-level employees; average age 32 years; 45% male; 58% White, 27% Hispanic, 6% Black

Year of development: 2011

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

2-Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Linda L. Neider and Chester A. Schriescheim

Narrative

Description: The Authentic Leadership Inventory (ALI) is a self-administered survey that measures followers' perceptions of a leader's behaviors related to authentic leadership. The ALI has 14 items across four subscales: Self-Awareness (3 items), Relational Transparency (3 items), Internalized Moral Perspective (4 items), and Balanced Processing (4 items). The ALI is not specific to a particular field and can be used in a wide range of settings. The ALI reflects the "What leaders do" and "What leaders bring" components of the ExCELS theory of change.

Uses of Information: The ALI can be used for leadership research. The developers explain that their goal was to develop a measure of authentic leadership that is theory based, reliable, and valid. Such a measure could be used to study authentic leadership, including its relationship with other leadership constructs, such as transformational leadership.

Methods of Scoring: Each item is scored on a 5-point agreement scale about the item content: disagree strongly (1), disagree (2), neither agree nor disagree (3), agree (4), and agree strongly (5). The developers do not give further details on calculating subscale scores, although it appears a subscale score is the average of the scores of the items in the subscale.

Interpretability: Higher scores on the ALI correspond to higher levels of authentic leadership.

Reliability:

(1) Internal consistency reliability: The first validation study (with Sample 1) asked undergraduate students to fill out the ALI for the two major party United States 2008 presidential candidates; the second validation study (with Sample 2) asked graduate business students (who were employed full-time) and a snowball sample of mid-level employees referred by the students to rate their current supervisor with the ALI. In all three cases, the Cronbach's alphas were above 0.70. For Sample 1, alphas ranged from 0.74 (Self-Awareness) to 0.85 (Relational Transparency, Internalized Moral Perspective, and Balanced Processing). For Sample 2, alphas ranged from 0.70 (Self-Awareness) to 0.82 (Balanced Processing).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers used the same theoretical framework and set of four subscales developed for the Authentic Leadership Questionnaire (ALQ; Avolio et al. 2018). The developers drafted 16 items, 8 of which were paraphrased from the 8 ALQ items made publicly available (out of 16 total) by the ALQ developers, in order to closely follow the theoretical framework on authentic leadership they had established. A sample of 40 undergraduate and 32 graduate business students were asked to rate each item for how well it fit the ALQ definitions of each of the four subscales, using a 5-point scale of none (0), hardly any (1), some (2), much (3), very much (4), and almost completely or completely (5). For each item, the developers compared the average rating on the intended subscale to each of the other three subscales using *t*-tests, and then conducted a principal components analysis. From this analysis, 14 items had averages and factor loadings that were clearly strongest for the intended subscale. The remaining 2 items did not have averages and factor loadings that were clearly strongest for the intended subscale; one was intended for the Self-Awareness subscale and the other for the Relational Transparency subscale. Later, after Sample 1, the developers dropped these 2 items.

(2) Construct/concurrent validity:

Construct validity: The developers conducted confirmatory factor analyses (CFA) with different samples and different numbers of items to conclude a 14-item measure with a second-order factor structure with four primary factors was most appropriate. For Sample 1, the developers tested several models using the original 16-item measure, including models allowing cross-loadings on the two items previously flagged by the content validity analysis as not clearly loading as well as with the other 14 items. Based on these CFA results and the content validity analysis, the developers dropped the two flagged items. The model fit with the 14-item version was acceptable and factor loadings on the principal factor ranged from 0.61 to 0.86. Inter-factor correlations ranged from 0.59 to 0.89.

For Sample 2, respondents used the final, 14-item version to rate their supervisors. With high inter-factor correlations, the developers tested a second-order factor structure. Both the four-factor model and the second-order factor model (with the four first-order factors loading onto the second-order factor) had acceptable fit. Neither model had a significantly stronger fit than the other. Factor loadings on the principal factor ranged from 0.46 to 0.80. Inter-factor correlations between the first-order factors ranged between 0.70 and 0.78, and the first-order factors' correlations with the second-order factor ranged between 0.83 and 0.89 ($p \leq .001$).

Concurrent validity: For Sample 2, the developers tested the convergent and divergent validity of the ALI with a modified version of the ALQ featuring only the 8 items publicly available at the time. Using confirmatory factor analysis of multitrait-multimethod matrix to look at associations among the different subscales and measures simultaneously, the developers found evidence of convergent validity. The correlations among the trait factors ranged from 0.62 to 0.82 ($p \leq .01$).

For both Sample 1 and Sample 2, the developers tested the divergent validity of the ALI with a measure of transformational leadership—the Transformational Leadership Inventory or TLI, which has 23 items and six subscales (Podsakoff et al. 1990)—because their theoretical framework considers authentic and transformational leadership to be related but separate constructs.

Using confirmatory factor analysis with Sample 1, the developers found that a 10-factor model with 4 factors for the ALI and 6 factors for the TLI had a stronger fit compared to higher-order models that combined the factors underlying constructs of authentic and transformational leadership. However, with Sample 2 only some fit statistics were acceptable and, in both Sample 1 and 2, multiple inter-factor correlations between the four ALI and six TLI subscales were greater than 0.70, suggesting convergent rather than divergent validity as the developers hypothesized.

With Sample 2, the developers also examined the concurrent criterion validity of the ALI against three outcomes: satisfaction with supervision, general job satisfaction, and organizational commitment. Correlations were statistically significant at $p \leq .001$. The highest bivariate correlations with ALI subscales was with satisfaction with supervision ($r = 0.58$ [Self-Awareness] to 0.62 [Balanced Processing]) for the four ALI subscales), followed by general job satisfaction ($r = 0.39$ [Relational Transparency] to 0.48 [Balanced Processing]), and lowest with organizational commitment ($r = 0.28$ [Self-Awareness] to 0.33 [Moral Perspective]). Similarly, using a multivariate regression analysis for each outcome, where each model includes all four subscales, the developers found that all four subscales were significantly associated with satisfaction with supervision, but only the Balanced Processing subscale was significantly associated with general job satisfaction, and only Internalized Moral Perspective was significantly associated with organizational commitment.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: For Sample 1 and Sample 2, the developers compared ALI responses to responses on a measure of tendencies to offer socially desirable responses (the Balanced Inventory of Desirable Responding; Paulhus 1998). Of the 12 comparisons for each subscale across the two samples to the social desirability measure, one was significant, though substantively small ($r = 0.09$ [Self-Awareness; $p \leq .05$; $n = 499$]). The developers concluded that ALI responses are not affected by social desirability bias.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): The ALI uses general terminology and is not setting specific, so it likely could be used in ECE settings without substantive adaptation. However, the ALI was tested with samples of majority-White business and management students in bachelor's and master's degree programs and its findings might not be generalizable to ECE settings where respondents are likely to be more diverse and have lower levels of education. Because it asks respondents to assess a leader who is usually their direct supervisor, it is most applicable to those serving in formal leadership roles, such as program or center directors.

Previous Version: None.

References:

Neider, L.L., and C.A. Schriesheim. "The Authentic Leadership Inventory (ALI): Development and Empirical Tests." *The Leadership Quarterly*, vol. 22, no. 6, 2011, pp. 1146–1164.

Authentic Leadership Questionnaire (ALQ), 2018

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Management

Content

What leaders do (Foster respect and learning, establish vision)

What leaders bring (Values, beliefs, attributes)

Administration characteristics

Respondent: Manager, employees

Level of measure: Site, individual

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 10–15 minutes, 16 items

Administration interval: None described

Languages available: English, Spanish, other (see measure website for details)

Technical information

Development sample[8]

Locale: United States (Northeast)

Setting: Business office

Sample: 224 employees from one large manufacturing business, average age 45 years, 80% male, 100% college educated

Year of development: 2008

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available (see Narrative)

Availability

4-Permission required, with costs (no costs for certain noncommercial research uses)
Material, training, and scoring costs:

- $2.50 per person, min. 50 (paper or nonpublisher survey system)
- $2.50 or $4 per person, min. 20 (publisher online system, depending on form)
- $15 to $200 reports per person or group (publisher online system, depending on type of report)

Developer(s)/publisher contacts

Developer(s): Bruce J. Avolio, William L. Gardner, and Fred O. Walumbwa
Publisher: Mind Garden, Inc.
650-322-6300
www.mindgarden.com
Measure website: https://www.mindgarden.com/69-authentic-leadership-questionnaire

[8] Developers also used a second sample in China.

Narrative

Description: The Authentic Leadership Questionnaire (ALQ) is a self-administered survey that measures the leader's behaviors that relate to the construct of authentic leadership. There are two versions with the same items: a leader's self-report and employee ratings of the leader. The ALQ is a 16-item survey that takes 10 to 15 minutes to administer. It has four subscales: Transparency (5 items), Ethical/Moral (4 items), Balanced Processing (3 items), and Self-Awareness (4 items). The items are not specific to a particular field; the ALQ can be used in a wide range of settings. The ALQ was initially assessed for reliability and validity using two samples, one in the United States and one in China. The ALQ can be administered and scored online or on paper using the publisher's assessment system. The items in the ALQ's four subscales focus on authentic leadership behaviors, which reflect the "What leaders do" component of the ExCELS theory of change. However, the underlying construct of authentic leadership can also be seen as an attribute of leadership that would fit into the "What leaders bring" component of the theory of change.

Uses of Information: The developers state that the ALQ can be used to "assess and develop authentic leadership behaviors" (Avolio et al. 2018a, p. 4). The developers describe how organizations can use the measure to develop and improve these behaviors. Researchers can use the measure to explore the associations of authentic leadership with positive outcomes such as psychological capital.

Methods of Scoring: Each item is scored on a 5-point frequency scale for behavior happening not at all (0); once in a while (1); sometimes (2); fairly often (3); or frequently, if not always (4). Scores for each subscale are calculated by averaging the score for each item, producing a raw score ranging from 0 to 4. The developers list the percentile associated with each total raw score[9] in the norming sample.[10] Scores can be calculated for an individual or for a group, based on either leaders' self-report or on employees rating leaders.

Interpretability: Higher ALQ raw scores indicate higher levels of authentic leadership behaviors. The percentile ranks for raw scores show the leader's position relative to the norming sample. The publisher's online assessment system includes reports providing ALQ results for individuals or groups. These reports include interpretations of scores, including comparisons between self-ratings and ratings from others, and summaries of scores for groups of participants.

Reliability:

(1) Internal consistency reliability: The initial validation study published in 2008 examined Cronbach's alpha scores for each of the four subscales of the 16-item ALQ in a sample from the United States (described in the profile overview) and a sample from China. The United States sample's scores ranged from 0.76 (Ethical/Moral) to 0.92 (Self-Awareness).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

[9] Although this list includes total raw scores, there is no information on how those total scores are calculated (for example, if they are averages of all items or averages of the four subscales).

[10] The norming sample differs from the initial development sample used to assess reliability and construct validity per the ALQ manual (Avolio et al. 2018a).

Validity:

(1) Content validity: During initial development of the ALQ, the developers conducted a literature review and worked with a group of other leadership researchers to define the theoretical constructs making up authentic leadership. This review led to the four subscales included in the measure. The developers also assessed content validity by asking expert reviewers (other faculty and doctoral students) to categorize each item into one of the four subscale constructs. At this stage, the developers dropped 6 of 22 items because they lacked at least 80 percent agreement on subscale categorization. The remaining 16 items comprise the version of the ALQ that was used for all subsequent work.

(2) Construct/concurrent validity:

Construct validity: A recent update of the confirmatory factor analysis (CFA) from the initial United States validation study (Avolio et al. 2018b), intended to address criticism of the methods and reporting from the original CFA, determined that a higher-order model (four factors that correspond to the four subscales and are under an overall factor for authentic leadership) had a stronger fit than a single-factor model (one overall authentic leadership factor) or a first-order model (four factors corresponding to the four subscales, without a higher-order factor). However, the higher-order model did not outperform another type of first-order model or a bifactor model. The higher-order model had an acceptable fit on most, but not all, criteria. As part of this update, the developers also note that subsequent studies have found evidence that a higher-order model is appropriate and conclude that researchers should continue to explore which models fit best during future studies.

Concurrent validity: The developers summarize several studies that demonstrate criterion-related validity based on relations between authentic leadership—usually, but not always, measured by the ALQ—and employee attitudes and behaviors. A 2016 meta-analysis of 100 samples (total of 25,425 individuals) found relationships between authentic leadership measures (to include ALQ and others) and six key outcomes: (1) follower job satisfaction, (2) follower satisfaction with leader, (3) task performance, (4) organizational citizen behavior, (5) group or organizational performance, and (6) rated leader effectiveness. It also found relationships with other outcomes, such as positive relationships with trust in leadership and employee creativity, and a negative relationship with burnout and stress. Other studies, with samples ranging from 117 to more than 300 individuals, found relationships between authentic leadership and similar follower and organizational outcomes, such as job satisfaction, organizational commitment, satisfaction with supervisor, and psychological ownership.

(3) Predictive validity: Although the meta-analysis and other studies have examined relationships between authentic leadership and various outcomes, it is not clear if any studies measured outcomes later in time than they measured authentic leadership.

Bias Analysis: No information available.

Training Support: The developers provide information on administration and scoring. The ALQ can be administered through the publisher's online system, which handles data collection and analysis and reporting of results.

Key Considerations for Early Care and Education (ECE): The ALQ uses general terminology and is not setting specific, so it likely can be used in ECE settings without substantive adaptation. Because it either asks leaders to rate themselves or others to rate their leader, it is most applicable to those serving in formal leadership roles, such as program or center directors. The language of a few items refers to

"others"; for ECE settings this might need to be clarified as "other adults" or "other staff in my organization" to avoid respondents thinking about adult-child interactions. Also, the primary reliability and validity evidence for the ALQ comes from a sample of predominantly male individuals who all had college degrees. As a result, its findings might not be generalizable to ECE settings where respondents are overwhelmingly female and likely to have lower levels of education.

Previous Version: None.

References:

Avolio, B.J., W.L. Gardner, and F.O. Walumbwa. *Authentic Leadership Questionnaire (ALQ) Manual*. Menlo Park, CA: Mind Garden, Inc., 2018a.[11]

Avolio, B.J., T. Wernsing, and W.L. Gardner. "Revisiting the Development and Validation of the Authentic Leadership Questionnaire: Analytical Clarifications." *Journal of Management*, vol. 44, no. 2, 2018b, pp. 399–411.

[11] We obtained additional details about the development sample, which is otherwise described in the ALQ manual, from the initial validation study (Walumbwa, F.O., B.J. Avolio, W.L. Gardner, T.S. Wernsing, and S.J. Peterson. "Authentic Leadership: Development and Validation of a Theory-Based Measure." *Journal of Management*, vol. 34, no. 1, 2008, pp. 89–126).

Collective Leadership Survey, 2006

Purpose and context

Purpose: Research/evaluation

Field: Management

Content

What leaders do† (Promote quality practices)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

† Includes what staff as leaders do

Administration characteristics

Respondent: Other (road crew team member)

Level of measure: Site/team

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 1 (not described)

Time/length: 25 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States

Setting: Other (state department of transportation)

Sample: 277 road crew team members in 52 teams (5% female, mean crew size of 5.5 employees, mean tenure with organization of 14.3 years)

Year of development: 2006

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70).

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

Availability

2 – Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Nathan J. Hiller, David V. Day, and Robert J. Vance

Narrative

Description: The Collective Leadership Survey is a self-administered survey, developed for a sample of roadwork teams, which measures how often team members share in tasks and relationships that are relevant to collective leadership. The survey has 25 items total, comprising four subscales: Planning and Organizing, Problem Solving, Support and Consideration, and Development and Mentoring. There are 6 items per subscale, except for Problem Solving, which has 7 items. The Planning and Organizing and Problem Solving subscales measure aspects of "What leaders do" across the team. The Support and Consideration and the Development and Mentoring subscales capture information on the workplace aligned with "Center culture, climate, and communication."

Uses of Information: The developers state that that the Collective Leadership Survey can be used to examine the existence and performance correlates of collective team leadership.

Methods of Scoring: Respondents reported on their perceptions of collective leadership in their teams using a 7-point scale: never (1) to always (7). Each of the Collective Leadership Survey subscales is an average of items within a specific subscale. The developers then indicate that each subscale can be aggregated to the team level (but the exact approach is not described).

The developers considered the appropriateness of aggregating scores to the site level in several ways. The intraclass correlations (ICCs) were significant ($p \leq .05$) for three subscales (and ranged from 0.08 [Planning and Organizing; and Problem Solving] to 0.10 [Development and Mentoring]), the exception being Support and Consideration (ICC = 0.01). The reported ICCs fall within the typical range for adequate group reliability. Eta-squared statistics for group-level effects ranged from 0.20 (Support and Consideration) to 0.26 (Development and Mentoring) (with a threshold noted by the developers of 0.20). This indicates that 20 to 26 percent of the variance was between groups. The developers reported that corrected r_{wg} values for inter-rater agreement within teams were, on average, below 0.70 as a threshold. Collectively, the developers claimed adequate support for site-level scores across individuals, but acknowledge that the within group correlation in the study indicates limitations in the shared understanding of the individuals within the teams.

Interpretability: A higher score on the subscale indicates greater element attributes associated with team participation in collective leadership tasks and relationships.

Reliability:

(1) Internal consistency reliability: The developers reported composite reliability (CR) estimates for the subscales as a result of the confirmatory factor analyses: Planning and Organizing (CR = 0.96), Problem Solving (CR = 0.96), Support and Consideration (CR = 0.93), and Development and Mentoring (CR = 0.94). For comparative purposes, the developers reported Cronbach's alphas for the four subscales as well: Planning and Organizing (α = 0.96), Problem Solving (α = 0.96), Support and Consideration (α = 0.92), and Development and Mentoring (α = 0.93).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: Items were informed by Ohio State and Michigan research studies cited by the developers (Cartwright and Zander 1960; Hemphill and Coons 1957), which showed leadership involves task- and relationship-oriented behaviors (Yunker and Hunt 1976). To determine the specific behaviors relevant to collective leadership, the developers reviewed the Managerial Practices Survey (MPS; Yukl and Lepsinger 1990) as a source for roles of effective managers and leaders. The developers removed those behaviors outside the scope of their sample. Fourteen expert reviewers from the Industrial Organization Psychology Ph.D. program reviewed the survey and eliminated 3 of the 28 items, resulting in 25 items retained in the survey.

(2) Construct/concurrent validity:

Construct validity: The developers assessed the factor structure of the Collective Leadership Survey items through a comparison of six nested models using confirmatory factor analysis at the individual level. The model specifying correlations between four collective leadership subscales had acceptable fit and better fit than other models. The inter-factor correlations between the subscales were significant at $p \leq .05$ and ranged from 0.70 (Development and Mentoring with Planning and Organizing) to 0.88 (Problem Solving with Planning and Organizing). The standardized factor loadings per subscale ranged from 0.85 to 0.91 (Planning and Organizing), 0.83 to 0.92 (Problem Solving), 0.71 to 0.88 (Support and Consideration), and 0.83 to 0.88 (Development and Mentoring). The developers also tested a model with two second-order factors (task and relationship), which also demonstrated acceptable fit (information on the subscale loadings to the second-order factors and the correlations between them were not available).

Concurrent validity: The developers examined the correlations of collective leadership scores with team member reports of collectivism and power distance (at the team level) and with supervisor ratings of team effectiveness in planning and organizing, problem solving, support and consideration, and development and mentoring. Correlations between the collective leadership subscale scores and outcomes were as follows: 0.39 (Problem Solving) to 0.46 (Development and Mentoring) with collectivism (all correlations significant at $p \leq .05$), 0.08 (Problem Solving) to 0.12 (Planning and Organizing) with power distance, 0.18 (Planning and Organizing) to 0.30 (Development and Mentoring; $p \leq .05$) with team effectiveness in planning and organizing, 0.15 (Problem Solving) to 0.27 (Development and Mentoring) with team effectiveness in problem solving, 0.28 (Problem Solving) to 0.45 (Development and Mentoring; $p \leq .05$) with team effectiveness in support and consideration, 0.22 (Problem Solving) to 0.38 (Development and Mentoring; $p \leq .05$) with team effectiveness in development and mentoring, 0.25 (Support and Consideration) to 0.37 (Development and Mentoring; $p \leq .05$) with overall team effectiveness, and 0.26 (Problem Solving) to 0.41 (Development and Mentoring; $p \leq .05$) with a composite supervisor rating.

The developers also conducted stepwise hierarchical regressions using the four subscales to predict supervisor ratings of effectiveness in planning and organizing, problem solving, support and consideration, and development and mentoring. Results indicated that the Support and Consideration subscale and the Development and Mentoring subscale significantly predicted the corresponding supervisor-rated effectiveness (for example, the team Support and Consideration subscale score is a significant predictor of effectiveness on support and consideration), providing evidence of convergent validity the developers noted. In addition, for a given effectiveness rating, adding the other noncorresponding collective leadership subscales to the models did not explain additional variance in the effectiveness ratings, suggesting evidence of divergent validity the developers noted.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): The developers created the survey specifically for their sample and removed roles from the MPS that were outside the scope of the sample (such as performance evaluation) while developing the Collective Leadership Survey. Therefore, the subscales of Planning and Organizing, Problem Solving, Support and Consideration, and Development and Mentoring are all relevant to center-based ECE settings. Although the terminology about the type of work is general, some items may need revisions to better correspond to the duties of teaching staff in ECE settings (for example, on flow of work).

Previous Version: None.

References:

Hiller, N.J., D.V. Day, and R.J. Vance. "Collective Enactment of Leadership Roles and Team Effectiveness: A Field Study." *The Leadership Quarterly*, vol. 17, no. 4, summer 2006, pp. 387–397. doi:10.1016/j.leaqua.2006.04.004.

Conger-Kanungo Scale of Charismatic Leadership (C-K scale), 1997

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Management

Content

What leaders do (Foster respect and learning, establish vision)

What leaders bring (Interpersonal and team-building knowledge and skills; values, beliefs, and attributes)

Administration characteristics

Respondent: Employees

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 20 items

Administration interval: None described

Languages available: English, other (French)

Technical information

Development sample

Locale: United States and Canada

Setting: Business office

Sample:

Sample 1: 488 respondents out of a sample of 750 managers from four U.S. and Canadian corporations; 89% male, average age 42 years, 62% with English as first language and 25% with French as first language, about half with organizational tenure of 12+ years, widely varying education levels and incomes

Sample 2: 103 middle- and senior-level employees from one U.S. corporation; 66% male, average age 40 years, average organizational tenure 11 years, 97% college degree or higher

Sample 3: 252 managers from one U.S. corporation; 94% male, average age 43 years, average organizational tenure 14 years, 80% college degree or higher

Year of development: 1994, with Sample 1 (additional validation in 1997 with Sample 1 and 2, and in 2000 with Sample 3)

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

2-Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Jay A. Conger and Rabindra N. Kanungo

Narrative

Description: The Conger-Kanungo Scale of Charismatic Leadership (C-K scale) is a self-administered survey that measures followers' or subordinates' perceptions of a leader's behaviors that may involve charismatic leadership. The current version of the C-K scale has 20 items in five subscales: Strategic Vision and Articulation, Sensitivity to the Environment, Sensitivity to Members' Needs, Personal Risk, and Unconventional Behavior.[12] Strategic Vision and Articulation has 7 items; Sensitivity to the Environment has 4 items and remaining subscales each have 3 items. The C-K scale is not specific to a particular field; it has primarily been used with managerial staff in corporate and business firms for rating their supervisors. The C-K scale reflects the "What leaders do" and "What leaders bring" components of the ExCELS theory of change. Items in all five subscales reflect both components to some degree, but measure different combinations of components. The Sensitivity to Members' Needs items primarily measure "What leaders do;" the Sensitivity to the Environment, Personal Risk, and Unconventional Behavior subscales primarily measure "What leaders bring;" and the Strategic Vision and Articulation subscale measures both components in similar amounts.

Uses of Information: The C-K scale is primarily intended for leadership research. It was developed to measure charismatic leadership, based on a model of charismatic leadership behavior created by the same developers. That model was based on increased interest in organizational research on transformational leadership and Max Weber's historical theories of charismatic leaders. The developers also note organizations could use the C-K scale to train, develop, and select leaders by focusing on the attributes and behaviors of the scale items.

Methods of Scoring: Each item is scored on a 6-point scale where the lowest score equals "very uncharacteristic" and the highest score equals "very characteristic." The developers report total and subscale scores. It appears subscale scores are the average of the scores of the items in the subscale (but not explicitly stated).

Conger et al. (1997) administered the C-K scale to a non-U.S. sample of 49 pairs of employees who reported to the same leader in a large national corporation in India. This study found high levels of agreement within the pairs regarding the leaders' C-K scale scores for the overall scale ($r = 0.84$) and for four of the five subscales ($r = 0.81$ for Strategic Vision and Articulation, 0.82 for Sensitivity to the Environment, 0.79 for Personal Risk, and 0.71 for Unconventional Behavior). Agreement for the Sensitivity to Members' Needs subscale was moderate ($r = 0.59$), suggesting different perceptions among some employees. All correlations were significant at $p \leq .001$.

Interpretability: Higher scores on the C-K subscales scale correspond to higher levels of charismatic leadership.

[12] The information from this profile is drawn primarily from Conger et al. (1997), who reanalyzed data from Sample 1 (originally studied by Conger and Kanungo [1994]) using the current version of the measure (the original study used a previous version of the measure that had a sixth subscale and more items in some of the current subscales). Conger et al. (1997) also conducted three new studies (Sample 2 and two non-U.S. samples) with the current version. See the Previous Version section for more details on the previous version of the measure.

Reliability:

(1) Internal consistency reliability: Using Sample 1 (originally analyzed by Conger and Kanungo 1994), the developers reanalyzed reliability using only the subscales and items that comprised the current version of the C-K scale (Conger et al. 1997). This analysis found that the Cronbach's alpha ranged between 0.74 (Unconventional Behavior) and 0.87 (Strategic Vision and Articulation) for the five subscales and was 0.88 for the overall scale. For Sample 2 (Conger et al. 1997), the developers found similar results, with alphas ranging from 0.72 (Sensitivity to the Environment) to 0.86 (Strategic Vision and Articulation) for the five subscales and 0.87 for the overall scale. For Sample 3 (Conger et al. 2000), the developers found slightly lower reliability scores: the Sensitivity to the Environment subscale had an alpha of 0.64, although the other four ranged from 0.71 (Unconventional Behavior) to 0.84 (Personal Risk), and the alpha for the overall scale was 0.82.

(2) Test-retest reliability: The original 25-item version of the C-K scale was administered twice, two weeks apart (Conger and Kanungo 1994). The test-retest reliability coefficients ranged from 0.69 (Does Not Maintain Status Quo) to 0.84 (Sensitivity to Members' Needs) for the six subscales and 0.69 for the overall scale. This study was separate from the original study of Sample 1, and had a sample of 75 respondents; the developers do not provide any information about the setting or characteristics of this separate sample.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers originally constructed 49 items based on a literature review and other previous research. After piloting the items with a sample of 120 business employees, the developers eliminated ambiguous or redundant items, leaving the original, 25-item version of the C-K scale. Later, the developers dropped five more items based on reanalysis of the original validation study and examination of the items for overlap. The current version of the scale has 20-items.

(2) Construct/concurrent validity:

Construct validity: Conger et al. (1997 and 2000) conducted three confirmatory factor analyses of the five-factor model with the current 20-item C-K scale, first reanalyzing data from Sample 1 (488 managers, some of whom spoke French and were given a French-language version of the measure), second using data from Sample 2 (103 employees), and third using Sample 3 (252 managers). Separate analyses of data from Sample 1 and Sample 2 both found that this model had an adequate fit based on one fit statistic (TLI) but not a second (NFI). For Sample 3, the developers compared the five-factor model to a single-factor model, finding the five-factor model to have a better fit.

For Sample 3, the developers performed an exploratory factor analysis of the five-factor model before conducting the confirmatory factor analysis. The exploratory factor analysis found that the items loaded onto the five factors as expected. Of the 20 factor loadings for the primary factor, 12 were 0.70 or higher, and all 20 were 0.50 or higher (the developers do not report factor loadings for any of the confirmatory factor analyses).

For Sample 3, the developers also examined correlations between each of the five subscales. Of the 10 correlations, 8 were statistically significant, positive, and ranged from 0.16 (Personal Risk with Sensitivity to Members' Needs) to 0.42 (Sensitivity to the Environment with Strategic Vision and Articulation). The other 2 correlations were negative (−0.04 and −0.20, although only the latter was statistically significant); both involved the Unconventional Behavior subscale, which the developers did not have theoretical reasons to expect either positive or negative associations with other leadership measures, and Sensitivity to the Environment, and Sensitivity to Members' Needs, respectively.

Concurrent validity: For the reanalysis of Sample 1 and the analysis of Sample 2, the developers analyzed relationships between the C-K subscales and other measures of leadership: items from the Bass charisma scale (Bass 1985), subscales from the Managerial Practices Survey (Yukl 1988), and measures focused on task orientation, people orientation, and participative orientation. The developers hypothesized that some relationships between C-K subscales and these other leadership measures would demonstrate convergent validity and that others would show divergent validity based on whether they covered similar constructs or not. The developers expected to find statistically significant relationships with certain other leadership measures for the subscales for Strategic Vision and Articulation, Sensitivity to the Environment, and Sensitivity to Members' Needs, but not for the Personal Risk and Unconventional Behavior subscales, which they regard as representing constructs not found in other leadership measures.

For both samples, most although not all hypotheses about statistically significant relationships were confirmed by the results, providing evidence of convergent and divergent validity. The Strategic Vision and Articulation, Sensitivity to the Environment, and Sensitivity to Members' Needs subscales were in most cases correlated with the other leadership measures for which there was a theoretical connection. Contrary to the developers' expectations, the Personal Risk and Unconventional Behavior subscales were in some cases significantly associated with other leadership measures (negatively for some). However, these relationships were weaker compared to those involving the other C-K subscales.

Conger et al. (1997) also examined convergent validity with a non-U.S. sample of 49 pairs of employees who reported to the same leader in a large national corporation in India. Each pair independently assessed the leader on both the C-K scale and the Bass charisma scale. The correlations between an employee's rating of his or her leader on the C-K scale and on the Bass charisma scale were 0.72 for half of each pair and 0.60 for the other half, evidence of convergent validity.

For Sample 3, the developers studied relationships between charismatic leadership, as measured by the C-K scale, and six follower outcomes involving their feelings about their leader, their work identity and performance, and their empowerment. The study found that half of the follower outcomes were significantly and directly related to charismatic leadership, whereas the other half were mediated through other follower outcomes. These relationships were driven by the C-K subscales of Strategic Vision and Articulation, Sensitivity to the Environment, and Sensitivity to Members' Needs; the Personal Risk and Unconventional Behavior subscales were not independently related to the outcomes.

Finally, Conger et al. (1997) conducted discriminant analysis using a sample of Canadian political party members to see if the C-K scale could distinguish between charismatic and noncharismatic leaders. A panel of 10 party members rated five leadership candidates on a single item of general

charisma. A second group of 71 party members filled out the C-K scale for the most and least charismatic candidates. The most charismatic candidates received significantly higher scores than the least charismatic candidates on four of the five C-K subscales, although the difference was smaller for Personal Risk. There was effectively no difference in scores on the fifth subscale, Unconventional Behavior.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): The C-K scale uses general terminology and is not setting specific, so it likely can be used in ECE settings without substantive adaptation. However, the C-K scale was tested with samples of predominantly male, highly educated corporate employees in United States settings or in international settings. As a result, its findings might not be generalizable to United States ECE settings where respondents are almost all female and tend to have lower levels of education. Because it asks respondents to assess their manager or supervisor, it is most applicable to those serving in formal leadership roles, such as program or center directors.

Previous Version: The original version of the C-K scale was longer, with 25 items comprising six subscales. Further development led to dropping the sixth, 2-item subscale ("Does Not Maintain the Status Quo") and 3 other items whose content was already reflected in other items.

References:

Conger, J.A., and R.N. Kanungo. "Charismatic Leadership in Organizations: Perceived Behavioral Attributes and Their Measurement." *Journal of Organizational Behavior*, vol. 15, 1994, pp. 439–452.

Conger, J.A., R.N. Kanungo, S.T. Menon, and P. Mathur. "Measuring Charisma: Dimensionality and Validity of the Conger-Kanungo Scale of Charismatic Leadership." *Canadian Journal of Administrative Sciences*, vol. 14, no. 3, 1997, pp. 290–302.

Conger, J.A., R.N. Kanungo, and S.T. Menon. "Charismatic Leadership and Follower Effects." *Journal of Organizational Behavior*, vol. 21, 2000, pp. 747–767.

Distributed Leadership Inventory (DLI), 2009

Purpose and context

Purpose: Research/evaluation

Field: K–12 education

Content

Who leaders are (Structure, leadership roles)

What leaders do† (Promote quality practices, foster respect and learning, establish vision)

† Includes what teachers as leaders do

Administration characteristics

Respondent: Principal, teachers

Level of measure: Site, group, individual

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 23 items (13 of the items can be asked multiple times, separately for different role types of leadership team)

Administration interval: None described

Languages available: English[13]

Technical information

Development sample

Locale: United States (Central New Jersey), original sample was in Belgium

Setting: School

Sample: 162 middle school (grades 6–8) teachers from five schools of average size 1,060 students: 77% female, 78% more than 10 years' experience

Belgian sample: 1,902 staff [47 principals, 85 assistant principals, 248 teacher leaders, and 1,522 teachers of students ages 14–16] from 46 secondary schools of minimum size 600 students: 55% female, average age 41 years, average job experience 13 years

Year of development: 2009 (for Belgian sample; 2018 for U.S. sample)

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available (primarily for non-U.S. sample)

-Predictive validity: Not available

Availability

2-Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Hester Hulpia, Geert Devos, and Yves Rosseel

[13] The DLI was originally administered with a sample in Belgium; the measure and study results were published in English (Hulpia et al. 2009), but the language in which it was administered is unstated.

Narrative

Description: The Distributed Leadership Inventory (DLI) measures the distribution of leadership characteristics and functions between leadership team members. It is a self-administered survey that assesses the extent to which elements of leadership occur and are distributed within staff at K–12 schools. The DLI was originally developed and tested with samples from Belgium, but it has recently been used in the United States with teachers in a middle school setting (DeMarco 2018). During its original development, the measure assessed distributed leadership involving teacher leaders, assistant principals, and principals; all three types of staff were surveyed, as were other teachers at their schools (Hulpia et al. 2009). In the study conducted in the United States (DeMarco 2018), only classroom teachers were surveyed, and they were asked about the leadership team at their school as a whole. Because of this limitation, the profile provides more information on the Belgium sample. The measure comprises 23 items and features three subscales: Support (10 items), Supervision (3 items), and Coherent Leadership (10 items). The Support and Supervision subscales can be asked repeatedly for each individual or group leader role, whereas the Coherent Leadership subscale is always asked collectively about all staff serving as leaders. The items cover content from the "What leaders do" section of the ExCELS theory of change. The ExCELS theory of change's "Who leaders are" section is covered by scores that compare item results across the different types of administrator and teacher groups. For example, different scores for different members of the school's leadership team can indicate the degree to which each team member is involved in leadership activities (See the Methods of Scoring and Interpretability sections for more details).

Uses of Information: The DLI was designed to research leadership team characteristics and functions, as well as the distribution of leadership among school leaders, such as principals, assistant principals, and teacher leaders.

Methods of Scoring: Responses on items are rated on 5-point response scales: Supervision and Support subscales from never (0) to always (4) and Coherent Leadership from strongly disagree (0) to strongly agree (4).[14] Researchers can score the DLI by calculating the mean of the items within a subscale. To calculate a total score, one averages the subscale mean values. If the Support and Supervision subscales were asked separately for different types of staff (such as principal, assistant principal, and teacher leader), then scores can be calculated for each staff type. These scores can reflect the leadership characteristics of a group when there is more than one person per staff type or an individual when there is only one person for the staff type. Three kinds of site-level scores can be calculated for the Support and Supervision subscales: (1) an average of mean scores across all staff types (which, like those scores, can range from 0 to 4), (2) a maximum score that uses the score of the highest-rated staff type (also ranges from 0 to 4), and (3) a leadership distribution score that is based on how similar scores are for each staff type, ignoring the absolute values of the scores, meaning that it is a different kind of score with a different range (0 to 6). The Coherent Leadership subscale, which is asked once about all staff serving as leaders, provides a site-level score, based on the average of scores across all respondents at the site.

Interpretability: Higher scores on the Support and Supervision subscales indicate that leaders engage in those functions to a greater extent, whereas higher scores on the Coherent Leadership subscale indicate a better-functioning leadership team. For the leadership distribution score, higher scores indicate more equal distribution of leadership within the leadership team and lower scores indicate more centralization of leadership (Hulpia and Devos 2009).

[14] In the U.S. study (DeMarco 2018), the author used the strongly disagree (0) to strongly agree (4) response scale for all items, including the Supervision and Support subscales.

Reliability:

(1) Internal consistency reliability: For the United States sample, Cronbach's alpha was calculated for each subscale of the DLI to measure reliability (0.87 for Support, 0.84 for Supervision, and 0.92 for Coherent Leadership; DeMarco, 2018). For the original, non-U.S. sample, Cronbach's alphas were above 0.90 for all subscales (0.91 for teachers, 0.93 for assistant principals and principals under the Support subscale; and 0.91 for the Coherent Leadership subscale) except the Supervision subscale (0.83 for principals, 0.85 for assistant principals, and 0.79 for teacher leaders).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers used a theoretical framework based on relevant literature and existing measures to create the DLI. The developers sent a draft version of the DLI to a group of 16 teachers, teacher leaders, principals, and policymakers for review, and tested it in two schools. The developers assessed feedback related to item complexity and administration feasibility and refined the measure by making minor edits to item text.

(2) Construct/concurrent validity:

Construct validity: The United States sample did not examine the construct validity, but the original non-U.S. study conducted an exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA), each with a randomly selected half of the sample (n = 951; Hulpia et al. 2009). The developers conducted separate factor analyses for leadership function items and for the items on leadership team characteristics. They also examined item statistics and content similarity. Based on their analyses, the developers dropped six leadership function items and three leadership team characteristics items. CFA models with these final sets of items had acceptable fit confirming the current subscales in both the CFA subsample and the EFA subsample.

Because inter-factor correlations between the Support and Supervision factors were moderate to high (r = 0.55 for teacher leaders to 0.70 for principals, n = 951), the developers tested a one-factor model but found that the two-factor model outperformed the one-factor version. Finally, factor loadings for the leadership function model were similar for each staff type and all were greater than 0.60.

The developers also examined correlations involving the different kinds of total scores for the Support and Supervision subscales (n = 1,902). Correlations between the average scores and the maximum scores, and between average scores and leadership distribution scores were moderate to high (0.55 [maximum Supervision and average Support] to 0.86 [maximum Support and average Support] and 0.30 [leadership distribution Supervision and average Support] to 0.63 [leadership distribution Supervision and average Supervision]). The correlations between the leadership distribution scores and the maximum scores were less than 0.18. Correlations between the Coherent Leadership scores and the different kinds of total scores for the Support and Supervision subscales ranged from 0.08 to 0.67.

Concurrent validity: As part of the (non-U.S.) validation study, the developers examined the relationship between DLI constructs and school leaders' job satisfaction using responses from principals and assistant principals (n = 130). Multiple regression analysis models included the leadership distribution score for Support, the leadership distribution score for Supervision, the score for Coherent Leadership, the score from a separate measure on participative decision making, and characteristics of the leader and school. The developers found that the Coherent Leadership score had the strongest association with job satisfaction (evidence of concurrent validity) and that the leadership distribution scores for Support and Supervision were not significant for this sample.

In the United States study, the author studied the correlations between the DLI subscales and the Teacher Self-Efficacy Scale (Schwarzer et al. 1999) and the School Culture Survey (Gruenert and Valentine 2006), all based on teacher surveys (n = 162).[15] The Support and Supervision subscales were only asked once, about the entire leadership team at the school, instead of multiple times for different staff types. The author found significant bivariate correlations in all cases, ranging from 0.20 (Supervision) to 0.38 (Coherent Leadership) for correlations with the measure of teacher self-efficacy and ranging from 0.49 (Supervision) to 0.75 (Support) for correlations with the overall measure of school culture defined by the School Culture Survey developers (Gruenert and Valentine 2006) as the "shared values/beliefs, the patterns of behavior, and the relationships in the school."

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): The DLI was developed for use in K–12 schools, and most items use general language, so revisions needed to item wording for use in ECE settings are likely minimal. The types of staff assessed would need to be updated. The developers only asked about staff types within the school for the Support and Supervision subscales, and the Coherent Leadership subscale asks about the leadership team at the school, so the equivalent staff types for ECE would be those within the center or other site.

Previous Version: None.

References:

DeMarco, A.L. "The Relationship Between Distributive Leadership, School Culture, and Teacher Self-Efficacy at the Middle School Level." Doctoral dissertation. South Orange, NJ: Seton Hall University, 2018.

Hulpia, H., and G. Devos. "Exploring the Link Between Distributed Leadership and Job Satisfaction of School Leaders." *Educational Studies*, vol. 35, no. 2, 2009, pp. 153–171.

Hulpia, H., G. Devos, and Y. Rosseel. "Development and Validation of Scores on the Distributed Leadership Inventory." *Educational and Psychological Measurement*, vol. 69, no. 6, 2009, pp. 1013–1034.

[15] The Teacher Self-Efficacy Scale is a 10-item scale measuring four domains, for job accomplishment, job skill development, social interaction with students, parents, and colleagues, and coping with job stress. The School Culture Survey has 35 items in 6 subscales, for Collaborative Leadership, Teacher Collaboration, Professional Development, Unity of Purpose, Collegial Support, and Learning Partnership. In the United States study, the author compared the DLI subscales only to the overall School Culture Survey score and not to any of its subscale scores.

Early Childhood Work Environment Survey, Third Edition (ECWES), 2016

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Early care and education (ECE)

Content

What leaders do (Promote quality practices)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

Center practices (Operational procedures and policies, family engagement)

Center structures and staff supports (Training and professional development, collaborative planning time, accountability structures)

Administration characteristics

Respondent: Director, teachers (lead and assistant)/support staff

Level of measure: Site

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

***Technology*:** Required

Staff level of qualifications

Personnel for administration: Test publisher or computer-scored program required

Training for administration: None

Ease of administration and scoring: 5 (administered or scored by publisher)

Time/length: 15 minutes, 185 items

Administration interval: None described for full measure, annual for short form

Languages available: English

Technical information

Development sample

Locale: United States (multiple states) and Canada

Setting: ECE center

Sample:

2016 normative sample: 2,580 staff (96% female) within 187 center-based early childhood programs (mean program size of 93 children, 47% nonprofit, 21% received Head Start funding)

Measure performance samples:

1985: 739 staff (94% female) within 65 center-based early childhood programs (mean program size of 86 children, 86% nonprofit)

1987: 423 staff (96% female) within 45 center-based early childhood programs

1996: 5,251 staff (95% female) within 421 center-based early childhood programs (mean program size of 99 children, 87% nonprofit)

2010: 3,980 staff (96% female) within 363 center-based early childhood programs (mean program size of 101 children)

Year of development: 2016

Measure performance

Reliability: 3 (meets minimum acceptability rating—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available (see Narrative)

Availability

4-Permission required, with costs

Material, training, and scoring costs: The ECWES costs $15 per survey through New Horizons. Cost per survey includes administration and scoring costs.

Developer(s)/publisher contacts

Developer(s): Paula J. Bloom

Publisher: New Horizons
P.O. Box 863
Lake Forest, IL 60045
847-295-8131
www.newhorizonsbooks.net/assessment-tools-2/early-childhood-work-environment-survey/

Measure website: www.mccormickcenter.nl.edu/library/the-early-childhood-work-environment-survey-ecwes/

Narrative

Description: The Early Childhood Work Environment Survey, third edition, (ECWES) is a self-administered online survey administered by New Horizons. It is designed to measure staff perceptions of center policies and practices and staff work attitudes in early care and education settings. The measure assesses 10 dimensions of organizational climate: Collegiality, Professional Growth, Supervisor Support, Clarity, Reward System, Decision Making, Goal Consensus, Task Orientation, Physical Setting, and Innovativeness. The ECWES includes these dimensions as 10 subscales (with 10 items each) to characterize the organizational climate of center-based early care and education programs. The Collegiality, Task Orientation, and Innovativeness subscales include content on "Center culture, climate, and communication." The Clarity subscale includes questions about "Center practices." The Goal Consensus subscale has questions relating to both the "Center culture, climate, and communication" subscale and the "Center practices" subscale. The Professional Growth and Rewards System subscales have content on "Center structures and staff supports." Both the Supervisor Support and Decision Making subscales include content on "What leaders do," but the Decision Making subscale also has items relating to "Center culture, climate, and communication." The Physical Setting subscale has items regarding the center's physical environment, which fall outside the Compendium's content areas. In addition, the ECWES assesses teaching staff's decision-making influence (5 items for desired and perceived influence, respectively) and perceptions of the center as a place to work. The ECWES also captures work attitudes, educational goals and objectives, and demographic information. Please review Chapter 3 of Bloom (2016) for more information on these measures, as they are not the focus of this compendium.

The survey averages 15 minutes to complete and should be completed by all paid staff, including administrators, coordinators, teachers (defined as lead teachers, teachers, and assistant teachers), and support staff (such as administrative assistants or cooks) who work at least 10 hours a week at the center. The developer recommends distributing a memo to staff inviting them to complete the survey. The developer also created a short version of the ECWES, though it cannot be used to describe the individual dimensions of organizational functioning so it will not be described here.

Uses of Information: The developer states that the ECWES could be used by program administrators to understand and improve a center's work environment. The measure can also be used for research or evaluation purposes to study job satisfaction and organizational climate.

Methods of Scoring: Since 2015, the ECWES has been administered as an online survey through New Horizons. Once all respondents complete their surveys, New Horizons aggregates the results and creates a Work Environment Profile summarizing the results. Many of the organizational climate subscales contain one question stem and responses that are "check all that apply" (with those responses referred to as *items*). For most of the organizational climate subscales, with the exception of Professional Growth and Clarity, the scores are calculated by subtracting the number of negative items checked from the number of positive items that were checked, and adding 5 to that value. The scores for the Professional Growth and Clarity subscales can be calculated by adding the number of items that were checked, resulting in a range of scores from 0 to 10. The aggregate center score for each subscale is the mean of all respondents' scores. Researchers are not advised to create an overall global organizational climate index by summing the 10 subscale scores, because the overall index score cannot show the variation in the 10 subscales.

The educational goals and objectives items are rank ordered from most important (1) to least important (6). The number of respondents that provided each ranking for these items is reported. Each of the five Decision-Making Influence items for desired decision-making influence and perceived influence are rated

by the respondent as having very little influence (scored as 0), some influence (scored as 5), or considerable influence (scored as 10). The score for each item is added together to calculate an overall desired influence and overall perceived influence score ranging from 0 to 50. This subscale can also be reported as a discrepancy score, which is the difference between the perceived difference score and the desired influence score. Staff perceptions of the center as a place to work are reported as the overall frequency of each word selected by respondents from a list of 30 words.

Interpretability: In general, higher scores on the organizational climate subscales indicate more favorable staff perceptions. New Horizons aggregates results into a Work Environment Profile for the center to provide overall staff perceptions of the center's organizational practices. To accompany the Work Environment Profile, New Horizons also provides an interpretation of the results. The interpretation document includes national norms relative to the 2016 sample for each organizational climate subscale. Bloom (2016) provides additional details on how to interpret ECWES results.

Reliability:

(1) Internal consistency reliability: The Cronbach's alpha coefficients were reported for each subscale and the total scale with the earlier editions of the measure (1985 and 1987) but were not reported for the 2016 normative sample. The total scale alpha coefficient for organizational climate was 0.93 and 0.95 (1985 and 1987, respectively). The developer also reported alpha coefficients for each of the six subscales (1985 and 1987, respectively): Collegiality (0.80 and 0.79); Professional Growth (0.75 and 0.69); Supervisor Support (0.84 and 0.83); Clarity (0.73 and 0.78); Reward System (0.68 and 0.75); Decision Making (0.83 and 0.80); Goal Consensus (0.75 and 0.82); Task Orientation (0.74 and 0.81); Physical Setting (0.65 and 0.77); and Innovativeness (0.73 and 0.70). The alpha coefficient was 0.66 for the Decision-Making Influence subscale (1987).

(2) Test-retest reliability: Test-retest reliability testing occurred within a two-month interval for both the 1985 and 1987 samples (n = 80 and 120, respectively). Correlations between the two administrations ranged from 0.60 (Clarity) to 0.93 (Decision Making) in 1985 and 0.60 (Physical Setting) to 0.89 (Decision Making) in 1987 for the 10 subscales, though statistical significance was not provided by the developers.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The individual items for each subscale were initially collected through interviews with early childhood teachers and directors and from other previously validated organizational climate scales. As part of the development of the first edition of the ECWES (1985), an unspecified number of expert reviewers conducted a Q-sort of the initial 150 items included in the organizational climate subscales. Items with less than 80% of agreement were removed, which resulted in the ECWES having 100 items in the measure across the 10 subscales.

(2) Construct/concurrent validity:

> *Construct validity:* The developer mentioned that factor analysis was not done on the subscales. Rather, they were derived from theory and previous sociology and psychology research.

In the second edition (1987), the developer redesigned the measure and made minor changes to the scale items. After making these changes, they conducted correlations between the subscales to determine the extent to which each subscale measured unique but inter-related aspects of early care and education work environments. The correlations ranged from 0.20 (for Professional Growth and Physical Setting) to 0.63 (for Supervisor Support and Decision Making) suggesting convergent validity (as may be expected for certain constructs). Furthermore, the magnitude of the mean correlation of a subscale with the other nine subscales ranged from 0.33 to 0.53 (subscale details not provided).

Concurrent validity: To test concurrent validity, the ECWES subscales were correlated to the Moos Work Environment Scale (Moos Scale), the Hay Group Organizational Climate Survey (HGOCS), and the CFK Climate Audit (CFKCA). The correlations between the ECWES and the Moos Scale ranged from 0.20 (Task Orientation subscale for the ECWES and the Moos Scale) to 0.90 (ECWES Decision Making subscale and the Moos Scale Supervisor Support subscale) in absolute value. The correlations between the ECWES and the HGOCS ranged between 0.25 (Task Orientation ECWES subscale and Performance Orientation HGOCS subscale) and 0.76 (Innovativeness ECWES subscale and Organizational Vitality HGOCS subscale). The correlations between the ECWES and the CFKCA ranged from 0.39 (Professional Growth ECWES subscale and Continuous Academic and Social Growth CFKCA subscale) to 0.86 (Physical Setting ECWES subscale and Suitability of School Plant CFKCA subscale). The developer did not indicate in what year these analyses were conducted (n = 120).

The developer compared the subscales on the ECWES to those on the Early Childhood Job Satisfaction Survey (ECJSS) (n = 120). The developer found correlations ranging from 0.02 (Clarity ECWES subscale and Pay and Promotion ECJSS subscale) to 0.84 (Physical Setting ECWES subscale and Working Conditions ECJSS subscale).

The developer also compared the Professional Growth subscale of the ECWES against the Program Administration Scale subscales (Talen and Bloom 2011) as a test of convergent validity (n = 67). The correlations ranged from 0.05 (Child Assessment and Marketing and Public Relations PAS subscales) to 0.43 (Family Partnerships PAS subscale). Dennis and O'Connor (2013) found the ECWES and the Organizational Climate Description Questionnaire had a correlation of 0.63. Schneider (1995) studied the relationship between work environment and burnout using the ECWES and the Maslach Burnout Inventory. The author reported that nine of the ECWES subscales were negatively correlated with the depersonalization burnout subscale, and five ECWES subscales were negatively correlated with the emotional exhaustion burnout scale.

The developer conducted differential statistical techniques to determine if responses to the ECWES would vary by role. The analyses demonstrated that administrators were more likely to view a center's organizational climate more positively than teachers do (1985). The developer also found that responses to items within some subscales varied based on program size, with larger programs receiving lower ratings on items related to team spirit, cooperation, and group cohesiveness (1985). In addition, the developer found that NAEYC-accredited centers scored higher on all 10 organizational climate subscales than those not NAEYC accredited (1996).

(3) Predictive validity: Although studies have examined relationships between organizational climate and various outcomes, it is not clear if any studies measured outcomes later in time than they measured organizational outcomes.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Designed for use in ECE. Validated using a sample of 2,580 staff within 187 center-based early childhood programs.

Previous Version: The ECWES was initially published in 1985. It was revised three times, in 1987, 1996, and 2010, before its current publication in 2016. In 2016, the developer made minor revisions to the individual items, generated new norms, and converted the ECWES to an online measure from a paper format.

References:

Bloom, P.J. *Measuring Work Attitudes: Technical Manual for the Early Childhood Job Satisfaction Survey and Early Childhood Work Environment Survey, 3rd edition*. Lake Forest, IL: New Horizons, 2016.

New Horizons. "Early Childhood Work Environment Survey." 2019. Available at http://newhorizonsbooks.net/assessment-tools-2/early-childhood-work-environment-survey/. Accessed July 22, 2019.

Essential 0-5 Survey (Previously Early Education Essentials)[16], 2018

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Early care and education (ECE) (intended for publicly funded programs serving preschool children)

Content

Who leaders are (Participation in decision making)

What leaders do† (Promote quality practices, foster respect and learning, promote family/community partnerships)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff; family relationships)

Center practices (Operational policies and procedures, family engagement)

Center structures and staff supports (Training and professional development)

† Includes what teachers as leaders do

Administration characteristics

Respondent: Teachers/other staff, parents

Level of measure: Site

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Required

Staff level of qualifications

Personnel for administration: Self- or computer administered, computer scored

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 5 (administered or scored by publisher)

Time/length: 20–30 minutes for teacher survey, 10-15 minutes for parent survey

Administration interval: None described

Languages available: English (teacher and parent survey), Spanish (parent survey)

Technical information

Development sample

Locale: United States (Chicago, IL)

Setting: ECE center, school (pre-K only)

Sample: 81 sites (41 school-based and 40 community-based serving preschoolers; on average children at sites were 52% male; 51% Hispanic and 38% Black; 13% special education); 746 teachers; 2,464 parents (34% completed Spanish version)

Year of development: 2016

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

4-Permission required, with costs
Material, training, and scoring costs: No cost information available. Direct consultation on using measure results has associated costs.

Developer(s)/publisher contacts

Developer(s): Stacy Ehrlich, Debra Pacchiano, Amanda Stein, and Maureen Wagner
Publisher: Start Early (formerly known as the Ounce of Prevention)
312-922-3863
www.startearly.org
Measure website: https://startearly.org/resource/the-essential-survey/

[16] The measure name and publisher changed since the profile review. We have updated linkages as of the time of publication where possible.

Narrative

Description: The Essential 0-5 Survey measure features a pair of teacher and parent surveys that measure the organizational conditions that support ECE teachers and other staff as well as teacher, child, and family relationships. The Essential 0-5 Survey consists of six essentials: (1) Effective Instructional Leaders, (2) Collaborative Teachers, (3) Involved Families, (4) Supportive Environment, (5) Ambitious Instruction, and (6) Parent Voice. The first five are based on the teacher survey (that can be completed by classroom and other staff working with children and families), and the sixth is based on the parent survey. The measure is designed for use in ECE settings, specifically school-based and center-based settings that receive public funding (Head Start or state pre-K) and serve preschool-age children.[17] It is adapted from the 5Essentials measure that was developed for K–12 education settings and uses a similar framework of essentials. Each essential consists of three to five subscales, referred to as measures. Based on the version used in the validation study (Ehrlich et al. 2018), each subscale consists of 3 to 8 items. The current measure has 21 subscales for the five essentials based on the teacher survey, and 4 subscales for the Parent Voice essential that is based on the parent survey; the number of items is not available. The teacher survey takes an average of 20 to 30 minutes to complete; the parent survey is shorter, at an average of 10 to 15 minutes.

The Essential 0-5 Survey reflects several categories of the ExCELS theory of change. The subscales in the Effective Instructional Leaders essential primarily contain content on "What leaders do," and individual subscales cover "Who leaders are" and "Center practices." The Collaborative Teachers essential mainly reflects the "Center culture, climate, and communication" category, with one subscale also measuring "Center structures and staff supports." The Supportive Environment and Ambitious Instruction essentials do not reflect any leadership constructs in the ExCELS theory of change; instead, their content involves child and center outcomes and instruction in classrooms. Finally, the Involved Families and Parent Voice essentials cover family-related constructs under the "What leaders do," "Center culture, climate, and communication," and "Center practices" categories, along with family outcomes. Across the measure, some of the subscales involving the "What leaders do" category specifically address leadership behaviors that teachers and other staff exhibit. In addition, the subscale under Effective Instructional Leaders measuring "Who leaders are" describes teacher leadership.

Uses of Information: Per the measure website, ECE programs can use the Essential 0-5 Survey to strengthen their organizational processes and practices, which should help continuously improve the quality of their teaching and enhance child outcomes. Along with the surveys, the publisher provides reports analyzing survey results and tools to help programs use the results to make improvements. The developers also recommend that researchers use the essentials to broaden the definition of quality in ECE and to explore connections between organizational conditions and other areas, such as leadership; staff and family experiences; program, staff, and child characteristics; classroom practice and quality; and child and family outcomes. In contrast, the developers caution against using the Essential 0-5 Survey for monitoring and accountability, because that could interfere with their intended use as improvement tools.

Methods of Scoring: Individual items use response scales such as level of agreement—for example, from strongly disagree (1) to strongly agree (4)—or frequency of behavior—for example, from never (1) to daily (5). For the validation study, the developers produced model-predicted site-level subscale scores (which is each center or school's deviation from the overall mean subscale score across all sites) on each

[17] The developers are currently adapting the teacher and parent surveys for use in infant-toddler settings; as of July 2020, the surveys had been pilot tested in Early Head Start-Child Care Partnership programs in Colorado, Florida, and the District of Columbia. (Start Early 2020)

subscale using a three-level measurement model. The developers then standardized the model-based subscale scores and averaged across the subscales under each essential to create site-level essential scores on each of the six essentials. Per the measure website, the scoring process is conducted by the publisher and involves converting individual responses (from teachers and parents) into site-level scores on each essential and subscale.

The developers calculated the intraclass correlation (ICC) for each subscale in a version of the measure that had 24 subscales in the teacher survey and 9 subscales in the parent survey.[18] The ICCs assess the degree to which teachers' and parents' responses were more related to those from other teachers and parents at the same center or school than they were to responses from other centers or schools. Five teacher survey subscales and six parent survey subscales had an ICC under 0.05, below the typical range for adequate site reliability. Those five teacher survey subscales were focused on classroom-level constructs instead of broader organization-level constructs.[19] Thirteen teacher survey subscales and three parent survey subscales had values within the typical range demonstrating adequate site reliability, whereas the remaining six teacher survey subscales also were adequate, with ICCs above 0.20.

Interpretability: Higher scores indicate stronger use of the essential practices. Information on the scores themselves, such as the range used, or their interpretation, such as what can be characterized as an average score, is not available. The publisher provides reports analyzing survey results to help programs interpret the meaning of their scores as well as tools that programs can use to improve their organizational conditions. The publisher notes the availability for direct support on using the results through web meetings and phone calls.

Reliability:

(1) Internal consistency reliability: In the validation study, the developers examined Rasch person reliability scores for each of the 24 subscales in the version of the teacher survey being tested and each of the 9 subscales in the version of the parent survey being tested. All subscales—those created from the teacher and the parent survey responses—had reliability scores of at least 0.70, and all but two subscales had scores of at least 0.80.

The developers also examined site reliability scores of the consistency across responses of teachers and parents from the same center or school. For the teacher survey, site reliability scores were lower for the subscales measuring classroom-level constructs, which ranged from 0.12 to 0.28, than for the remaining subscales, which ranged from 0.35 to 0.83 (same range for the 21 subscales that remain in the current version of the measure). For the parent survey, site reliability scores ranged from 0.33 to 0.73 (from 0.33 to 0.49 for the 4 subscales that remain in the current measure).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

[18] Later in the development process, the developers dropped some subscales, leaving the 21 subscales based on the teacher survey and 4 subscales based on the parent survey in the current measure.

[19] Of the five subscales, Positive Learning Climate is a subscale in the Supportive Environment essential. The other four, Quality of Student Interaction, Early Language Development, Early Cognitive Development, and Early Social-Emotional Development, constitute four of the five subscales in the Ambitious Instruction essential in the current measure.

Validity:

(1) Content validity: The developers pursued content validity by using or adapting most teacher survey questions from the 5Essentials teacher survey, because the Essential 0-5 Survey is based on that measure. Because the parent survey was new, the developers conducted one-on-one cognitive interviews with parents to determine if the questions and response categories were understandable and appropriate.

The developers also conducted a pilot study in spring 2015, before the validation study, using samples of preschool teachers in the Chicago public school system and a nationally representative set of Head Start center teachers for the teacher survey, and parents from 16 school- and center-based preschool sites for the parent survey. The developers analyzed the pilot study results for reliability, construct validity, and internal validity, including Rasch reliability coefficients, item difficulty, and item fit statistics (Ehrlich et al. 2016). Based on the analyses, the developers revised the Essential 0-5 Survey before proceeding with the validation study.

(2) Construct/concurrent validity:

Construct validity: The developers used Rasch analysis to examine item fit and spread of difficulty and removed some items that had a large misfit or did not differentiate respondents beyond other items. All items retained for the full analysis had acceptable infit mean squares (between 0.7 and 1.3). However, some parent survey items did not provide enough differentiation because they had very high rates of favorable responses.

During the full analysis, the developers used exploratory factor analysis to examine the grouping of subscales into higher-level essentials. In the version of the measure tested, the parent survey subscales had been initially placed within the five essentials. The analysis demonstrated that the most appropriate number of factors was five, although the factors did not fully align with the five essentials. The developers reported all factor loadings greater than 0.50 (including negative values below −0.50). They found that most Effective Instructional Leaders and Collaborative Teachers subscales loaded onto one factor (with loadings ranging from 0.55 to 0.89); most Supportive Environment and Ambitious Instruction subscales loaded onto a second factor (with loadings ranging from 0.53 to 0.80); almost all the parent survey subscales across the five essentials loaded onto a third factor (with loadings ranging from 0.63 to 0.85); and the fourth and fifth factors did not map strongly to any essential. Seven of the 33 subscales cross-loaded onto more than one factor. For the most part, the developers preserved the existing placement of subscales within essentials because the constituent subscales for a given essential tended to load onto a factor together (even though more than one essential loaded onto a single factor) and so they would align with the K–12 version (5Essentials). However, because the subscales based on the parent survey all loaded strongly into one factor, the developers moved all those subscales to create the new sixth essential, Parent Voice.

Concurrent validity: The validation study examined associations between the Essential 0-5 Survey scores and the CLASS Pre-K (which measures classroom quality) and student attendance (as a child outcome of interest) (n = 120 classrooms in school-based sites and 150 classrooms in community-based sites). The measures were all collected during the 2015–16 school year, although the precise timing varied by site (the surveys for the essentials were all collected in spring 2016; the CLASS Pre-K scores were primarily from the 2015–16 school year, with a small percentage coming from the previous school year; the student attendance was for the 2015–16 school year). The Effective

Instructional Leaders and Collaborative Teachers essentials were significantly positively related to CLASS scores; the Ambitious Instruction essential was significantly negatively related to CLASS; and the other essentials were not significantly related. Four essentials—Effective Instructional Leaders, Collaborative Teachers, Supportive Environment, and Involved Families—were significantly, positively related to student attendance. The strength of these associations was lower when adjusted for student background characteristics, but most associations remained significant, providing evidence of concurrent validity.

The developers also conducted what they refer to as a "qualitative validation study" by selecting four sites with especially strong or weak scores on the Essential 0-5 Survey. They then visited these sites to interview leaders, teachers, and parents, and to informally observe common areas of the sites, such as drop-off and pick-up areas, hallways, and outdoor activity areas.[20] Site visitors did not know which of the sites they visited had high or low Essential 0-5 Survey scores. The developers concluded from the site visits that the Essential 0-5 Survey measures differentiated between programs with varying levels of supports for their staff and the families and children they serve[21].

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: The validation study examined differential item functioning (DIF) between school-based versus community-based settings and between the English and Spanish versions of the parent survey. The developers considered removing items with large, significant DIFs, although items were retained if the fit was deemed acceptable within each group. For the subscales retained for the final analysis, 25 percent of the teacher survey subscales had more than one item with large, significant DIFs between setting type. This suggests that caution is needed in making any comparisons between setting types. The same score may have different meaning across setting types. One subscale, School Commitment (part of the Collaborative Teachers essential) had all four items exhibit large, significant DIFs; the developers noted they would monitor this subscale in future use. None of the parent survey subscales had items with large, significant DIFs between setting types, and only one subscale (Including Parents as Partners, which remains in the current measure) had more than one item with large, significant DIFs between survey languages. As part of the initial development process before the pilot study, the developers conducted cognitive interviews with parents to ensure different language groups interpreted the questions similarly.

Training Support: The publisher does not list any required training. Several survey administration supports are available, including a survey administration manual and presentation, and recruitment materials (flyers and email templates) to use with respondents.

Key Considerations for Early Care and Education (ECE): Although the original 5Essentials measure was developed for K–12 education, a primary aspect of developing the Essential 0-5 Survey measure was adapting items from the 5Essentials for ECE settings. To date, the Essential 0-5 Survey has only been validated in publicly funded school- and center-based settings that serve preschool-age children. The developers have adapted the measure for settings serving children from birth to age 2, and are currently investigating the validity of this adapted version.

[20] The developers did not conduct formal observations of classrooms.

[21] For more information about the qualitative study findings, see https://startearly.org/app/uploads/pdf/Early-Ed-Essentials-Snapshot-Mar2018-Ounce-Consortium.pdf.

Previous Version: Although the Essential 0-5 Survey is drawn heavily from the 5Essentials measure used in K–12 education, it is a different measure. It has been revised throughout the development and validation process.

References:

Ehrlich, S.B., D.M. Pacchiano, A.G. Stein, and S. Luppescu. "Essential Organizational Supports for Early Education: The Development of a New Survey Tool to Measure Organizational Conditions." Chicago, IL: University of Chicago Consortium on School Research and the Ounce of Prevention Fund, 2016.

Ehrlich, S.B., D.M. Pacchiano, A.G. Stein, M.R. Wagner, S. Luppescu, S. Park, E. Frank, H. Lewandowski, and C. Young. "Organizing Early Education for Improvement: Testing a New Survey Tool." Chicago, IL: University of Chicago Consortium on School Research and the Ounce of Prevention Fund, 2018.

Ehrlich, S.B., D. Pacchiano, A.G. Stein, M.R. Wagner, S. Park, E. Frank, S. Luppescu and C. Young. "Early Education Essentials: Validation of Surveys Measuring Early Education Organizational Conditions." *Early Education and Development*, vol. 30, no. 4, 2019, pp. 540–567. doi:10.1080/10409289.2018.1556969.

Ounce of Prevention Fund. "Early Education Essentials." Chicago, IL: Ounce of Prevention Fund, July 2020. Available at https://www.startearly.org/resources-professionals/professional-development/essential-survey/. Accessed July 15, 2020.

Implementation Leadership Scale (ILS), 2014

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Health

Content

What leaders do (Promote quality practices, foster respect and learning)

What leaders bring (Knowledge of practices; values, beliefs, attributes)

Administration characteristics

Respondent: Manager, employees

Level of measure: Site/team, individual

Data sources: Survey (mode–self-administered, report level–self-report or report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 5 minutes, 12 items

Administration interval: None described

Languages available: English

Technical information[22]

Development sample

Locale: United States (Southern California)

Setting: Mental health programs

Sample: 459 mental health clinicians; Mean age 36.5; 79% female; 54% White, 23.4% Hispanic, 6.7% Black, 5% Asian, 0.5% American Indian, and 10% Other.

Year of development: 2014

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

Availability

1-Public domain *(Distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited)*

Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Gregory A. Aarons, Mark G. Ehrhart, and Lauren R. Farahnak

[22] Information available for staff version only.

Narrative

Description: The Implementation Leadership Scale (ILS) is a self-administered survey to measure leadership behaviors that support successful evidence-based practice (EBP) implementation in mental health service settings. The survey can be administered either online or on paper and takes about five minutes to complete. There are two versions of the ILS: one for staff report of supervisors and another one for supervisor report of oneself. Each version of the ILS contains four subscales: Proactive Leadership, Knowledgeable Leadership, Supportive Leadership, and Perseverant Leadership. Each subscale contains 3 items for a total of 12 items in the measure. The Proactive, Supportive, and Perseverant subscales measure "What leaders do." The Knowledgeable and Perseverant subscales capture information on "What leaders bring." The developers examined the psychometric properties of the staff version of the ILS.

Uses of Information: The developers designed the ILS to identify leadership behaviors that may help create a supportive EBP implementation climate in the teams and facilitate EBP implementation and sustainability. It is also used in the evaluation of an intervention to improve EBP implementation leadership.

Methods of Scoring: Each item in the ILS is scored on a 5-point scale for the extent one agrees with a given statement: not at all (0), slight extent (1), moderate extent (2), great extent (3), or very great extent (4). Subscale scores are the mean of the items within each of the subscales. The total ILS score is the mean of the subscale scores.

The developers validated the measure within two samples, representing two separate sectors, with similar results. They note that the scales can be considered for site-level reporting (across staff). They examined the average agreement within group for each item, $a_{wg(1)}$, and the subscales, $a_{wg(j)}$. Estimates of a_{wg} greater than 0.60 indicate acceptable agreement. In the study with mental health service providers, the a_{wg} estimates ranged from 0.67 to 0.73 for individual items and from 0.68 (Proactive) to 0.72 (Knowledgeable) for subscales. In the study with alcohol and drug use (AOD) treatment service providers, the a_{wg} estimates ranged from 0.73 to 0.78 for individual items and from 0.74 (Supportive) to 0.76 (Proactive, Knowledgeable, and Perseverant) for subscales.

Interpretability: Higher scores on the ILS indicate stronger perceptions of implementation leadership behaviors. Higher subscale scores indicate stronger perceptions toward proactive, knowledgeable, supportive, or perseverant leadership behaviors based on what subscale is being completed.

Reliability:

(1) Internal consistency reliability: Cronbach's alpha was 0.98 for the ILS total score, and the alphas ranged from 0.95 (Proactive and Supportive) to 0.96 (Knowledgeable and Perseverant) for the subscales in the study with mental health service providers (Aarons et al. 2014). In another validation study of the ILS in a sample of service providers in AOD treatment agencies, Cronbach's alpha was 0.97 for the total score, and the alphas ranged from 0.93 (Supportive) to 0.97 (Knowledgeable) for the subscales (Aarons et al. 2016).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: Three phases occurred for the development of items in the ILS. First, the developers reviewed literature on leader behaviors on organizational climate, implementation, and culture change. Second, the developers conducted expert review of items. Experts in the field included a mental health program leader, an EBP trainer, a community development team consultant, and four mental health program managers. Third, expert reviewers assessed potential items for face validity.

(2) Construct/concurrent validity:

Construct validity: The developers conducted exploratory factor analysis (EFA) to identify the factor structure of the ILS and confirmatory factor analysis (CFA) to assess the model fit of the factor structure across two samples. The EFA factor loadings were strong for each scale – all greater than 0.43 (n = 229). Inter-factor correlations were significant and ranged from 0.73 (Proactive and Knowledgeable) to 0.80 (Perseverant and Supportive) (Aarons et al. 2014) (n = 229). The CFA model with a second-order factor for overall implementation leadership demonstrated acceptable fit in both samples, with first-order factors corresponding to the four subscales identified in the EFA (Aarons et al. 2014, 2016). First-order factor loadings (of items to a factor represented by a subscale) ranged from 0.90 to 0.97 in the mental health providers sample (n = 229) and from 0.85 to 0.97 with AOD treatment service providers. Second-order factor loadings (of the subscales to overall implementation leadership) ranged from 0.90 to 0.94 with mental health providers (n = 230) and 0.87 to 0.92 with AOD treatment service providers.

Concurrent validity: The correlations between the ILS subscales and total score and the Multifactor Leadership Questionnaire (MLQ; Bass and Avolio 1995) subscales on transformational and transactional leadership were significant at $p \leq .01$ and ranged from 0.62 (Proactive ILS subscale and Individualized Consideration MLQ subscale) to 0.75 (ILS total score and Idealized influence MLQ subscale), indicating evidence of convergent validity (n = 459). The correlations between the ILS subscales and total score and the Organizational Climate Measure (OCM; Patterson et al. 2005) subscales on general organizational climate (autonomy, formalization, efficiency, and performance feedback) ranged from 0.05 (Proactive ILS subscale and Autonomy OCM subscale) to 0.41 (Supportive ILS subscale and Feedback OCM subscale), indicating evidence of divergent validity with very low correlations between subscales measuring very different constructs. All ILS and OCM correlations were significant ($p \leq 0.05$ or higher), with the exception of 0.05 for the Proactive ILS subscale and Autonomy OCM subscale and 0.08 for the Perseverant ILS subscale and Autonomy OCM subscale (n = 459).

The study with AOD treatment service providers demonstrates similar evidence of convergent and divergent validity (n varied from 316 to 323). Correlations between total score and the MLQ subscales were all significant and ranged from 0.57 (Knowledgeable ILS subscale and the Transactional Leadership Contingent Reward MLQ subscale) to 0.77 (ILS total score with multiple MLQ subscales including Intellectual Stimulation, Individual Consideration, and Idealized Influence) providing evidence for convergent validity. Correlations between the ILS subscales and the OCM subscales were significant and ranged from 0.19 (Knowledgeable ILS subscale and Autonomy OCM subscale) to 0.57 (ILS total score and Support subscale with the Feedback subscale of the OCM), showing divergent validity.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: The ILS is within public domain with instructions on how to score it.

Key Considerations for Early Care and Education (ECE): The ILS was validated across two sectors, with similar results indicating the measure can be generalized to other populations. The terminology for the measure focuses on "implementing EBP" within mental health clinical settings and some items would need to be revised to ECE practices.

Previous Version: None.

References:

Aarons, G.A., M.G. Ehrhart, and L.R. Farahnak. "The Implementation Leadership Scale (ILS): Development of a Brief Measure of Unit Level Implementation Leadership." *Implementation Science*, vol. 9, article no. 45, 2014. doi:10.1186/1748-5908-9-45.

Aarons, G.A., M.G. Ehrhart, E.M. Torres, N.K. Finn, and S.C. Roesch. "Validation of the Implementation Leadership Scale (ILS) in Substance Use Disorder Treatment Organizations." *Journal of Substance Abuse Treatment*, vol. 68, 2016, pp. 31–35.

Leadership Practices Inventory (LPI), 2016

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Multiple (K–12 education, management, health)

Content

What leaders do (Promote quality practices, foster respect and learning, establish vision, manage efficient operations)

What leaders bring (Personal development or critical-thinking knowledge and skills, interpersonal and team-building knowledge and skills; values, beliefs, attributes)

Administration characteristics

Respondent: Manager/director (LPI-Self), employees (LPI-Observer), other (manager/director's supervisor) (LPI-Observer)

Level of measure: Site, individual

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills with some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills

Time/length: 8 to 10 minutes, 30 items

Administration interval: None described

Languages available: English, Spanish (Latin American), and other (simplified Chinese, Arabic, and Brazilian Portuguese)

Technical information

Development sample

Locale: United States and non-U.S. (Australia, China, Hong Kong, Jordan, Lebanon, Mexico, Philippines, Saudi Arabia, Thailand, Uganda)

Setting: Multiple (for example, schools, hospitals, business offices)

Sample: 2.8 million respondents (74% from U.S.) to the online LPI, including 17% leaders

LPI-Self: 56.4% male; 58.4% 33 to 49 years old; 42.4% completed college; 72.3% White (only asked of U.S. respondents)

LPI-Observer (leader's supervisor, direct reports to the leader, co-workers, or other): 54.2% male; 51.7% 33 to 49 years old; 44.7% completed college; 72.8% White (only asked of U.S. respondents)

Year of development: 2007–2015

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available (see Narrative)

Availability

4-Permission required, with costs

Material, training, and scoring costs: Three different packages are available. All include online administration of the survey, scoring with personalized reports, and participant materials. Prices vary depending on the number of licenses purchased.

- LPI-Self: $64 to $80 for online survey
- LPI 360: includes the LPI-Self and LPI-Observer assessments; $160 to $200 per leader
- LPI 360+: offers the LPI-Self and LPI-Observer assessments, and reassessment within 18 months; $176 to $220 per leader

The developers offer a Leadership Challenge facilitator training ranging from $1,695 to $4,995. Participants learn how to facilitate a Leadership Challenge workshop, which includes the LPI.

Developer(s)/publisher contacts

Developer(s): Barry Z. Posner and James M. Kouzes

Publisher: The Leadership Challenge, a Wiley brand
(866) 888-5159
Leadership@wiley.com

Measure website: www.leadershipchallenge.com/professionals-section-lpi.aspx

Narrative[23]

Description: The Leadership Practices Inventory (LPI) was developed to measure the Five Practices of Exemplary Leadership framework, a transformational leadership model also developed by Kouzes and Posner (2017c). The LPI comprises two forms: one for a leader's self-report (LPI-Self) and the second for others' report of a leader (LPI-Observer). It is part of the Leadership Challenge Workshop, a workshop designed to help attendees develop the needed skills to meet leadership challenges, but it can also be administered independently. Participants are asked to complete the LPI-Self, and they select 5 to 10 people who are familiar with their leadership behavior to complete the LPI-Observer, such as the leader's supervisor, employees that report directly to the leader, or other co-workers. The LPI is administered as a paper or web survey by an independent facilitator. The framework identifies five subscales (referred to as leadership practices): Model the Way, Inspire a Shared Vision, Challenge the Process, Enable Others to Act, and Encourage the Heart. The Model the Way, Challenge the Process, and Enable Others to Act subscales contain content on "What leaders do" and "What leaders bring." The Inspire a Shared Vision and Encourage the Heart subscales only measure "What leaders do." Each subscale has six items, with a total of 30 items across subscales. Each form takes between 8 to 10 minutes to complete.

Uses of Information: The developers note that the original purpose of the LPI was for development or improvement by helping the respondents of the LPI-Self become more effective leaders. The LPI can also be used for research or evaluation. The developers cite studies looking at the association of the LPI with organizational and managerial effectiveness and other outcomes such as group performance, team cohesion, and job satisfaction.

Methods of Scoring: The measure uses a 10-point frequency scale: almost never do (1), rarely (2), seldom (3), once in a while (4), occasionally (5), sometimes (6), fairly often (7), usually (8), very frequently (9), and almost always (10). A subscale score can be calculated by adding the responses to each of the six items within a subscale; subscores range from 6 to 60. The subscale scores are reported separately for the leader based on his or her responses to the subscale items in the self-report, and then for each individual who completed the LPI-Observer, known as Observers. The leader receives an average subscale score across all Observers, and average scores by Observer type such as the leader's supervisor, direct reports, co-workers, and other. Both the LPI-Self and LPI-Observer forms are scored by The Leadership Challenge. They can be scored by hand or by a computer program. If the web version of the LPI is used, the forms are automatically scored, and reports are created for the leader. See Kouzes and Posner 2017a and 2017b for sample reports with additional scoring details.

Interpretability: In general, low scores on the items indicate less frequent use of a leadership practice, and higher scores indicate more frequent engagement of the practice. Consistent scores across all respondents for a particular item indicate a level of agreement on how frequently the leader engages in the leadership behavior. The developers suggest comparing the leader's LPI-Self score to the average Observer score and the leader's supervisor score for each individual item. If the difference in scores is greater than 1.5 points, the behavior "merits attention" (Posner 2017b, p. 4). Leaders receive a Self-Report and Individual Feedback Report to help them interpret their results. The reports include the raw scores from the LPI-Self and LPI-Observer forms and averages across all Observers and by Observer type. They both provide a ranking of leadership behaviors, from most to least engagement. Both reports also compare the leader's scores within each subscale to Observer scores for other leaders in the LPI database. See Kouzes and Posner 2017a and 2017b for sample reports and additional details on how to interpret results.

[23] The information summarized in this profile was taken from Posner 2016 unless otherwise noted.

Reliability:

(1) Internal consistency reliability: For the LPI normative data, Cronbach's alphas ranged from 0.81 (Model the Way) to 0.90 (Inspire a Shared Vision) for the leadership practices subscales in the LPI-Self, and 0.86 (Model the Way) to 0.92 (Encourage the Heart) for the leadership practices subscales in the LPI-Observer. The developers also cite other studies that showed strong internal reliability across diverse samples both within and outside the United States, including a study among K–12 education teachers, with Cronbach's alphas ranging from 0.78 to 0.95.

(2) Test-retest reliability: The developers note that in an earlier version of the measure subscales for the LPI showed test-retest reliability scores at the 0.90 level and above. The developers highlighted a study among school administrators that reported test-retest reliabilities of 0.86 for superintendents and 0.79 for school principals (Kouzes and Posner 2002). The developers did not specify the time period between the two administrations of the LPI.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers mention that past qualitative and quantitative research informed the Five Practices of Exemplary Leadership framework and the LPI specifically. Interviews and iterative feedback sessions with respondents and expert reviewers also informed the individual items of the LPI. Psychometric tests further refined the measure.

(2) Construct/concurrent validity:

Construct validity: The developers conducted exploratory factor analysis using principal factoring with iteration and varimax rotation on an earlier version of the measure and found the LPI generally contains five factors consistent with the LPI subscales. The factor loadings ranged from 0.46 to 0.72 for Enable Others to Act; 0.53 to 0.73 for Encourage the Heart; 0.48 to 0.71 for Inspire a Shared Vision; 0.39 to 0.64 for Challenge the Process; and 0.37 to 0.61 for Model the Way (Posner and Kouzes 1990). The developers also cite other studies that used the LPI and confirmed the five-factor structure through confirmatory factor analysis.

Concurrent validity: The developers cite studies that show significant associations of LPI Observer scores and organizational and managerial effectiveness, and other outcomes, such as group performance, team cohesion, and job satisfaction (evidence of concurrent validity).

(3) Predictive validity: Although studies have examined relationships between leadership practices and various outcomes, it is not clear if any studies measured outcomes later in time than they measured leadership practices.

Bias Analysis: The developers cite a study (Zagorsek et al. 2006) that conducted multigroup confirmatory factor analysis on the LPI that demonstrated a consistent five-factor structure across different cultural settings.

Training Support: The developers developed an LPI facilitator's guide, which includes a paper version of the LPI Self and Observer forms, instructions on how to administer the LPI, and scoring software. The developers also hold Leadership Challenge facilitator training sessions, where participants learn how to facilitate a Leadership Challenge workshop, which includes the LPI.

Key Considerations for Early Care and Education (ECE): The developers describe the use of the LPI across various industries, including K–12 school settings and higher education, but not ECE. Considering its use in these settings however, the LPI could be used within ECE without needing adaptations.

Previous Version: It is not clear how many versions of the LPI there have been since it was developed in 1988. The developers mention that the original version of the LPI used a 5-point scale, and the developers transitioned to a 10-point frequency scale in 1999. The developers note that they developed a specific LPI form for high school and college students.

References:

John Wiley & Sons, Inc. "The Leadership Challenge. LPI: Leadership Practices Inventory." 2019. Available at https://www.leadershipchallenge.com/Resources.aspx. Accessed September 20, 2019.

Kouzes, J.M., and B.Z. Posner. "LPI: Leadership Practices Inventory Self Report, Sample Self Report (V5)." The Leadership Challenge, A Wiley Brand, 2017a. Accessed at www.leadershipchallenge.com. Accessed September 20, 2019; V5 no longer available.

Kouzes, J.M., and B.Z. Posner. "LPI: Leadership Practices Inventory Individual Feedback Report, Sample Assessment (V5)." The Leadership Challenge, A Wiley Brand, 2017b. Accessed at www.leadershipchallenge.com. Accessed September 20, 2019; V5 no longer available.

Kouzes, J.M., and B.Z. Posner. *The Leadership Challenge.* 6th Edition. Hoboken, NJ: John Wiley & Sons, Inc., 2017c.

"The Leadership Practices Inventory: Theory and Evidence behind the Five Practices of Exemplary Leaders." May 12, 2002. Accessed at: www.leadershipchallenge.com. Accessed September 20, 2019; no longer available.

Posner, B.Z., and J.M. Kouzes. "Development and Validation of the Leadership Practices Inventory." *Educational and Psychological Measurement,* vol. 48, 1988, pp. 483–496.

Posner, B.Z., and J.M. Kouzes. "Leadership Practices: An Alternative to the Psychological Perspective." In *Measures of Leadership*, edited by K.E. Clark and M.B. Clark. West Orange, NJ: Leadership Library of America, 1990.

Posner, B.Z., "Investigating the Reliability and Validity of the Leadership Practices Inventory." *Administrative Sciences*, vol. 6, no. 17, 2016. doi:10.3390/admsci6040017.

Multifactor Leadership Questionnaire, Third Edition (MLQ [5X-Short]), 2011

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Management

Content

What leaders do (Promote quality practices, foster respect and learning, establish vision)

What leaders bring (Interpersonal and team-building knowledge and skills; values, beliefs, attributes)

Administration characteristics

Respondent: Manager, employees

Level of measure: Site, individual

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 15 minutes; 45 items

Administration interval: 3 to 12 months

Languages available: English, Spanish, other (see measure website for details)

Technical information

Development sample

Locale: United States

Setting: Business office

Sample: 27,285 participants from organizations using publisher's database; contains self-rating leaders (14%) and employees at higher (16%), lower (44%), same (19%), and other (7%) levels compared to leaders they rated

Year of development: 2004

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Available

Availability

4-Permission required, with costs

Material, training, and scoring costs:

- $2.50 per person, min. 50 (paper or nonpublisher survey system)
- $2.50 or $8 per person, min. 20 (publisher online system, depending on form)
- $15 to $200 reports per person or group (publisher online system, depending on type of report)

Developer(s)/publisher contacts

Developer(s): Bruce J. Avolio and Bernard M. Bass

Publisher: Mind Garden, Inc.
650-322-6300
www.mindgarden.com

Measure website: https://www.mindgarden.com/16-multifactor-leadership-questionnaire

Narrative

Description: The Multifactor Leadership Questionnaire (MLQ—also known as the MLQ 5X-Short or the standard MLQ) is a self-administered survey that measures the leader's behaviors and attributes that are components of various leadership styles. It is a 45-item survey that takes 15 minutes to administer. There are two versions with the same items: a leader's self-report or employee ratings of the leader. The MLQ includes nine subscales measuring leadership styles and each contain 4 items:

- Transformational leadership is measured by five subscales: Idealized Attributes, Idealized Behaviors, Inspirational Motivation, Intellectual Stimulation, and Individual Consideration.
- Transactional leadership is measured by two subscales: Contingent Reward and Management by Exception–Active.
- Passive/avoidant leadership is measured by two subscales: Laissez-Faire and Management by Exception–Passive

The MLQ also features three remaining subscales that measure leadership outcomes: Extra Effort (3 items), Effectiveness (4 items), and Satisfaction (2 items).

The MLQ is not specific to a particular field; it can be used in a wide range of settings. Another version of the MLQ, the 5X-Long, adds two items to each leadership subscale for a total of 63 items. However, the 5X-Long is no longer in print; except when explicitly mentioned, this profile describes the MLQ 5X-Short and refers to it as the MLQ. The MLQ can be administered and scored online or on paper using the publisher's assessment system.

The MLQ's items focus on leadership behaviors, reflecting the "What leaders do" component of the ExCELS theory of change. The items in the Idealized Attributes subscale also relate to "What leaders bring," and the items in the Extra Effort, Effectiveness, and Satisfaction subscales involve outcomes such as staff motivation and staff satisfaction with leadership; however, some of the items in these subscales are connected to "What leaders do."

Uses of Information: The developers state that the MLQ can be used for both leadership development and leadership research (Avolio and Bass 2011). The developers describe specific applications for both purposes. For development, organizations can use the MLQ to transfer or promote staff into leadership positions, including those that fit best with their leadership style. Organizations can also use it to identify leaders to receive training or other professional development opportunities, or more directly, to train or coach leaders or groups of leaders specifically using their MLQ scores as a guide to improving their leadership style and behaviors. The publisher offers tools to be used with the MLQ for development and improvement. The developers suggest a retest in three months to a year for assessing change for development and improvement purposes.

For research, the MLQ can be used to study how leadership styles and effectiveness are shaped by leaders' backgrounds and experiences, or how leadership might affect organizational outcomes. It can also be used as an outcome measure to evaluate the effectiveness of organizational initiatives to improve its leadership quality.

Methods of Scoring: Each item is scored on a 5-point frequency scale for behavior happening not at all (0); once in a while (1); sometimes (2); fairly often (3); and frequently, if not always (4). Scores for each of the 12 subscales are calculated by averaging the score for each item, producing raw scores ranging from 0 to 4. The developers list the raw score for every 5th percentile for the norming sample, which contains responses accumulated over time from leaders and employees whose organizations had administered the MLQ through the publisher's online system. Scores can be calculated for an individual or for a group, based on either leaders' self-reports or on employees' ratings of leaders.

Interpretability: Higher raw scores correspond to higher levels of the leadership style reflected in the subscale. However, the developers say scores should be interpreted not in absolute terms (as in "this leader is transformational") but relative to normal (as in "this leader is more transformational than the norm"). The percentile ranks for raw scores show the leader's position relative to the norming sample. The publisher's online assessment system includes reports providing MLQ results for individuals or groups. These reports include interpretations of scores, including comparisons between self-ratings and ratings from others, and summaries of scores for groups of participants.

Reliability:

(1) Internal consistency reliability: The most recent reliability and validation study occurred in 2004, using data from the norming sample that the publisher accumulated over time. With this study's overall sample, reliability scores[24] for the 12 subscales ranged from 0.69 (Contingent Reward) to 0.83 (two subscales: Inspirational Motivation and Extra Effort); only one subscale, Contingent Reward, had a score below 0.70. The study also examined reliability scores for five subgroups: self-raters; raters at higher, lower, and the same level as the leader they rated; and raters whose level relative to leaders could not be compared. For self-ratings, reliability scores ranged from 0.60 (two subscales: Contingent Reward and Laissez-Faire) to 0.79 (Extra Effort), and 7 of 12 were lower than 0.70: Idealized Behaviors, Intellectual Stimulation, Individual Consideration, Contingent Reward, Management by Exception–Passive, Laissez-Faire, and Effectiveness. Reliability scores for raters at higher levels ranged from 0.48 (Idealized Behaviors) to 0.83 (three subscales: Inspirational Motivation, Extra Effort, and Effectiveness), with 3 of 12 ratings less than 0.70: those for the Idealized Behaviors, Contingent Reward, and Management by Exception–Passive subscales. Finally, for the reliability scores for raters at the same and lower levels as the leader and those at noncomparable levels, the 36 scores (12 subscales for these three subgroups) ranged from 0.68 (Contingent Reward) to 0.85 (Extra Effort) and only one was below 0.70: Contingent Reward among raters at the same level.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: To develop the current version of the MLQ, the developers used results from factor and other analyses of the previous version to select items, developed new items based on recent studies of leadership, and received recommendations from expert reviewers about items to modify or drop.

[24] The MLQ manual did not identify the statistic used for the reliability score.

(2) Construct/concurrent validity:

Construct validity: The most recent validation study used confirmatory factor analysis to examine models ranging from a one-factor model (all items) to a full nine-factor model (one per leadership subscale). The nine-factor model demonstrated an acceptable fit, which was also the strongest fit among the models tested. With the full sample, the 20 item factor loadings for the five transformational leadership subscales ranged from 0.44 to 0.81: the 8 item factor loadings for the two transactional leadership subscales ranged from 0.51 to 0.70, and the 8 item factor loadings for the two passive/avoidant leadership subscales ranged from 0.34 to 0.80.

The study also examined inter-factor correlations for the 12 subscales. With the full sample, examining correlations among subscales within a leadership style (n = 27,285):

- The five transformational leadership (Idealized Attributes, Idealized Behaviors, Inspirational Motivation, Intellectual Stimulation, and Individual Consideration) inter-factor correlations ranged from 0.59 (Intellectual Stimulation with Idealized Behaviors and Inspirational Motivation; and Individual Consideration with Inspirational Motivation) to 0.71 (Individual Consideration with Idealized Attributes) (all statistically significant), suggesting they are all measuring the same leadership style.
- The two passive/avoidant leadership subscales (Laissez-Faire and Management by Exception–Passive) were correlated (r = 0.61, statistically significant), suggesting they are measuring the same leadership style.
- The two transactional leadership subscales (Contingent Reward and Management by Exception–Active) were not correlated (r = 0.01, not statistically significant) suggesting that they are measuring different leadership styles.

Second, to look at relations among different leadership styles, correlations involve subscales under different leadership styles. All but one of these correlations were statistically significant, although this also reflects the massive sample size (n = 27,285):

- The transformational and passive/avoidant leadership subscales were negatively correlated, ranging from −0.27 (Idealized Behaviors with Management by Exception–Passive) to −0.49 (Idealized Attributes and Laissez-Faire), providing evidence of divergent validity.
- Consistent with the absence of moderate inter-factor correlations, the two transactional leadership subscales demonstrated different associations with subscales in the other types of leadership.
- Correlations between Contingent Reward (theoretically a transactional leadership subscale) and the five transformational leadership subscales ranged from 0.61 (two subscales: Idealized Behaviors and Intellectual Stimulation) to 0.68 (Individual Consideration). Alternatively, Contingent Reward had a negative relation with two passive/avoidant leadership subscales (r = −0.32 with Management by Exception–Passive and −0.44 with Laissez-Faire). These correlations provide evidence that the Contingent Reward subscale is more similar to the transformational leadership subscales instead of belonging under a different leadership style.
- The other purported transactional leadership subscale (Management by Exception—Active) ranged from almost no relation to a negative relation (−0.12 [Individual Consideration] to 0.08 [Inspirational Motivation]; one correlation of −0.01 was not statistically significant) with the five transformational leadership subscales, and had similarly low correlations with the

passive/avoidant leadership scales (r = 0.08 [Laissez-Faire] and 0.10 [Management by Exception–Passive]). These low correlations suggest that this subscale does represent a different leadership style than the transformational and the passive/avoidant styles.

Finally, examining correlations between the nine subscales (measuring leadership styles) and the three subscales measuring leadership outcomes (Extra Effort, Effectiveness, and Satisfaction) (n = 27,285):

- The transformational leadership subscales with the three outcome subscales had moderate to strong positive correlations ranging from 0.54 (Idealized Behaviors with Satisfaction) to 0.75 (Idealized Attributes with Satisfaction), providing evidence of concurrent validity.
- The passive/avoidant leadership subscales were negatively correlated with the outcome subscales (r ranged from −0.33 [Management by Exception–Passive with Extra Effort] to −0.56 [Laissez-Faire with Effectiveness]), suggesting some evidence of concurrent validity.
- The transactional leadership subscales differed in their relation to outcomes: Similar to the transformational leadership subscales, the Contingent Reward subscale had moderate positive correlations with outcomes (r ranged from 0.63 [Extra Effort] to 0.67 [Effectiveness]), providing evidence of concurrent validity. The other transactional leadership subscale Management by Exception—Active had very low negative correlations with outcomes (r ranged from −0.06 [Extra Effort and Effectiveness] to −0.12 [Satisfaction]), suggesting lack of concurrent validity.

Concurrent validity: The developers summarize results from a number of studies, including meta-analyses, showing relationships between leadership styles as measured by the MLQ and performance and outcome measures, such as organizational commitment from staff or role conflict and interpersonal relationships. The studies represent a variety of settings, including the military and school systems, and were conducted in the United States and internationally. The developers note that the studies demonstrated positive associations between transformational leadership subscales and positive outcomes involving organizational effectiveness and performance (evidence of concurrent validity). The study reported weaker associations between the transactional leadership subscales and these organizational outcomes. Finally, the studies showed weak or negative associations between the passive/avoidant subscales and these outcomes.[25]

(3) Predictive validity: The developers describe studies comparing the nine leadership subscale scores with later outcome measures. These studies have found that the MLQ predicts combat readiness in the military one month later, as well as market share, customer satisfaction, and firm performance in financial institutions at least one year later. In particular, transformational leadership subscale scores predicted favorable outcomes.

Bias Analysis: No information available.

Training Support: The developers provide information on administration and scoring. The MLQ can be administered through the publisher's online system, which handles data collection, analysis, and reporting of results.

[25] Studies have also found relationships between transformational leadership and other domains, such as ethics, creativity, organizational culture, and personality traits; some of these studies may have used measures other than the MLQ.

Key Considerations for Early Care and Education (ECE): The MLQ uses general terminology and is not setting specific, so it likely can be used in ECE settings without substantive adaptation. Although psychometric evidence for the MLQ is drawn from a very large sample, members of that sample likely resemble samples used to develop other measures in this compendium from the business and management field. Those samples tend to differ from leaders and teaching staff in ECE settings along characteristics such as gender and level of education; therefore, the MLQ findings might not be generalizable to ECE settings. Because it either asks leaders to rate themselves or others to rate their leaders, it is most applicable to those serving in formal leadership roles, such as program or center directors. The language of a few items involves terms with specific meanings in ECE settings, such as teaching, coaching, or standards; those might need to be clarified to ensure ECE respondents think of the general meaning of the term when taking the survey.

Previous Version: The current version is the third edition. Earlier versions (the MLQ Form 5R and the MLQ Form 1) demonstrated theoretical and psychometric issues in research use, including grouping items on behaviors and outcomes within subscales, weak divergent validity among the leadership styles measured, and failure to replicate the original factor structure. The developers investigated these issues and revised the MLQ, including splitting two of the original seven subscales (Charisma was divided into Idealized Behavior and Idealized Attributes; Management by Exception was divided into Active and Passive versions) and adding the three leadership outcome subscales.

References:

Avolio, B.J., and B.M. Bass. *Multifactor Leadership Questionnaire, Third Edition: Manual and Sample Set*. Menlo Park, CA: Mind Garden, Inc., 2004.

Organizational Climate Descriptive Questionnaire (OCDQ-RE), 1991

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Multiple (K–12 education [elementary], early care and education [ECE])

Content

What leaders do (Foster respect and learning)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

Administration characteristics

Respondent: Teachers

Level of measure: Site

Data sources: Survey (mode–group-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills with some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 10 minutes, 42 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States (New Jersey)

Setting: School

Sample: 1,071 teachers were randomly selected across 70 elementary schools representing a broad range of schools

Year of development: 1991

Measure performance

Reliability: 3 (meets minimum acceptability rating—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

Availability

2-Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Wayne K. Hoy, C. John Tarter, and Robert B. Kottkamp
Measure website: https://www.waynekhoy.com/ocdq-re/

Narrative

Description: The revised Organizational Climate Descriptive Questionnaire (OCDQ-RE) is a group-administered teacher survey developed to measure organizational climate in elementary schools. The measure consists of 42 statements across six subscales (referred to as dimensions) that together describe elementary school principal behavior—(1) Supportive; (2) Directive; (3) Restrictive—and teacher behavior—(4) Collegial; (5) Intimate; and (6) Disengaged. The three subscales of teacher behavior are used to define the openness of teacher interactions, whereas the three subscales of principal behavior together define the openness (or closedness) of a principal's leadership behavior. The Supportive and Directive OCDQ-RE subscales contain content on "What leaders do," whereas the other four subscales (Restrictive, Collegial, Intimate, and Disengaged) have content on "Center culture, climate, and communication." The survey is administered on paper and averages 10 minutes to complete. Someone other than the principal should administer the survey. A similar measure (OCDQ-RS) was also developed in 1991 for secondary schools. The OCDQ-RE has been adapted for the ECE setting serving infants, toddlers, and preschoolers (Dennis and O'Connor 2013; Cannon et al. 2019).

Uses of Information: The developers state that the OCDQ-RE could be used for understanding and improving a school's work environment. The measure provides a snapshot of the school climate, which could be used as a basis for planning changes and implementing new programs. Furthermore, it could be used to assess the results of those changes or initiatives. The measure can also be used for research and evaluation purposes to study the relationship between organizational climate and student outcomes or classroom quality.

Methods of Scoring: Each item is scored on a 4-point frequency scale by assigning a value of 1 through 4 to the items: rarely occurs (1), sometimes occurs (2), often occurs (3), and very frequently occurs (4). Though teachers are surveyed, the level of measure is intended to be the school so responses are aggregated and divided by the number of teachers to calculate an average score for the school for each of the 42 items. These school-level average item scores can be added together within their corresponding subscales to obtain a subscale score for the six subscales, which represents the school's climate profile. The subscale scores can then be assigned a standard score with a mean of 500 and a standard deviation of 100, using the sample of New Jersey elementary schools that participated in the testing. Scores for principal openness (or closedness) and teacher openness can also be calculated based on the three relevant behavior subscales and indexed. The developers created a computer scoring program to assist with scoring, but it is not required. Hoy et al. (1991) provides details on how to calculate each openness index (Chapter 2, p. 34–35).

Interpretability: The developers provide a range to help interpret the standard subscale scores and openness scores. For example, a standard score of 500 for any of the subscales is average as compared to the sample of New Jersey elementary schools that participated in the testing. A score of 400 or 600 is one standard deviation away from the average score in New Jersey schools in 1991, and is lower or higher, respectively, than 84 percent of the schools in the New Jersey sample. The openness scores can be interpreted in the same way, with a score of 500 being average.

The developers also created four prototypic profiles of school climate—open, engaged, disengaged, and closed—based on the results of their second-order factor analysis conducted on the New Jersey sample. A school with an open climate will have high supportive, collegial, and intimate scores and low directive, restrictive, and disengaged scores. Conversely, a school with a closed climate will have low supportive, collegial, and intimate scores, and high directive, restrictive, and disengaged scores. All of the standard scores are easy to calculate, and can be interpreted based on the ranges provided. The developers envision principals or other school administrators interpreting their schools' results. Please review Chapter 7 of Hoy et al. (1991) for more information on scoring the OCDQ-RE and interpreting the results.

Reliability:

(1) Internal consistency reliability: The developers reported alpha coefficients for each of the six subscales: Supportive (0.95), Directive (0.89), Restrictive (0.80), Collegial (0.90), Intimate (0.85), and Disengaged (0.75) based on their elementary school sample in New Jersey. Cronbach's alpha coefficient for the Supportive Behavior subscale was 0.94 at the center level within the national Early Head Start Family and Child Experiences Survey (Baby FACES) – 2018 (infant and toddler programs).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternative form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: A pilot study was conducted with 152 teachers across 38 elementary schools in New Jersey to refine the OCDQ-RE measure. Items with low factor loadings (< 0.30) were removed, and 42 items remained in the measure.

(2) Construct/concurrent validity:

Construct validity: The developers conducted an exploratory factor analysis with a six-factor solution. The factor loadings ranged from 0.46 to 0.90.

The developers conducted a second-order factor analysis on the subscales. The Disengaged, Intimate, and Collegial subscales loaded strongly on one factor, teacher openness, whereas the Restrictive, Supportive, and Directive subscales loaded strongly on the second factor, principal openness (or closedness). The factor loadings ranged from −0.84 (Disengaged) to 0.77 (Collegial) for teacher openness, and −0.65 (Supportive) to 0.83 (Directive) for principal closedness.

Concurrent validity: The original OCDQ index of openness was used to correlate each subscale of openness in the revised measure. The new index of teacher openness and the new index of principal openness significantly correlated ($p \leq .01$) with the original general school openness index ($r = 0.67$ and $r = 0.52$, respectively), suggesting evidence of convergent validity.

The developers examined bivariate correlations between the six subscales of the OCDQ-RE and faculty trust in others ($n = 44$). Three subscales—Collegiality ($r = 0.13$), Intimate ($r = 0.25$, $p \leq .05$), and Supportive ($r = 0.58$, $p \leq .01$) had a positive relationship with faculty trust in the principal, while the three subscales of Restrictive ($r = -0.13$), Directive ($r = -0.15$), and Disengaged ($r = -0.28$, $p \leq .05$) had a negative relationship. Similarly, the three subscales of Collegiality ($r = 0.67$, $p \leq .01$), Intimate ($r = 0.43$, $p \leq .01$), and Supportive ($r = 0.43$, $p \leq .01$) had a positive relationship with faculty trust in the colleagues, while the three subscales of Restrictive ($r = -0.22$), Directive ($r = -0.06$), and Disengaged ($r = -0.60$, $p \leq .01$) had a negative relationship. All correlations except the ones of Directive and Restrictive with faculty trust in the principal and colleagues and the one of Collegiality with faculty trust in the principal provide evidence of concurrent validity.

Principal openness correlations with trust ranged from 0.37 (for trust in colleagues) to 0.49 (for trust in principal), both statistically significant ($p \leq .01$) ($n = 44$). Teacher openness correlations with trust ranged from 0.25 (for trust in principal) to 0.72 (for trust in colleagues), both statistically significant ($p \leq .05$ and $p \leq .01$, respectively). The subscales of Directive, Restrictive and Disengaged consistently showed a negative relationship with the dependent variable, while Supportive, Collegial, and Intimate were positive, suggesting concurrent validity.

The developers conducted regression analyses with all subscales in association with principal and faculty trust to examine how much variance is explained (n = 44). The climate subscales explained 0.66 of the variance in faculty trust in the principal, with supportive principal behavior the only significant predictor. The climate subscales also explained 0.75 of the variance in faculty trust in colleagues, with the Collegial and reverse-coded Disengaged subscales significantly contributing to the variance. No controls appear to be included.

The developers also studied the association between a school's organizational climate (as measured through the OCDQ-RE) and perceived organizational effectiveness (n = 44). Four subscales were positively correlated to perceived effectiveness—Directive (r = 0.06), Supportive (r = 0.29 $p\leq.05$), Intimate (r = 0.36, $p\leq.01$), and Collegial (r = 0.54, $p\leq.01$). The other two subscales—Restrictive (r = -0.23) and Disengaged (r = -0.54, $p\leq.01$)—were negatively correlated. A regression model demonstrated that the subscales explained 0.64 of the variance in school effectiveness, though only the disengaged subscale was a significant predictor. These provide evidence of concurrent validity.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Dennis and O'Connor (2013) and Cannon et al. (2019) adapted the OCDQ-RE for the ECE setting. Some of the terminology within the 42 items needed to be adapted to be relevant to the center setting, director role, and teaching staff and director responsibilities that are different than at elementary schools. For example, Dennis and O'Connor (2013) revised items that used the word *principal* to *director*, and *faculty* to *staff*.

Previous Version: The original version of the OCDQ, with 64 items, was published in 1962 by Halpin and Croft. Hoy et al. (1991) discarded 24 of the original 64 items and added two new items to comprise the 42-item OCDQ-RE. The final OCDQ-RE was based on a review of the original items, a pilot study that included new items, review of the validity evidence for the items, and a factor analysis. No additional changes appear to have been made to the OCDQ-RE after validation.

References:

Hoy, W.K., C.J. Tarter, and R.B. Kottkamp. *Open Schools/Healthy Schools: Measuring Organizational Climate*. Newbury Park, CA: Sage, 1991.

Dennis, S.E., and E. O'Connor. "Reexamining Quality in Early Childhood Education: Exploring the Relationship Between the Organizational Climate and the Classroom." *Journal of Research in Childhood Education*, vol. 27, no. 1, 2013, pp. 74–92. doi:10.1080/02568543.2012. 739589

Cannon, J., K. Schellenberger, A. Defnet, A. Bloomenthal, Y. Xue, and C.A. Vogel. *Baby FACES 2018: Data Users' Guide.* Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2019.

Preschool Instructional Leadership Survey, Version 2 (PILS), 2017

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Early care and education (ECE)

Content

What leaders do (Promote quality practices)

Center structures and staff supports (Training and professional development, collaborative planning time)

Administration characteristics

Respondent: Other (instructional leader)

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–self-report)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 17 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States (Illinois)

Setting: ECE center

Sample: 318 school-based and community-based early childhood instructional leaders; 71% White; 90% had a bachelor's degree or higher; majority had 6–10 years of experience

Year of development: 2017

Measure performance*

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

*Version examined based on measure adjustments that do not appear to have been revalidated with final set of items.

Availability

3-Permission required from developer(s), no known costs
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Heather L. Horsley and Karen Fong

Narrative

Description: The Preschool Instructional Leadership Survey (PILS) is a new measure initially developed as part of the Ounce of Prevention Lead Learn Excel Instructional Supports initiative, a three-year evaluation funded through the federal Race to the Top Early Learning Challenge grant program, with a focus on supporting instruction in pre-K classrooms. It measures how often early childhood leaders demonstrate instructional leadership behaviors through three subscales (referred to as domains): Effective Leadership, Professional Capacity, and Instructional Guidance. The PILS is a self-administered self-report with 17 items across three subscales—Effective Leadership (n = 6), Instructional Guidance (n = 6), and Professional Capacity (n = 5). The Effective Leadership subscale relates to "What leaders do," whereas the Instructional Guidance and Professional Capacity subscales relate to "What leaders do" and "Center structures and staff supports." The PILS is still under development. Please contact the developers for additional information.

Uses of Information: The developers developed the PILS as part of an evaluation of the Ounce of Prevention Lead Learn Excel Instructional Supports initiative; therefore, it can be used for research or evaluation. The PILS can also be used for development or improvement purposes, as the developers describe using the PILS before and after the evaluation, indicating the PILS can be used to measure how instructional leadership behavior may change as a result of program improvement initiatives.

Methods of Scoring: The PILS uses a four-point frequency scale. The instructional leader assigns a value of one through four to the items: less than once a month (1), once or twice a month (2), once or twice a week (3), and more than twice a week (4). Average scores can be calculated for each subscale for the individual leader.

Interpretability: The average subscale scores can range from 1 through 4, with higher scores indicating that the behaviors associated with a particular subscale are occurring more frequently.

Reliability:

(1) Internal consistency reliability: As a unidimensional scale (that is, a single factor or scale), the PILS person reliability[26] was 0.84. When treated as multidimensional with three subscales, person reliability estimates ranged from 0.66 to 0.72 for the three subscales.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers conducted a literature review on leadership practices and used the Head Start Performance Standards to inform the survey content. The PILS originally included 30 items, but expert reviewers advised removing items that did not focus on leadership behavior. The PILS survey was fielded with 20 items, however, only 18 of those items were part of the validation study. The survey items were further refined as a result of the study–removing an item from Instructional Guidance and moving one item under the Professional Capacity subscale to the Instructional Guidance subscale, resulting in the final 17-item measure.

[26] Person reliability in Rasch is similar to Cronbach's alpha.

(2) Construct/concurrent validity:

Construct validity: The developers initially validated the PILS using Rasch principal components analysis (PCA) of residuals to assess dimensionality and determined the measure to be multidimensional rather than unidimensional. The correlations between the subscales ranged from 0.87 (Effective Leadership with Professional Capacity) to 0.93 (Effective Leadership with Instructional Guidance, and Instructional Guidance with Professional Capacity). The developers did not indicate sample size or statistical significance. No convergent or divergent validity information was provided by the developers of this measure.

The developers also examined the Andrich threshold and found the rating scale did not work as intended, with response options four and five underused. As a result, the developers removed a fifth response option (daily) to settle on a four-point rating scale (see Methods of Scoring).

Concurrent validity: No concurrent validity information was provided by the developers of this measure.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Designed for use in ECE with a sample of 318 school-based and community-based early childhood instructional leaders. Respondents were 71 percent White, 90 percent had a bachelor's degree or higher, and a majority had 6-10 years of experience. The developers made changes to the measure as a result of the validation study and intended to field the revised measure in a future study.

Previous Version: An earlier version of the PILS was used as part of the Lead Learn Excel evaluation. The developers made revisions to the items and rating scale described in the validity section above.

References:

Fong, K., and H.L. Horsley. "A Multidimensional Rasch Analysis of the Preschool Instructional Leadership Survey." Chicago, IL: University of Illinois, 2017.

Horsley, H.L., J.M. Vasquez, and W.H. Teale. "Evaluation of the Ounce Lead Learn Excel Final Report 2014–2017." Chicago, IL: University of Illinois at Chicago Center for Literacy, June 2017.

Principal Instructional Management Rating Scale (PIMRS), 2015

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: K–12 Education

Content

What leaders do (Promote quality practices, foster respect and learning, establish vision, manage efficient operations)

Center culture, climate, and communication (Family relationships)

Administration characteristics

Respondent: Principal, teachers, other (principal's supervisor)

Level of measure: Site, individual

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 50 items

Administration interval: Not described

Languages available: English, Spanish, other (Malay, Chinese, Arabic, Thai, five other languages)

Technical information

Development sample

Locale: United States and non-U.S. (primarily Asia)

Setting: School

Sample: (1) 43 studies with a combined 2,508 principals (44% from U.S. studies) and 12,064 teachers (71% from U.S. studies) in primary and secondary schools (meta-reliability study); (2) 19 studies with a combined 649 principals (65% from U.S. studies) and 4,370 teachers (% from U.S. studies unknown[27]) in primary and secondary schools (meta-validation study)

Year of development: 1984–2013

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

4-Permission required, with costs
Material, training, and scoring costs: Contact developer

Developer(s)/publisher contacts

Developer(s): Philip Hallinger
Measure website: http://philiphallinger.com/purchasing-pimrs/

[27] The list of studies for the teachers in sample 2 omitted the largest individual study, so it is not clear if it was a U.S. or international study. If the former, then the percent of teachers from U.S. studies is 89%; if the latter, then it is 53%.

Narrative

Description: The Principal Instructional Management Rating Scale (PIMRS) is a self-administered survey that measures how often a school principal engages in functions, practices, and behaviors related to instructional leadership. The PIMRS can be completed as a self-report by the principal or as a report about the principal by the teachers at the principal's school or the district supervisor who oversees the principal. The PIMRS has parallel forms for each of these three types of respondents; the item content is the same. The PIMRS has a total of 50 items evenly divided between 10 subscales, which in turn are organized under three dimensions of instructional leadership:

1. The Define the School Mission dimension has two subscales: Frame the School Goals and Communicate the School Goals.
2. The Manage the Instructional Program dimension has three subscales: Supervise and Evaluate Instruction, Coordinate the Curriculum, and Monitor Student Progress.
3. The Develop a Positive School Learning Climate dimension has five subscales: Protect Instructional Time, Maintain High Visibility, Provide Incentives for Teachers, Promote Professional Development, and Provide Incentives for Learning.

The PIMRS was developed for use in K–12 education settings. There is also a short form available for teachers, with 22 items. The content of the PIMRS primarily falls under the "What leaders do" component of the ExCELS theory of change, as all 10 subscales consist of items describing a principal's behaviors. The Provide Incentives for Learning subscale and some items from other subscales (Communicate the Schools Goals, Monitor Student Progress, Protect Instructional Time, and Maintain High Visibility) also reflect the "Center culture, climate, and communication" component.

Uses of Information: The PIMRS was originally developed for research on instructional leadership to meet the need for a validated measure in this area. The developer notes that it can also be used for development and improvement. For example, school districts can use the PIMRS to assess principals' needs for professional development and support and to evaluate principals.

Methods of Scoring: Each item is scored on a 5-point frequency scale with the options almost never (1), seldom (2), sometimes (3), frequently (4), and almost always (5). Scores can be calculated and analyzed at the item, subscale, or dimension level, or as a total score. The developer recommends subscale- and dimension-level scores as the most appropriate. Because each subscale has five items, dimension-level scores are the same if calculated by averaging item scores or subscale scores.

When multiple respondents fill out the PIMRS for a single principal (usually this is the teachers at the school, but this could also occur if multiple district officials rate a principal), score distributions can be calculated in addition to averages. For example, two principals could have the same average scores, but one could have very high scores from some teachers and very low scores from others, whereas the other could have consistently middle-range scores from almost all teachers. When more than one type of respondent fills out the PIMRS, scores can also be compared to each other. For example, the principal's self-rating score could be compared to the average scores from the teachers' rating the principal. However, scores from different respondent types should not be combined—for example, a principal's self-reported score should not be included in the average of scores from teachers.

Interpretability: Higher scores reflect greater engagement by principals in practices and behaviors that constitute instructional leadership. However, as the developer cautions, scores are not a direct measure of principals' effectiveness. For example, the highest response option ("almost always") is not necessarily required for a principal to be most optimally engaged in some behaviors. Accordingly, the developer recommends interpreting average scores of 4 or higher as reflecting a high degree of engagement. The

developer also notes that school, principal, teacher, and student characteristics—such as school grade span, principal tenure, teacher experience, and student achievement—should be considered contextually when interpreting results. Finally, studies using the PIMRS have found that principal self-reported ratings are higher than teacher-reported ratings of the principal, although this is consistent with other results from the education and management fields. As part of publishing the PIMRS, the developer provides additional information on interpreting PIMRS scores.

Reliability:

(1) Internal consistency reliability: The developer led an analysis of reliability through a meta-analysis of data from 43 studies that have used the PIMRS since its creation (Hallinger et al. 2013). Limiting to studies conducted in the United States, Cronbach's alpha for the principal self-report samples ranged from 0.75 (Provide Incentives for Teachers) to 0.86 (Frame the School Goals) for the 10 subscales, 0.90 (Define the School Mission) to 0.93 (Manage the Instructional Program) for the three dimensions, and was 0.96 for the whole measure. These results were based on statistics calculated from raw data for some studies and statistics published in the study's report for other studies. For teacher reports on principals from United States studies, the developer and co-author of that study reported two sets of results with different statistics: (1) estimates using generalizability theory coefficients calculated from raw data from some studies ranged from 0.91 (Maintain High Visibility and Protect Instructional Time) to 0.96 (Frame the School Goals and Communicate the School Goals) for the 10 subscales, were 0.98 for all three dimensions, and was 0.99 for the whole measure; (2) estimates based on published Cronbach's alpha results from other studies ranged from 0.83 (Maintain High Visibility and Provide Incentives for Teachers) to 0.91 (Provide Incentives for Learners) for the 10 subscales, 0.90 (Define the School Mission) to 0.94 (Develop a Positive School Learning Climate) for the three dimensions, and was 0.97 for the whole measure.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developer initially used theory and research on instructional leadership and semi-structured interviews with principals and district administrators to create the list of three dimensions and, at the time, 11 subscales for the PIMRS. The same sources were used to create a list of instructional leadership-related job behaviors that was converted into a draft set of 93 potential items. Four expert reviewers (three principals and a vice principal) who had not been involved in creating the list of behaviors were asked to review the potential items and assign each item to 1 of the 11 subscales. After this process led to assignments for 81 items, the developer consulted with the superintendent whose district participated in the initial validation study to drop 10 items to balance the subscale lengths and reduce the overall measure length. This left the 71 items that constituted the initial version of the PIMRS, which was used in the initial validation study (Hallinger and Murphy 1985). The average agreement by the judges on item placement ranged from 80 to 100 percent for each subscale, exceeding the developer's target of 80 percent agreement. Based on the results of the initial validation study, the developer dropped one subscale and several other items, leading to the current version of 10 subscales and 50 items; the rest of the evidence in this profile refers to the current version of the PIMRS.

(2) Construct/concurrent validity:

Construct validity: In a 2015 meta-analysis of data from 19 studies that have used the PIMRS in the preceding decade (Hallinger and Wang 2015), the developer and co-author of the study conducted Rasch analysis of both principal self-ratings and teacher ratings of principals. When they assessed item fit statistics of the principal self-ratings, all but 2 of the 50 items met the developer and co-author's standard of the outfit mean square (between 0.6 and 1.4, because the expected value should be 1). Thirty-six of the 50 items met the standard for the correlation between the item and total score (greater than 0.50). The remaining 14 item-to-total score correlations, which did not meet the standard, were between 0.42 and 0.49. All of the 16 items that did not meet both standards were in the Develop a Positive School Learning Climate dimension (which has a total of five subscales and 25 items). Similarly, for the teacher ratings of principals, all but 4 items met the standard for outfit mean square—3 in the Develop a Positive School Learning Climate dimension and 1 in the Manage the Instructional Program dimension. All 50 items met the standard for correlation between item and total score, with a range of 0.53 to 0.80.

As part of the Rasch analysis, the developer and co-author of the study noted comparing theoretical construct maps to Wright maps generated from study data to assess item difficulty. For both principal self-ratings and teacher ratings of principals, the developer and co-author found a high degree of alignment between the construct maps and Wright maps on all three dimensions (suggesting that these dimensions are measuring the intended constructs), although the alignment for the Develop a Positive School Learning Climate dimension was weaker than for the other two dimensions. Overall, these results indicate there is weaker evidence of construct validity for the subscales in the Develop a Positive School Learning Climate dimension compared to the subscales in the other two dimensions.

Concurrent validity: In the 2015 validation meta-analysis, the developer and co-author of the study reported that several studies have compared the PIMRS to other leadership measures, specifically the Multifactor Leadership Questionnaire and Leadership Practices Inventory, and that the developer and co-author are reanalyzing data from two of those studies to assess convergent and divergent validity. The developer and co-author noted that preliminary results support the convergent and divergent validity of PIMRS, but the results are not yet available. To examine criterion-related validity, the developer and co-author also reported they are conducting a research synthesis and meta-analysis of studies that have used the PIMRS to investigate relationships between instructional leadership and either student achievement or school effectiveness. Results from this study are also not yet available.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: As part of Rasch analysis in the 2015 meta-analysis, the developer and co-author of the study conducted differential item functioning (DIF) analysis between subgroups of principals of primary and secondary schools. The analysis did not find a consistent pattern of results favoring either subgroup. For principal self-ratings, differences in scores between primary and secondary principals were small for each of the three dimensions (differences in average scores between the two subgroups of close to 0 for Define the School Mission, 0.24 for Manage the Instructional Program, and 0.14 for Develop a Positive School Learning Climate, with secondary principals scoring higher on the latter two). However, for teacher ratings of principals, differences in scores were large, moderate, and small, respectively (differences in average scores of 1.03 for Define the School Mission, 0.53 for Manage the Instructional Program, and 0.12 for Develop a Positive School Learning Climate, with primary principals scoring higher on the first dimension and the direction not stated for the latter two).

Examining individual items, the developer and co-author found that based on principal self-ratings, 6 of the 50 items had DIF sizes (differences in average scores between the two subgroups) larger than the authors' selected threshold of 0.50, and based on teacher ratings of principals, 3 items had DIF sizes larger than 0.50. Most of these items were only slightly above the threshold; 0.68 was the largest DIF size mentioned. However, the dimensions with larger overall differences in scores did not necessarily have more individual items with large DIF sizes; for example, teacher ratings on Define the School Mission had the largest overall difference between primary and secondary principals but also did not have any individual items with a DIF size above 0.50.

The developer and co-author also compared inter-dimension correlations and within-dimension variances for each of the three dimensions for elementary-grade and secondary-grade principals. Inter-dimension correlations were very similar for each dimension for both principal self-ratings and teacher ratings of principals (the six differentials ranged from nearly 0 to approximately 0.09), leading the developer and co-author to conclude that because the structure of the PIMRS was stable across elementary and secondary principals, any DIF effects are canceled out in overall scores. The within-dimension variances were consistently much larger for the self-rating of the secondary-grade principals, indicating that secondary principals had a wider range of scores than elementary principals, perhaps related to differences in roles in school serving older compared to younger. The developer and co-author did not recommend dropping any items based on the DIF analysis.

Training Support: Materials from the developer, such as a technical report and a user manual, include guidance on using the measure, including analyzing and interpreting results.

Key Considerations for Early Care and Education (ECE): The PIMRS was developed for K–12 education settings and describes specific behaviors involving instructional leadership in these settings. Some PIMRS items could apply to ECE settings with minimal revisions, such as items involving school goals, supervising instruction and curriculum, and professional development. Other items, particularly those around monitoring student progress and communicating with students, would likely need to be revised or replaced to focus on communication and interactions with families to reflect children's age or developmental levels. More broadly, several items refer to a school's academic goals or performance; for ECE settings, this might need to be broadened to include goals or performance in multiple developmental domains, such as social-emotional development.

The PIMRS assumes a single principal at a school who both has considerable authority over the school's goals and policies and engages in frequent interaction with teachers and instructional activities. This approach might need to be adapted for ECE settings.

Previous Version: The initial version of the PIMRS contained 11 subscales and 71 items. Based on the results of the initial validation study (Hallinger and Murphy 1985), one subscale and several other items were dropped, leading to the current version of 10 subscales and 50 items that is described throughout this profile. The dropped subscale was Maintain High Academic Standards, which had been part of the Develop a Positive School Learning Climate dimension.

References:

Hallinger, P., and J. Murphy. "Assessing the Instructional Leadership Behavior of Principals." *Elementary School Journal*, vol. 86, no. 2, 1985, pp. 217–248.

Hallinger, P., W-C. Wang, and C-W. Chen. "Assessing the Measurement Properties of the Principal Instructional Management Rating Scale: A Meta-Analysis of Reliability Studies." *Educational Administration Quarterly*, vol. 49, no. 2, 2013, pp. 272–309.

Hallinger, P., and W-C. Wang. *Assessing Instructional Leadership with the Principal Instructional Management Rating Scale.* Dordrecht, Netherlands: Springer, 2015.

Program Administration Scale, Second Edition (PAS), 2011

Purpose and context

Purpose: Development/improvement, monitoring, research/evaluation

Field: Early care and education (ECE)

Content

What leaders do (Promote quality practices, promote family/community partnerships, manage efficient operations)

What leaders bring (Administrative, business, and management knowledge and skills; education and experience)

Center culture, climate, and communication (Culture of respect, shared growth, and learning)

Center practices (Operational procedures and policies; regular assessment of program, classroom, and children; family engagement)

Center structures and staff supports (Training and professional development, conflict resolution, accountability structures)

Administration characteristics

Respondent: Manager/director/principal

Level of measure: Site

Data sources: Survey (mode–interview, report level–report of others), direct observation, document review

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Highly trained individual

Training for administration: Extensive > 2 hours

Ease of administration and scoring: 3 (administered and/or scored by a highly trained individual)

Time/length: 4 hours, 25 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States (25 states)

Setting: ECE center

Sample: 564 ECE centers, mean licensing capacity of 90 children; 69% nonprofit; 35% received Head Start funding, 36% received pre-K funding; 23% affiliated with faith-based organizations; 31% accredited by the National Association for the Education of Young Children (NAEYC)

Year of development: 2006–2009

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available (see Narrative)

Availability

4-Permission required, with costs

Material, training, and scoring costs: $25 fee to purchase manual; $50 to $100 registration fee to participate in optional online "getting ready for the PAS" depending on the number of modules purchased

Developer(s)/publisher contacts

Developer(s): Teri N. Talan and Paula J. Bloom

Publisher: Teachers College Press
1234 Amsterdam Avenue
New York, NY 10027

Measure website: https://mccormickcenter.nl.edu/library/program-administration-scale-pas-2nd-ed/

Narrative

Description: The Program Administration Scale (PAS) was developed to measure leadership and management functions in the center-based early care and education setting through interviews, observations, and document reviews. Leadership functions involve goal setting, clarifying and affirming an organization's values and vision, and setting a path to achieving that vision. Management functions are task oriented and involve implementing the organization's mission. The measure consists of 25 items, 2 of which are optional, grouped within 10 subscales: (1) Human Resources Development, (2) Personnel Cost and Allocation, (3) Center Operations, (4) Child Assessment, (5) Fiscal Management, (6) Program Planning and Evaluation, (7) Family Partnerships, (8) Marketing and Public Relations, (9) Technology, and (10) Staff Qualifications. The Human Resources Development, Fiscal Management, and Marketing and Public Relations subscales include items related to "What leaders do" and "Center structures and staff supports." The Personnel Cost and Allocation subscale asks about "What leaders do," "Center practices," and "Center structures and staff supports." The Staff Qualifications subscale includes items relating to "What leaders bring" only. The Child Assessment, Program Planning and Evaluation, Family Partnerships, and Technology subscales have content within the "Center practices" area only. The Center Operations subscale has items relating to "Center culture, climate, and communication," "Center practices," and "Center structures and staff supports." The Staff Qualifications subscale has additional content referring to center characteristics, and the Fiscal Management subscale has additional content referring to policy, regulatory, and fiscal infrastructure, which are all outside the scope of this compendium. The PAS measures the quality of each item through two to five indicators. Although the developers designed the PAS to be used by a senior program administrator, a trained independent assessor can also administer the PAS. If the PAS is being administered by an independent assessor, it will take approximately four hours to complete including a two-hour interview with the site administrator (director, manager, coordinator, or principal) and two hours for a facility observation and document review. If a senior program administrator is using the PAS, it may take less amount of time to complete.

Uses of Information: The developers state that the PAS could be used for program development and self-improvement by using the profile to benchmark the progress a center makes in meeting program goals. State and local quality improvement initiatives can also include the PAS as a monitoring and technical assistance tool. Some quality rating and improvement systems use the PAS as part of their rating process for indicators on program administration, management, and leadership (BUILD Initiative and Child Trends 2017). PAS can also be used for research and evaluation as a way to benchmark program quality.

Methods of Scoring: Each indicator for items 1 through 21 are rated by writing a Y (yes) or N (no) across four columns that represent a quality scale from 1 to 7: inadequate (1), minimal (3), good (5) and excellent (7). Scores for items 1 through 21 are then calculated based on how many Y and N responses were provided for the indicators within the four quality categories. For example, a score of 1 is given for an item if all indicators under the inadequate (1) column are rated Y (yes). A score of 7 is given to an item if all indicators under the first column—inadequate (1)—are rated an N (no) and the indicators under the second, third, and fourth columns representing minimal (3), good (5), and excellent (7), are rated Y (yes). Three worksheets—the Administrator Qualifications Worksheet, the Teaching Staff Qualifications Worksheet, and the Summary of Teaching Staff Qualifications Worksheet—accompany the PAS and are used to score items 22 through 25, which refer to staff qualifications. To calculate a total PAS score for the center, the score for each of the 25 items are added together. To calculate an average PAS score for the center, the total PAS score is divided by the number of items scored.

Interpretability: The subscale scores can range from 1 to 7, with higher scores indicating higher quality of program administration practices. Once a center's scores are calculated, a PAS profile can be completed for the center to summarize the results, and the original PAS profile can be used as a benchmark as a center works on meeting internal goals.

Reliability:

(1) Internal consistency reliability: Cronbach's alpha coefficient for the PAS total scale was 0.86 for the current version and 0.85 in the first edition.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternative form.

(4) Inter-rater reliability: The PAS assessors in the validation study averaged an overall inter-rater reliability score of 94% on the 25 items, with a range of 86% to 100%. Assessors participated in a four-day training, and during which they were videotaped and rated on their ability to match the PAS item scores within one point. In the first edition, the inter-rater reliability score ranged from 81% to 95%.

Validity:

(1) Content validity: A panel of 10 expert reviewers established content validity for the first edition of the PAS. The panel evaluated each subscale, item, and indicator to ensure they related to key leadership and management practices. Because the changes between the first and second edition of the PAS were considered minimal by the developers, they did not establish content validity again for the second edition.

(2) Construct/concurrent validity:

Construct validity: To confirm that the subscales measure distinct aspects of early childhood administration, the developers conducted correlation analysis between the 10 subscales. The intercorrelations for the 10 subscales ranged from 0.04 (Human Resources Development subscale with the Staff Qualifications subscale) to 0.72 (Personnel Cost and Allocation subscale with the Child Assessment subscale), with a median value of 0.33. The findings mirror the first edition, with the intercorrelations ranging from 0.09 to 0.63, and a median value of 0.33.

Concurrent validity: The developers analyzed correlation between the PAS first edition subscales and the subscales of two other measures—the Professional Growth subscale of the Early Childhood Work Environment Survey (ECWES) and the Parents and Staff subscale of the Early Childhood Environment Rating Scale-Revised (ECERS-R)—that measure organizational effectiveness in the early care and education setting ($n = 67$). The correlation analysis showed that the PAS first edition was related to the ECWES and ECERS-R but was not redundant with those measures' subscales. The correlations between the PAS and the ECWES Professional Growth subscale ranged from 0.05 (two subscales: Child Assessment subscale and Marketing and Public Relations subscale) to 0.43 (Family Partnerships subscale) for the PAS subscales and 0.52 for the PAS total score. The correlations between the PAS and the ECERS-R Parents and Staff subscale ranged from 0.10 (Marketing and Public Relations subscale) to 0.47 (Fiscal Management subscale) from the PAS subscales and 0.53 for the PAS total score. The developers note a similar study by Kagan et al. (2008) that confirmed the PAS and ECERS-R captured distinct dimensions of quality based on correlational analysis ($r = 0.52$; $p \leq .01$) and factor analysis (with two distinct factors aligned to the measures). The moderate correlations suggest the PAS measures similar, but still distinct constructs than the ECERS-R and ECWES, evidence of divergent validity.

The developers summarized three studies that show positive associations between the PAS and organizational climate (as measured by the ECWES), classroom quality (as measured by the ECERS-R and the Early Language and Literacy Classroom Observation), and director qualifications (Lower and Cassidy 2007; MCECL 2010; Rous et al. 2008), evidence of concurrent validity.

The developers analyzed variance models to determine if the PAS could differentiate between programs varying on numerous characteristics. The analyses demonstrated that the PAS could differentiate between programs of varying quality, size, and type, with programs accredited by NAEYC, medium- and large-sized programs, and nonprofit centers receiving higher PAS scores, showing evidence of concurrent validity.

The developers also summarized four studies that showed the PAS differentiated between different types of programs, further evidence of concurrent validity. Miller and Bogatova (2007) found that programs that met state accreditation standards received higher PAS scores than programs only meeting minimal licensing requirements. MCECL (2007) discovered that programs with high use of local initiative funds averaged higher PAS scores than programs that use fewer funds. Rous et al. (2008) found that programs that used a state professional development framework scored higher on the PAS than programs that did not implement the framework. Arend (2010) found that the quality of management practices (as measured through items from the PAS) were higher for managers with more training.

(3) Predictive validity: Although studies have examined relationships between leadership and management functions and various outcomes, it is not clear if any studies measured outcomes later in time than they measured leadership and management functions.

Bias Analysis: No information available.

Training Support: Although formal training is not required, a "Getting Ready for the PAS" online module is available for a $50 to $100 registration fee, depending on the number of modules being purchased. The module provides a general overview of the PAS, describes each of the 25 PAS items in detail, provides opportunities to practice scoring, and includes instructions on how to prepare for the PAS. The module takes eight hours to complete.

Key Considerations for Early Care and Education (ECE): Designed for use in ECE; validated with a sample of 563 ECE centers with a mean licensing capacity of 90 children.

Previous Version: The first edition of the PAS was published in 2004 by Talan and Bloom. The second edition consists of refined wording of some of the item indicators and additional information to support the measure's reliability and validity.

References:

McCormick Center for Early Childhood Leadership at National Louis University. "Getting Ready for the PAS | Online Module" 2019. Available at https://mccormickcenter.nl.edu/events/getting-ready-for-the-pas-online-module/. Accessed July 19, 2019.

Talen, T.N., and P.J. Bloom. *Program Administration Scale*, 2nd edition. New York: Teachers College Press, 2011.

Program Quality Assessment Form B – Agency Items for Infant-Toddler and Preschool Programs (PQA Form B), 2013

Purpose and context

Purpose: Development/improvement, monitoring, research/evaluation

Field: Early care and education (ECE)

Content

What leaders bring (Pedagogical knowledge, education and experience)

Center culture, climate, and communication (Family relationships)

Center practices (Operational procedures and policies; regular assessment of program, classroom, and children; family engagement)

Center structures and staff supports (Training and professional development)

Administration characteristics

Respondent: Director

Level of measure: Site

Data sources: Survey (mode–interview, report level–report of others) (see Narrative), direct observation

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Highly trained individual

Training for administration: Self-training < 1 hour (see Narrative)

Ease of administration and scoring: 3 (administered and/or scored by a highly trained individual)

Time/length: 24 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States

Setting: ECE center

Sample: 96 infant and toddler group care settings, described as diverse by the authors (Infant-Toddler Form B); 19 classrooms participating in the Michigan School Readiness Program (MSRP) evaluation and 253 classrooms in Head Start and private child care settings (121 in fall 2000, 132 in spring 2001) participating in the Michigan Full-Day Preschool Comparison Study (Preschool Form B)

Year of development: Infant-Toddler Form B, 2001; Preschool Form B, 2000-2001

Measure performance*

Reliability: 2 (all or mostly under minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

*Versions examined based on measure adjustments that do not appear to have been revalidated with final set of items.

Availability

4-Permission required, with costs

Material, training, and scoring costs: A "starter" pack can be purchased from the publisher for $27.95 each, which includes Form A, Form B, and the administrator manual. There is an unknown cost associated with the online use of the PQA, which includes automated scoring, reports, and technical assistance. HighScope also offers online training and reliability tests to certify PQA assessors (see Narrative in Training Support).

Developer(s)/publisher contacts

Developer(s): Ann Epstein, Suzanne Gainsley, Mary Hohmann, Ted Jurkiewicz, Shannon Lockhart, Beth Marshall, and Jeanne Montie

Publisher: HighScope Educational Research Foundation
600 North River Street
Ypsilanti, MI 48198
(734) 485-2000 or (800) 40.PRESS
press@highscope.org

Measure website: www.highscope.org/our-practice/pqa/

Narrative

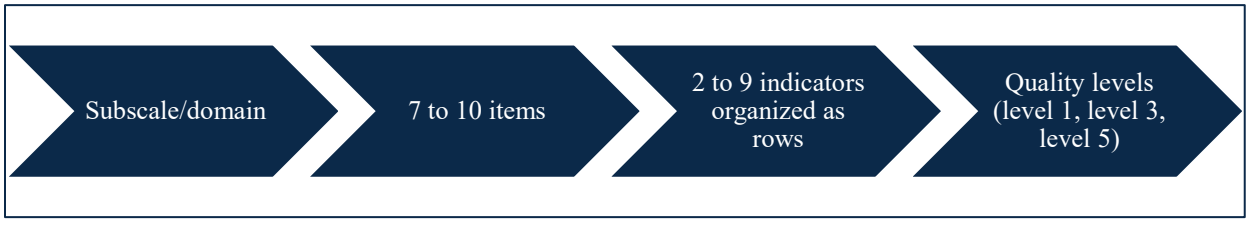
Description: The Program Quality Assessment (PQA) Form B—Agency Items for Infant-Toddler and Preschool Programs was developed by the HighScope Educational Research Foundation to measure program quality in early care and education programs. Although Form A/Form B typically refers to alternate versions of the same measure, in this case, Form A and Form B measure different concepts. Specifically, the PQA Form B relates to program operations or agency-level practices and can be used in programs serving infants, toddlers, and preschool-age children. The PQA can be administered by an independent, trained rater through interviews with the program director or other knowledgeable staff and supplemental observations, or it can be administered as a self-assessment by program staff. HighScope also has a Form A that relates to curriculum implementation; there is an Infant-Toddler version (for programs and classrooms with children from 6 weeks to 36 months old (Hohmann et al. 2016) and a Preschool version (HighScope 2019a). The PQA Form B for Infant-Toddler and Preschool Programs is the focus of this compendium.

The PQA Form B measures three subscales (referred to as domains): Parent Involvement and Family Services; Staff Qualifications and Staff Development; and Program Management. Form B can be administered to programs not using the HighScope curriculum. Each of the subscales consists of 7 to 10 items further divided into a series of indicators organized as rows (some of them are only relevant for particular age groups for example, infant and/or toddler rooms or preschool rooms). All three subscales in the PQA Form B have content relating to "Center practices." The Parent Involvement and Family Services subscale also has content relating to "Center culture, climate, and communication," whereas the Staff Qualifications and Staff Development subscale also includes items relating to "Center structures and staff supports" and "What leaders bring." The PQA Form B also includes content on nonleadership constructs, including center characteristics; policy, regulatory, and fiscal infrastructure; and staff outcomes.

Uses of Information: The developers state that the PQA can be used for development or improvement purposes by identifying staff professional development needs and assessing the quality of implementation of program practices. The PQA can be used to monitor goals and program implementation. The PQA can also be used for research or evaluation purposes to study quality at early care and education centers. The developers emphasize the use of the PQA at center-based Early Head Start and Head Start programs because they used the Early Head Start and Head Start Program Performance Standards as a primary reference when developing the Infant-Toddler and Preschool measures.

Methods of Scoring: The PQA includes descriptions of quality for the different indicators of the subscale items at three anchor points, level-1, level-3, and level-5 visualized in Exhibit I.

Exhibit I. PQA methods of scoring



Raters select the quality level for each indicator (row) in an item. The quality level for each item is then assigned on a 5-point scale based on the number of level-1, level-3, and level-5 indicators associated with

the item. Item scoring varies if an item only includes two indicators or if an item has three or more. For two-indicator items, if both are the same, the item value is the same as the indicator values. If the indicator scores vary, the score assigned to the item depends on the score of the lowest and next highest indicator. For example, an item would have a quality level of 2 if one indicator was marked a 1 but the other indicator was marked a 3 or 5, and an item would have a quality level of 4 if one indicator was marked a 3 and the other was marked a 5. The scoring for items with three or more indicators is similar. Epstein et al. (2013) provides detailed scoring instructions. Once all the items are scored, the rater can complete the summary sheet, which includes the quality scores for the 24 items, a total agency score (a sum of item scores), and an average agency score (total agency score divided by the number of rated items). HighScope developed an online version of the assessment to assist with scoring and analysis, but it is not required.

Interpretability: The item scores range from 1 to 5, with higher scores indicating higher levels of program quality in terms of parent involvement and family services, staff qualifications and staff development, and program management.

Reliability:

(1) Internal consistency reliability: The developers conducted reliability testing of the first edition of the Infant-Toddler PQA Form B. Cronbach's alpha coefficient for the Infant-Toddler PQA Form B total score was 0.44. The alpha coefficients for the subscales were 0.13 for Program Management, 0.52 for Staff Qualifications and Staff Development, and 0.67 for Parent Involvement and Family Services.

The developers also conducted reliability testing of the third version of the Preschool PQA Form B. Cronbach's alpha coefficient for the Parent Involvement and Family Services subscale was 0.91 and 0.90, for the fall and spring Michigan Full-Day Preschool Comparison sample, respectively, and 0.74 for the MSRP sample. The developers did not report the alpha coefficient for the other two subscales because of insufficient sample sizes, and they did not report an alpha coefficient for the Form B total score.

The developers revised both the Infant-Toddler and Preschool Form B after this initial testing, and ultimately combined both forms into one, but details on the revisions were not provided in the documentation, and they did not conduct new reliability testing.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: No information available.

Validity:

(1) Content validity: The developers created the content of the subscales based on theory, research, and best practices in early care and education. They also consulted HighScope's infant-toddler curriculum manual, *Tender Care and Early Learning: Supporting Infants and Toddlers in Child Care Settings* (Post and Hohmann 2000; Post et al. 2011), and resources from early childhood professional organizations.

(2) Construct/concurrent validity:

Construct validity: No construct validity information was provided by the developers for the Infant-Toddler Form B. The developers did conduct confirmatory factor analysis (CFA) on the Preschool PQA Form B and found factor loadings from 0.57 to 0.75 for the Parent Involvement and Family

Services subscale (n = 141–150). The developers did not conduct CFA on the other two Form B subscales because of insufficient sample sizes.

Concurrent validity: No concurrent validity information was provided by the developers for the Infant-Toddler Form B. The developers found that the Parent Involvement and Family Services subscale of the Preschool PQA was correlated with the Teacher Beliefs Scale. As the developers expected, it was positively correlated with appropriate practices (0.28, $p \leq .10$, n = 41) and negatively correlated with inappropriate practices (-0.43, $p \leq .01$, n = 40).

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: As of October 2019, HighScope offers online reliability tests for Form A and Form B of the PQA. The training costs $75. Upon passing the online reliability test, HighScope certifies participants as PQA assessors.

Key Considerations for Early Care and Education (ECE): Designed for use in ECE with a sample of 96 infant and toddler group care settings and 272 preschool group care settings.

Previous Version: The PQA used to exist as two separate measures—one for infants and toddlers, and another version for preschool. Both versions were ultimately combined, though it is unclear when that occurred into one PQA that can be administered to infants, toddlers, and preschoolers. The Infant-Toddler PQA was first developed in 2001, Form B was later revised after initial internal consistency testing. The developers removed some program management items or combined them with other agency items. The Preschool PQA was first developed in 1988 and was revised five years later to improve administration procedures, required raters to score the items separately before assigning the total score, revised and consolidated items, and officially broke out the Preschool PQA into forms A and B.

References:

Epstein, A.S., S. Gainsley, M. Hohmann, T. Jurkiewicz, S. Lockhart, B. Marshall, and J. Montie. *Infant-Toddler Program Quality Assessment (PQA) Form B—Agency Items for Infant-Toddler and Preschool Programs.* Ypsilanti, MI: HighScope Press, 2013.

HighScope. *PQA Preschool Program Quality Assessment Administration Manual.* Second Edition. Ypsilanti, MI: HighScope Educational Research Foundation, 2003.

HighScope. *Preschool Program Quality Assessment—Revised (PQA-R) Manual*. Ypsilanti, MI: HighScope Educational Research Foundation, 2019a.

HighScope. "Program Quality Assessment." 2019b. Available at https://highscope.org/our-practice/pqa/. Accessed August 25, 2019.

Hohmann, M., S. Lockhart, and J. Montie. *Infant-Toddler Program Quality Assessment (PQA) Form A—Observation Items.* Ypsilanti, MI: HighScope Press, 2016.

Lockhart, S., A.S. Epstein, J. Claxton, and Z. Xiang. *Infant-Toddler Program Quality Assessment (PQA) Administration Manual.* Ypsilanti, MI: HighScope Press, 2014.

Program Sustainability Index (PSI), 2004

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Health

Content

Who leaders are (Participation in decision making)

What leaders do (Establish vision, manage efficient operations)

What leaders bring (Values, beliefs, attributes; education and experience)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

Center practices (Operational procedures and policies; regular assessment of program)

Administration characteristics

Respondent: Other (community-based program professionals)

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills with some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 53 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States

Setting: Community-based programs

Sample: 243 (Human development and family life professionals in local-, regional-, and national-level program development and evaluation roles)

Year of development: 2001

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

Availability

2 – Published source, contact developer(s) about permission requirements
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Jay A. Mancini and Lydia I. Marek

Narrative

Description: The Program Sustainability Index (PSI) is a self-administered, self-report survey. It is a measure of elements that support the longevity of the programs developed in the health field and administered at an annual meeting of the United States Department of Agriculture's Children, Youth and Families at Risk initiative. The PSI contains 53 items across seven subscales: Leadership Competence (7 items), Effective Collaboration (12 items), Understanding the Community (9 items), Demonstrating Program Results (7 items), Strategic Funding (5 items), Staff Involvement and Integration (10 items), and Program Responsivity (3 items). Leadership Competence and Strategic Funding measure aspects of "What leaders do." Three subscales contain items that provide information on "Who leaders are" (Effective Collaboration, Understanding the Community, and Staff Involvement and Integration). The Effective Collaboration subscale also measures the "Center culture, climate, and communication." Some subscales also provide information on "Center practices" (Understanding the Community and Demonstrating Program Results). One subscale (Program Responsivity) provides context information outside the leadership content identified by ExCELS; however, given the items are part of overall measure analyses we include information on this subscale in this profile. The developers' research on performance focused on six subscales (excluding Understanding the Community) and 29 items, but the developers recommend using the full set of items.

Uses of Information: The PSI was developed to be an assessment tool for program planning and implementation and for research on programs.

Methods of Scoring: Respondents reported on a list of project attributes using a 3-point scale: not at all (0), somewhat (1), and very much (2). Each of the PSI subscale scores is an average of items with a specific subscale.

Interpretability: A higher score on the subscale indicates greater extent that elements are in place that support program sustainability.

Reliability:

(1) Internal consistency reliability: Based on factor analyses (see Validity section), the developers reported the internal consistency for six subscales based on 29 of the 53 items: Leadership Competence (α = 0.81, 5 items), Effective Collaboration (α = 0.88, 10 items), Staff Involvement and Integration (α = 0.76, 4 items), Demonstrating Program Results (α = 0.85, 4 items), Strategic Funding (α = 0.76, 3 items), and Program Responsivity (α = 0.67, 3 items).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers drafted the items and seven subscales based on previous research to identify elements consistently contributing to sustainability. The developers conducted qualitative interviews with over 100 community program personnel (Mancini and Marek 1998) and over 4,000 program professionals (Betts et al. 2001). The qualitative results informed a survey on sustainability administered from 1999 to 2003 with 153 community-based programs.

(2) Construct/concurrent validity:

Construct validity: The developers used an exploratory factor analysis approach and identified six factors on a total of 29 out of the original 53 PSI items that capture all the sustainability elements in the conceptual model except the element of community understanding. Based on these analyses, items associated with the subscale on Understanding the Community did not load of any given factor and were excluded. Factor loadings for each subscale ranged as follows: 0.57 to 0.71 for Leadership Competence, 0.57 to 0.73 for Effective Collaboration, 0.73 to 0.80 for Demonstrating Program Results, 0.59 to 0.90 for Strategic Funding, 0.45 to 0.79 for Staff Involvement and Integration, and 0.49 to 0.85 for Program Responsivity. Subsequent confirmatory factor analysis showed acceptable model fit. The correlations between the factors range from 0.14 to 0.54 (Demonstrating Program Results and Leadership Competence). Ten of the 15 inter-factor correlations are 0.30 or above, with Leadership Competence correlated most highly with other factors.

Concurrent validity: The developers used bivariate correlations (Pearson's *r* and eta[28]—a measure of nonlinear association) to examine how the PSI subscale scores are related to middle-range program results reported by the respondents in terms of meeting the needs of at-risk families ($n = 224$), program sustainability planning ($n = 193$), and confidence in program survival in five years ($n = 223$). Pearson correlations were all low and ranged from 0.04 (Effective Collaboration and confidence in program survival) to 0.24 (Strategic Funding and program sustainability planning). Most of the correlations were significant at $p \leq .05$ or $p \leq .01$, with the exception of the Effective Collaboration subscale with program sustainability planning (0.08), and confidence in program survival (0.04); and Program Responsivity with program sustainability planning (0.05), and confidence in program survival (0.10). Most of the eta estimates were statistically significant ($p \leq .05$ or $p \leq .01$) and ranged from 0.17 (Program Responsivity subscale with meeting the needs of at-risk families, and confidence in program survival) to 0.33 (Leadership Competence and confidence in program survival). Six eta estimates were insignificant and ranged from 0.11 (Effective Collaboration with program sustainability planning) to 0.21 (Demonstrating Program Results with program sustainability planning). These analyses provide evidence of concurrent validity.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Individual items would need to be reworded to apply to ECE more generally, moving away from "project" language and more toward "program" or "center" language.

[28] The eta (η) only ranges from 0 to 1 (no negative values).

Previous Version: None.

References:

Mancini, J.A., and L.I. Marek. "Patterns of Project Survival and Organizational Support: The National Youth at Risk Program Sustainability Study." Virginia Cooperative Extension Publication 350–800. Blacksburg, VA: Virginia Polytechnic Institute and State University, 1998.

Mancini, J.A., and L.I. Marek. "Sustaining Community-Based Programs for Families: Conceptualization and Measurement." *Family Relations,* vol. 53, no. 4, 2004, pp. 339–347.

Betts, S.C., D.J. Peterson, M.S. Marczak, and L.S. Richmond. "System-Wide Evaluation: Taking the Pulse of a National Organization Serving Children, Youth, and Families at Risk." *Children's Services: Social Policy, Research, and Practice*, vol. 4, no. 2, 2001, pp. 87–101.

Relational Coordination Survey (RC Survey), 2018

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Multiple (management, health, early care and education [ECE])

Content

Center culture, climate, and communication
(Culture of respect, shared growth, and learning; collaboration among staff)

Administration characteristics

Respondent: Employees

Level of measure: Site, employee group, individual

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills with some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 3 (administered and/or scored by a highly trained individual)

Time/length: 7 items per type of role (see Narrative)

Administration interval: As frequently as desired

Languages available: English, other (see RC guidelines for details)

Technical information

Development sample

Locale: United States

Setting: Hospital

Sample: 9 orthopedics units in large urban hospitals; 666 eligible medical staff in six different roles, of which 338 (51%) responded to survey

Year of development: 2002

Measure performance*

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available (see Narrative)

*Version examined based on measure adjustments that do not appear to have been revalidated with final set of items.

Availability

4-Permission required, with costs
Material, training, and scoring costs: Unknown; contact publisher (Relational Coordination Analytics)

Developer(s)/publisher contacts

Developer(s): Jody Hoffer Gittell
Publisher: Relational Coordination Analytics (RCA); Relational Coordination Research Collaborative (RCRC)
RCA: 617-892-8653; RCRC: 781-736-3680
RCA: http://rcanalytic.com/; RCRC: https://heller.brandeis.edu/relational-coordination/index.html
Measure website: RCA: http://rcanalytic.com/rc-survey/; RCRC: https://heller.brandeis.edu/relational-coordination/survey/index.html

Narrative

Description: The Relational Coordination Survey (RC Survey) is a self-administered survey that asks employees to report on dimensions of how well they communicate and form professional relationships that support coordination with other employees in different roles within or across organizations. It can also measure relational coordination with clients. It was originally developed for studies of air travel and surgical care and has been used extensively throughout the broader management and health fields; it has begun to be used in early care and education. It measures seven dimensions: four about communication between employees (the extent to which it is frequent, accurate, timely, and focuses on problem solving) and three about relationships between employees (the extent to which there are shared goals, shared knowledge, and mutual respect). Each dimension only has one survey item, but the survey asks separately about communication and relationships with employees in each role involved in a focal work task in the organization. This means that before the RC Survey can be administered, the focal work task and the roles involved in the task must be defined and included in the text of the survey items. Accordingly, the total number of items on a given survey varies and equals seven times the number of roles involved. The time also varies; in one study when the RC Survey asked about 12 roles in an organization (but only for six of the dimensions), the typical completion time was 20 minutes (Gittell 2018, Section 4.7). Because the RC Survey focuses on dimensions of relational coordination between different groups of employees, it involves the "Center culture, climate, and communication" component of the ExCELS theory of change. Leaders can be included in the measure if they are listed as one or more of the roles in the survey, but as with other employee groups, the items address aspects of organizational culture and communication that are influenced by leaders.

Uses of Information: The RC Survey was originally developed to conduct research on relational coordination theory, including organizational conditions that predict relational coordination as well as its relationships with organizational outcomes, and it has been used extensively for this purpose. The RC guidelines (Gittell 2018) include results from a systematic review of 83 articles that studied relational coordination theory, 72 of which used a version of the RC Survey. The Relational Coordination Research Collaborative (RCRC) and its spinoff Relational Coordination Analytics (RCA), both founded by the developer, support researchers using the RC Survey as well as organizations that wish to use it as part of interventions designed to improve their work environment and performance. Although its use for development and improvement is less well-established, the RCRC offers training and coaching to organizations on using the RC Survey for this purpose. For both types of uses, if the survey is given at multiple points in time, scores from each time can be used to assess changes in relational coordination over time.

Methods of Scoring: Each item is scored on a scale of 1 to 5, with the corresponding response options varying by question. For example, for the items on timely, accurate, and problem-solving communication, the response options measure frequency: never (1), rarely (2), occasionally (3), often (4), and always (5). Other response scales focus on quantity. For shared goals and mutual respect, the response options are not at all (1), a little (2), somewhat (3), a lot (4), and completely (5). Shared knowledge uses similar response options of nothing (1), little (2), some (3), a lot (4), and everything (5). Finally, frequent communication uses a scale with relative response options of far too little (1), too little (2), just right (3), too much (4), and far too much (5); for scoring the items are recoded so that the highest score indicates the most-desired level of communication ("just right") in the order of far too little (1), far too much (2), too little (3), too much (4), just right (5).

Three types of scores can be calculated for an individual: (1) dimension scores can be averaged across roles providing a dimension-level score for that individual (for example, if there are five roles involved, the person's response on shared knowledge for each of the roles—including other employees in their role—would produce a shared knowledge score reflecting that person's perceptions of their shared knowledge with all of their relevant colleagues, based on 5 items); (2) role scores can be averaged across the dimensions providing an overall score for a role for that individual (for example, if one role is teachers, the person's response on all seven dimensions for teachers would produce a teacher score reflecting that person's perceptions of their overall relational coordination with teachers, based on seven items); and (3) total scores can be averaged across all dimensions and roles (for example, the person's responses to all seven dimensions for all five roles would produce a total score based on 35 items, reflecting that person's perceptions of their overall relational coordination with all of their relevant colleagues). All of these scores involve using equal weights for all items – that is, it is a simple average of the scores. The RC guidelines note that scores can alternatively be calculated using percentages of responses of 4 or 5 instead of an average, but that this is much less common.

The individual-level scores can also be aggregated to create group-, site-, or organization-level scores. The most common group-level scoring is to fill in a matrix showing the average score for each role's assessment of every other role. However, the developer notes that site- and organization-level scores should be weighted to account for missing responses, so each role's scores are weighted according to the group size within the organization. Otherwise, when roles have different propensities to engage in relational coordination and also have different response rates, site and organization scores will be biased.

For a study of patient care that used the RC Survey with different organizations (hospitals; Gittell et al. 2010), the developer calculated the intraclass correlation (ICC or the proportion of variance between hospitals). The ICC was 0.25, which is above the threshold for adequate group reliability.

Interpretability: Higher scores indicate greater levels of relational coordination among employees. The RC guidelines describe cut-points between the lowest, middle, and highest third of scores based on all scores collected by RCA through its online data collection system. These cut-points can be used to define weak, moderate, and strong levels of relational coordination, although they are relative terms based purely on existing data. The RC guidelines note there is not yet evidence that relational coordination has anything but a linear relationship with performance outcomes.[29] The role-based structure of the RC Survey means that patterns of stronger and weaker levels of relational coordination between (or within) particular roles can be analyzed through the matrix diagram of scores between role groups and a relational coordination network map. RCA provides score reports and can help organizations analyze and interpret their results.

Reliability:

(1) Internal consistency reliability: Primary validation of the RC Survey comes from additional analysis of a 2002 study of patient care in hospitals (Gittell et al. 2010). The Cronbach's alpha for the full scale was 0.86 (as discussed below under Validity, evidence supports the RC Survey as a single-factor scale). As recounted in the RC guidelines, a separate, earlier study of air travel found a Cronbach's alpha to be 0.80, but this involved a previous version of the RC Survey that did not include the item for accurate communication.

[29] An example of a nonlinear relationship would be if there is a threshold of relational coordination below which performance outcomes are poor regardless of the specific level of relational coordination.

The developer also calculated the hospital-level reliability of the scores, which was 0.81. Along with the ICC (0.25) that measured the proportion of variance between hospitals indicating adequate group reliability, this supports the use of site- and organization-level scoring of relational coordination.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The RC Survey was developed to align with relational coordination theory, which is reviewed in detail in the RC guidelines. A review of measures of teamwork in health care settings that included the RC Survey (Valentine et al. 2015) noted that it was "extensively tested through qualitative work," citing another publication by the developer.

(2) Construct/concurrent validity:

Construct validity: The primary validation study involved exploratory factor analysis, which supported a single-factor model. Factor loadings for the seven items ranged from 0.57 to 0.80. The earlier study of air travel also found evidence for a single-factor model; factor loadings for the six items in that version of the RC survey ranged from 0.54 to 0.72.

Concurrent validity: Studies using the RC Survey to assess patient care in hospitals, including the primary validation study, have found favorable associations between relational coordination as measured by the RC Survey and several patient outcomes, including higher quality of care (based on patient surveys) and reduced length of stay (based on hospital records), suggesting concurrent validity.

The RC guidelines include a systematic review by the developer of research on relational coordination. Most, although not all, articles included in the review used the RC Survey to measure relational coordination. They studied a variety of health and management settings and involved outcomes in several domains (efficiency and financial, quality and safety, client engagement, employee well-being, and learning and innovation). According to the developer, the large majority of findings covered in those studies were relationships between relational coordination and favorable outcomes in those domains (such as higher staff productivity, lower costs and higher profits, fewer customer complaints, improved patient well-being and satisfaction with care, higher family engagement, employee satisfaction and motivation, employee psychological safety, and knowledge creation), although a few findings demonstrated no relationships or relationships with unfavorable outcomes in those domains (often the same measures that were favorable in other studies, such as longer hospital stays, lower staff productivity, and lower patient quality and satisfaction).

(3) Predictive validity: It is not clear if any of the studies of relationships between relational coordination as measured by the RC Survey and outcomes involved outcomes that occurred later in time than when the RC Survey was conducted, though it is likely that length of hospital stay was obtained from records after the survey was conducted.

Bias Analysis: No information available.

Training Support: The RC guidelines provide information on administration and scoring. The RC Survey can be administered through RCA's online data collection system, which also offers analysis and reporting of results. The RC Survey can also be administered outside of RCA's system, either in person, by mail, or by email.

Key Considerations for Early Care and Education (ECE): The RC Survey is designed to be used in a wide range of settings. It has begun to be used in early care and education, although to date published research has involved conceptualizing relational coordination (Douglass and Gittell 2012) or measuring relational coordination via qualitative interviews in lieu of using the RC Survey (Douglass 2011).[30] As part of being used in any setting, the first step involves defining the work process around which relational coordination will be measured, and the roles involved, and inserted into the measure. In ECE, the work process could be the broad function of "improving infant and toddler child care" or a somewhat more specific aspect of care such as "meeting young children's social-emotional needs." Focusing on the center building, ECE roles might include center directors, teachers, assistant teachers, aides/floaters, coaches, specialists, and parents.

Previous Version: The RC Survey has had a few minor updates over time. The original version did not include the dimension for accurate communication. More recently, the item on frequent communication was updated to ask about how often employees communicate with the respondent (instead of the reverse) to make this item match the framing of the other items. The response options for the item on frequent communication were also updated from a 5-point frequency scale—never (1) to constantly (5). The updates to the frequent communication item occurred after the primary validation study was conducted.

Finally, the developer created a short form of a previous version of the RC Survey. The short form has fewer items and response options and is intended for respondents who have lower levels of education and are not native English speakers. The properties (reliability and validity) of the short form have not been examined to the same degree as the full measure.

References:

Douglass, A. "Improving Family Engagement: The Organizational Context and Its Influence on Partnering with Parents in Formal Child Care Settings." *Early Childhood Research & Practice*, vol. 13, no. 2, 2011.

Douglass, A., and J.H. Gittell. "Transforming Professionalism: Relational Bureaucracy and Parent-Teacher Partnerships in Child Care Settings." *Journal of Early Childhood Research*, vol. 10, no. 3, 2012, pp. 267–281.

Gittell, J.H.. "Relational Coordination: Guidelines for Theory, Measurement and Analysis." Waltham, MA: Relational Coordination Research Collaborative, revised August 2018.

Gittell, J.H., R. Seidner, and J. Wimbush. "A Relational Model of How High-Performance Work Systems Work." *Organization Science*, vol. 21, no. 2, March–April 2010, pp. 490–506. doi:10.1287/orsc.1090.0446.

Valentine, M.A., I.M. Nembhard, and A.C. Edmondson. "Measuring Teamwork in Health Care Settings: A Review of Survey Instruments." *Medical Care*, vol. 53, no. 4, April 2015, pp. e16–e30. doi:10.1097/MLR.0b013e31827feef6.

[30] The RC Survey has been used recently in a study of Early Head Start and child care partnerships; analyses are in progress (A. Douglass, personal communication, November 19, 2019).

Shared and Vertical Leadership Questionnaire (SVLQ), 2002

Purpose and context

Purpose: Development/implementation, research/evaluation

Field: Management

Content

What leaders do† (Promote quality practices, foster respect and learning)

What leaders bring (Values, beliefs, attributes)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

† Includes what staff as leaders do

Administration characteristics

Respondent: Manager, employees

Level of measure: Site/team

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Highly trained individual

Training for administration: Minimal 1–2 hours

Ease of administration and scoring: 3 (administered and/or scored by a highly trained individual)

Time/length: 70 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States (Mid-Atlantic area)

Setting: Manufacturing firm

Sample: 197 participants from 71 teams; 97.5% male, average age 49 years.

Year of development: 2002

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

***Validity*:**

-Construct/concurrent validity: Available

-Predictive validity: Available

Availability

2-Published source, contact developer(s) about permission requirements
Material, training and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Craig L. Pearce and Henry P. Sims Jr.

Narrative

Description: The Shared and Vertical Leadership Questionnaire (SVLQ) is a self-administered survey measuring perceptions of team leaders and team members in the work place. In practice, a researcher may need to be on site to manage administration. The SVLQ consists of two components: one on leadership behavior (70 items) and a second to subsequently assess outcomes of team effectiveness. Participants reported the leadership behavior for their team managers (vertical leadership) and their team members (shared leadership) using the same set of items based around five different subscales of leadership behavior for each: Aversive (6 items), Directive (6 items), Transactional (16 items), Transformational (20 items), and Empowering (22 items). All subscales capture information on "What leaders do." The Empowering subscale also measures aspects of "Center culture, climate, and communication," whereas the Transformational subscale also measures "What leaders bring." The developers collected data on the leadership behavior component first and on the team effectiveness component as a measure of later outcomes approximately six months after the data collection for the leadership behavior component.

Uses of Information: The developers created the SVLQ to help form recommendations for a manufacturing firm on what leadership areas the organization should continue to develop to increase the success of its teams. The developers used the leadership behavior subscales to examine the associations of team effectiveness scores with shared and vertical leadership styles.

Methods of Scoring: Items in the leadership behavior subscales are rated on a scale of 1 to 5, with the response options being definitely not true (1), not true (2), neither true nor untrue (3), true (4), and definitely true (5). Subscale scores are unit-weighted averages of the items in each of the factors. The developers also created composite scores for shared and vertical leadership, summing the subscales by reporter.

The developers reported the $r_{WG(J)}$ procedure (James et al. 1984) for inter-rater agreement within groups on the leadership subscales, which ranged from 0.84 (Transactional) to 0.91 (Aversive and Empowering) for shared leadership subscales and 0.86 (Transactional) to 0.94 (Empowering) for vertical leadership subscales.

Interpretability: Higher scores on the shared or vertical leadership subscale and composite scores indicate greater perceptions of leadership behavior exhibited by team members or team managers.

Reliability:

(1) Internal consistency reliability: Cronbach's alphas ranged from 0.72 (Transformational) to 0.87 (Transactional) for shared leadership subscales and from 0.77 (Aversive) to 0.87 (Directive and Transactional) for vertical leadership subscales.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers created the SVLQ based on an existing questionnaire, which they had done previous analysis on to identify five leader behavior strategies—aversive, directive, transactional, transformational, and empowering. The developers obtained additional source items through contacts with others conducting leadership research (such as House and Avolio), modifying items as needed to support asking items of different respondents. The developers also drafted new items.

(2) Construct/concurrent validity:

Construct validity: The developers conducted an exploratory factor analysis for shared and vertical leadership separately and identified five factors for leadership behavior, aligning to the five subscales. The factor loadings (in absolute value) ranged from 0.38 to 0.81 for shared leadership subscales (0.40 to 0.71 Transformational, 0.53 to 0.70 Directive, 0.42 to 0.81 Transactional, 0.38 to 0.70 Aversive, 0.39 to 0.68 Empowering). The factor loadings (in absolute value) ranged from 0.42 to 0.84 for vertical leadership subscales (0.46 to 0.76 Transformational, 0.56 to 0.70 Directive, 0.47 to 0.84 Transactional, 0.50 to 0.77 Aversive, 0.46 to 0.73 Empowering). The correlations ranged from 0.08 (Transformational and Aversive) to 0.67 (Empowering and Transformational) between the five subscales for shared leadership and from 0.20 (Transformational and Transactional) to 0.67 (Directive and Aversive) between the five subscales for vertical leadership. All but two of the correlations were significant at $p \leq .05$ or $p \leq .01$. The correlations between the same subscales for shared and vertical leadership (for example, shared aversive and vertical aversive) are highly correlated ($p \leq .01$), ranging from 0.78 (Transformational) to 0.90 (Transactional).

Concurrent validity: No convergent or divergent validity information was provided by the developers of this measure. The developers conducted discriminant analysis, comparing the 10 shared and vertical leadership subscales for low- and high-performing teams identified based on team effectiveness ratings. High-performing teams were higher than low-performing teams on 9 out of the 10 leadership subscale scores. Low-performing teams showed more vertical leadership than shared leadership compared to high-performing teams, whereas high-performing teams showed slightly more shared leadership than vertical leadership compared to low-performing teams.

(3) Predictive validity: The developers used multiple regression analysis to explore the associations of vertical and shared leadership subscales with the team effectiveness component (assessing overall effectiveness and output, quality, change, planning, interpersonal, and value). Team effectiveness was rated by three groups (and analyzed separately)—managers, team members, and customers. The developers found that both vertical and shared leadership were predictors of team effectiveness, but shared leadership was a more useful predictor relative to vertical leadership based on the percentage of variance explained in team effectiveness. With regard to specific types of leadership, vertical aversive leadership and shared aversive leadership were significantly negatively associated with team member reports of team effectiveness; vertical directive leadership and shared directive leadership were significantly negatively associated with manager reports of team effectiveness and internal customer reports of team effectiveness, respectively. In contrast, vertical transformational leadership and shared transformational leadership were significantly positively associated with manager and team member reports of team effectiveness.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): This measure was developed specifically for a research project with manufacturing, centered on looking at highly autonomous teams that possess a manager-worker style hierarchy. Adaptions may have to be made to terminology and roles to be used in an ECE setting.

Previous Version: None.

References:

Pearce, C.L., and H.P. Sims, Jr. "Vertical Versus Shared Leadership as Predictors of the Effectiveness of Change Management Teams: An Examination of Aversive, Directive, Transactional, Transformational, and Empowering Leader Behaviors." *Group Dynamics: Theory, Research, and Practice*, vol. 6, no. 2, 2002, pp. 172–197. doi:10.1037//1089-2699.6.2.172.

Supportive Environmental Quality Underlying Adult Learning (SEQUAL), 2019

Purpose and context

Purpose: Development/improvement, research/evaluation

Field: Early care and education (ECE)

Content

Who leaders are (Leadership roles, participation in decision making)

What leaders do† (Promote quality practices, foster respect and learning, manage efficient operations)

What leaders bring (Pedagogical knowledge)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

Center practices (Operational policies and procedures, regular assessment of children)

Center structures and staff supports (Training and professional development, collaborative planning time)

† Includes what teachers as leaders do

Administration characteristics

Respondent: Teaching staff (teacher-directors, lead or head teachers, assistant teachers, and aides)

Level of measure: Site

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 40 minutes, 134 items

Administration interval: Annual

Languages available: English, Spanish

Technical information

Development sample

Locale: United States (3 states and 4 counties)

Setting: ECE centers

Sample: 8 studies involving 1,280 teaching staff in diverse ECE settings (community-based, school-based, Head Start, pre-K) and backgrounds (race/ethnicity, languages spoken)

Year of development: 2012–2018

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

4-Permission required, with costs

Material, training, and scoring costs: Depend on project scope and publisher involvement; contact publisher for more information

Developer(s)/publisher contacts

Developer(s): Marcy Whitebook and Sharon Ryan

Publisher: Center for the Study of Child Care Employment, University of California, Berkeley
(510) 643-8293
https://cscce.berkeley.edu/

Measure website: https://cscce.berkeley.edu/topic/teacher-work-environments/sequal/

Narrative

Description: The Supportive Environmental Quality Underlying Adult Learning (SEQUAL) measure consists of a self-administered survey of teaching staff to assess their perceptions of their work environment and the supports and conditions that affect their practice. The SEQUAL is currently designed for use in center-based early care and education settings with any teaching staff who work with children from birth to age 5, including teacher-directors, lead or head teachers, assistant teachers, and aides. The publisher is currently adapting SEQUAL for family child care settings.

The SEQUAL has five subscales, each with multiple dimensions: (1) Teaching Supports (34 items), which covers curriculum, observations and assessments, materials, child and family supports, staffing, and professional responsibilities; (2) Learning Community (12 items), which asks about professional development; (3) Job Crafting (21 items), which addresses workplace decision-making, teamwork, and teaching staff input; (4) Adult Well-Being (39 items), which involves economic well-being, quality of work life, and wellness supports; and (5) Program Leadership (28 items), which covers perceptions of supervisors and program leaders. Each subscale also has a few related factual questions about topics such as work activities or job titles or open-ended reflections about the work environment and supports provided by programs. Aside from these factual and open-ended questions, the SEQUAL has a total of 134 items; the survey takes approximately 40 minutes to complete. In addition, teaching staff and administrators complete a short profile about their personal, professional, and job characteristics and, for administrators, about program characteristics. The SEQUAL can be administered as an online survey with publisher support, or as a paper version.

Each SEQUAL subscale covers several different topics and, therefore, aligns with multiple constructs in the ExCELS theory of change. (Relatedly, SEQUAL results are often presented by item instead of by subscale.) The Program Leadership subscale most directly assesses leadership; its items primarily cover "What leaders do" in the ExCELS theory of change. A couple of items also address elements of "What leaders bring" and "Who leaders are." Some items in the Job Crafting subscale also reflect "What leaders do" and "Who leaders are" from the perspective of teachers acting as leaders. Other items in Job Crafting and most items in Teacher Supports, Learning Community, and Adult Well-Being cover aspects of "Center culture, climate, and communication," "Center practices," and "Center structures and staff supports." Finally, the Teacher Supports and Adult Well-Being subscales also contain content that falls outside the ExCELS theory of change, such as items on other aspects of centers and on staff well-being outcomes.

Uses of Information: According to the publisher (Center for the Study of Child Care Employment 2014), the SEQUAL can be used for multiple purposes. Researchers can use it to study relationships between work environment and teaching staff and center characteristics (such as teacher education or center size), program improvement initiatives, and key outcomes (such as program and classroom quality, and children's learning). Directors, mentors, coaches, or others working with programs and teaching staff can use it to identify issues and guide improvements needed to the work environment and the policies, practices, and relationships that support teaching staff's growth and development. The publisher also suggests that the SEQUAL could be used by policymakers to inform decisions about early care and education policies and needs for funding and other resources.

Methods of Scoring: Each of the 134 items are scored on a 6-point agreement scale with the options strongly disagree (1), disagree (2), somewhat disagree (3), somewhat agree (4), agree (5), and strongly agree (6). Additional questions that collect respondent characteristics (for example, job titles, education) and open-ended questions are not scored but can be examined in connection with the rest of the SEQUAL results. The publisher notes that the SEQUAL is designed to provide site-level assessments of a program's work environment, based on aggregate information collected from teaching staff (Center for the Study of Child Care Employment 2019). The publisher does not recommend using or reporting individual teaching staff results, but it is currently exploring the use of the measure as an assessment of an individual's perceptions of the work environment. The publisher recommends reporting item-level results as frequencies, most commonly the percentage of teaching staff in the program that strongly agree or agree with an item. Subscale scores (and dimension scores within subscales) can be calculated using mean scores of items, for the purposes of studying associations between scores and teaching staff or program characteristics or other measures.

Interpretability: The publisher describes that SEQUAL results should be used to identify conditions and supports involving work environment that are currently in place and areas where additional supports or other approaches might be needed to improve the work environment. Higher subscale scores and items with larger percentages of agreement indicate stronger work environment conditions and supports; lower scores and smaller percentages indicate areas where additional supports could be needed.

Reliability:

(1) Internal consistency reliability: According to the publisher, Cronbach's alpha for the subscales in the SEQUAL range from 0.80 to 0.98. Across all studies, Program Leadership has the highest Cronbach's alpha scores (0.97 to 0.98) and Learning Community has the lowest scores (0.80 to 0.86), with the remaining scales in between: Teaching Supports (0.93 to 0.97), Job Crafting (0.91 to 0.94), and Adult Well-Being (0.92 to 0.94).

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers initially conducted a literature review and focus groups of teaching staff to better understand their perceptions of their work environments and how that influenced their teaching practice. The developers also asked expert reviewers to review the initial subscales and items. Pilot tests included assessing content validity, and the developers made several types of changes to address confusing and redundant items, including rewording, dropping, and reordering items, and moving items between subscales. In addition, the developers removed items that did not clearly apply to all types of ECE teachers and program settings.

(2) Construct/concurrent validity:

Construct validity: During the initial development of the SEQUAL, the developers conducted an exploratory factor analysis, which supported a five-factor model. Results from this analysis were used to revise the measure. The developers also conducted confirmatory factor analyses in studies using

more recent versions of the SEQUAL. According to the developers, each analysis has found a five-factor model has an acceptable fit. Additional analyses are underway.[31]

Concurrent validity: In several SEQUAL studies, the developers have examined associations of SEQUAL scores with measures of program and center quality (Quality Rating and Improvement System ratings) or with measures of observed classroom quality (the Classroom Assessment Scoring System or the Environment Rating Scales [ERS], including the Infant/Toddler Environment Rating Scale [ITERS]; the developers did not indicate which versions of these measures were used in these studies). Higher scores on the SEQUAL Adult Well-Being subscale was associated with high ratings on the CLASS Instructional Support domain. The SEQUAL Job Crafting subscale was associated with higher language and reasoning scores in the ERS, while the Learning Community subscale was also associated with a higher ERS activity score. Higher scores on the SEQUAL Learning Community and Leadership subscales predicted higher overall ERS scores. Higher scores on the SEQUAL Learning Community and Teaching Supports subscales were also associated with higher scores on the ITERS Interaction subscale. These analyses showed evidence of concurrent validity.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: The publisher can support programs by administering the SEQUAL using the publisher's online survey software, preparing a report with the results, and helping programs analyze the results. The SEQUAL can also be administered separately in a hard-copy format.

Key Considerations for Early Care and Education (ECE): Designed for use in ECE with 1,280 teaching staff in diverse ECE settings (community-based, school-based, Head Start, pre-K) and backgrounds (race/ethnicity, languages spoken).

Previous Version: The SEQUAL was most recently updated in 2018; a previous version was updated in 2015, and the initial version was developed in 2012. Updates have incorporated results from validity and other analyses conducted by the developers. The developers have reworded, reordered, moved, and dropped items.

References:

Center for the Study of Child Care Employment. "Supportive Environmental Quality Underlying Adult Learning (SEQUAL): Purpose and Context." Preliminary technical information provided to Mathematica Early Care and Education Leadership Study team, Sept. 16, 2019.

Center for the Study of Child Care Employment. "Supportive Environmental Quality Underlying Adult Learning (SEQUAL): A Tool for Program Improvement." Berkeley, CA: Center for the Study of Child Care Employment, University of California, Berkeley, 2014. Available at https://cscce.berkeley.edu/wp-content/uploads/2014/SEQUAL-Overview.pdf. Accessed May 22, 2019.

Whitebook, M., and S. Ryan. "SEQUAL 2.1 (Supportive Environmental Quality Underlying Adult Learning) Assessment." Berkeley, CA: Center for the Study of Child Care Employment, University of California, Berkeley, 2018.

[31] For additional information, readers may contact the publisher at cscceinfo@berkeley.edu.

Survey of Transformational Leadership (STL), 2010

Purpose and context

Purpose: Development/improvement, monitoring, research/evaluation

Field: Health

Content

What leaders do (Promote quality practices, foster respect and learning, establish vision, promote family/community partnerships)

What leaders bring (Personal development or critical-thinking knowledge and skills; interpersonal and team-building knowledge and skills; advocacy and community-building skills; values, beliefs, attributes)

Administration characteristics

Respondent: Employees

Level of measure: Individual

Data sources: Survey (mode–self-administered, report level–report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 2 (self-administered or administered and scored by someone with basic clerical skills)

Time/length: 30 minutes, 96 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States (Northwest, Gulf Coast, Southeast, Great Lakes regions)

Setting: Other (outpatient substance use treatment programs)

Sample: Counselors involved in the Treatment Costs and Organizational Monitoring project; 57 programs; 213 staff and 57 leaders; most participants were female, White, college educated, worked at least three years in the field and one year in current position; staff averaged 39 years of age and leaders averaged 48 years of age

Year of development: 2008

Measure performance

***Reliability*:** 3 (meets minimum acceptability ratings—0.70)

***Validity*:**

-Construct/concurrent validity: Available

-Predictive validity: Not available

Availability

2-Published source, contact developer(s)/publisher about permission requirements[32]
Material, training, and scoring costs: No known costs

Developer(s)/publisher contacts

Developer(s): Jennifer R. Edwards, Danica K. Knight, Kirk M. Broome, and Patrick M. Flynn
Publisher: Institute of Behavioral Research
Texas Christian University
TCU Box 298740
Fort Worth, TX 76129
irb@tcu.edu
(817) 257-7226
Measure website: https://ibr.tcu.edu/forms/organizational-staff-assessments/

[32] The developers grant permission to use the STL for nonprofit educational and nonprofit library purpose, following the developers' guidelines. Express written permission is required to use the STL for commercial purposes.

Narrative

Description: The Survey of Transformational Leadership (STL) was developed to measure transformational practices within substance use treatment organizations. The survey is a self-administered report of clinical directors by program staff such as clinicians and counselors. It includes a total of 96 items and takes 30 minutes to complete. The measure includes five subscales (with 83 scored items), referred to as components, with each subscale containing items from one to two different conceptual themes (a total of nine themes, listed in parentheses): Idealized Influence (integrity and sensible risk), Intellectual Stimulation (encourages innovation and demonstrates innovation), Inspirational Motivation (inspirational motivation), Individualized Consideration (develops others and respects others), and Empowerment (task delegation and expects excellence). The integrity and inspirational motivation conceptual themes have items across the "What leaders bring" and "What leaders do" content areas. The conceptual themes of sensible risk and expects excellence have items within the "What leaders bring" content area only, whereas respects others, develops others, task delegation, encourages innovation, and demonstrates innovation focus on "What leaders do" content only. The survey also includes 12 additional items generally relating to staff performance and the "What leaders do" content area.

Uses of Information: The developers state the STL can be used to develop or improve transformational leadership practices and strategies. It can be used for monitoring leader progress toward set goals. It can also be used for research or evaluation of the transformational leadership construct.

Methods of Scoring: Respondents are asked to rate various statements about leadership practices on a 5-point frequency scale ranging from not at all (0); once in a while (1); sometimes (2); fairly often (3); and frequently, if not always (4). Average scores can be calculated for each theme and subscale by adding the scores for each item within a theme or subscale and dividing that value by the number of items within the theme or subscale. That average score is then multiplied by 10, so final scores range from 0 to 40. The developers discuss how the STL can be used to measure transformational leadership globally. The developers calculate an average transformational leadership score, but they do not specify whether to average the items or the themes to calculate the score.

Interpretability: Generally, the higher the rating score, the more frequently the leader is perceived to fit the leadership style or practice described by the statement, or to possess the transformational leadership qualities identified in the conceptual themes and subscales.

Reliability:

(1) Internal consistency reliability: The Cronbach's alpha coefficients were calculated for each conceptual theme: integrity (0.95), sensible risk (0.89), encourages innovation (0.92), demonstrates innovation (0.86), inspirational motivation (0.97), develops others (0.89), supports others (0.78), task delegation (0.89), and expects excellence (0.95). The alpha coefficient for the global measure of transformational leadership was 0.96.

(2) Test-retest reliability: No information available.

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developers conducted three focus groups with counselors and directors of two Gulf Coast Addiction Technology Training Centers to evaluate item wording and overall utility of the measure. The focus groups helped the developers identify the most appropriate person to be rated, make minor revisions to items, and add new items.

(2) Construct/concurrent validity:

Construct validity: The developers conducted exploratory factor analyses within each of the five subscales and found nine first-order leadership factors pertaining to the conceptual themes, with item factor loadings ranging from 0.47 to 0.88. Second-order maximum likelihood factor analysis of the conceptual themes identified two factors within each of the subscales, except for inspirational motivation, which only had one factor. The conceptual theme factor loadings were all significant ($p \leq .001$) and ranged from 0.67 (expects excellence) to 0.98 (task delegation). The intercorrelations ranged from 0.40 (task delegation and develops others) to 0.93 (task delegation and inspirational motivation) indicating that the STL can be used to measure a higher order construct (transformational leadership).

Concurrent validity: The developers used four scales from the Multifactor Leadership Questionnaire (MLQ 5X), two scales from the Attributes of Leader Behavior Questionnaire (ALBQ), and six items measuring job satisfaction from the Survey of Organizational Functioning (SOF) as the criterion measures to validate the STL. The correlations between seven of the STL's conceptual themes most aligned with the four MLQ subscales were significant ($p \leq .001$) and ranged from 0.74 (supports others from the STL and individualized consideration from the MLQ) to 0.88 (inspirational motivation from the STL and inspirational motivation from the MLQ). The global measure of STL was correlated with the global measure of the MLQ at 0.95. The correlations between the other two STL conceptual themes with two ALBQ subscales were significant ($p \leq .001$) and ranged from 0.50 (expects excellence from the STL and assures competency from the ALBQ) to 0.86 (task delegation from the STL and opportunities for success from the ALBQ). The strong correlations between the subscales of the STL, MLQ, and ALBQ suggest convergent validity.

The developers also studied whether the STL conceptual themes were related to job satisfaction as measured through the SOF. The developers found that the mean scores for the conceptual themes varied by job satisfaction (by developing a dichotomous variable based on the median-split for job satisfaction).

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Most of the STL items use general terminology and would not need to be adapted for the ECE setting. Items reference the program and staff, which are terms used in ECE settings. The developers noted that it could be relevant to use in other service sector settings because the workplace practices and attitudes described may be similar in these settings.

Previous Version: None.

References:

Edwards, J.R., D.K. Knight, K.M. Broome, and P.M. Flynn. "The Development and Validation of a Transformational Leadership Survey of Substance Use Treatment Programs." *Substance Use and Misuse,* vol. 45, no. 9, July 2010, pp. 1279–1302. doi: 10.3109/10826081003682834

Institute of Behavioral Research. "Survey of Transformational Leadership: Program Staff Version (TCU STL-S)." Fort Worth, TX: Texas Christian University, Institute of Behavioral Research, 2009. Available at http://ibr.tcu.edu/wp-content/uploads/2013/06/tcom-STL-S.pdf. Accessed October 18, 2019.

Institute of Behavioral Research. "Survey of Transformational Leadership: Program Staff Version Scales and Item Scoring Guide." Fort Worth, TX: Texas Christian University, Institute of Behavioral Research, 2009. Available at http://ibr.tcu.edu/wp-content/uploads/2013/06/tcom-STL-S-sg.pdf. Accessed October 18, 2019.

Tripod Teacher Survey, 2014

Purpose and context

Purpose: Development/improvement, monitoring, research/evaluation

Field: K–12 education

Content

What leaders do (Promote quality practices, foster respect and learning, establish vision, promote family/community partnerships, manage efficient operations)

What leaders bring (Interpersonal and team-building knowledge and skills; administrative, business, and management knowledge and skills)

Center culture, climate, and communication (Culture of respect, shared growth, and learning; collaboration among staff)

Center practices (Operational procedures and policies; regular assessment of program, classroom, and children)

Center structures and staff supports (Training and professional development, accountability structures)

Administration characteristics

Respondent: Teachers

Level of measure: Site

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not specified

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills and some training

Training for administration: Self-training < 1 hour

Ease of administration and scoring: 5 (administered or scored by developer)

Time/length: 50 items

Administration interval: None described

Languages available: English

Technical information

Development sample

Locale: United States

Setting: School

Sample: 3,769 teachers in 294 schools across 25 school districts

Year of development: 2012–2014

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not available

Availability

4-Permission required, with costs
Material, training, and scoring costs: Contact publisher for information on support with administration and reporting results.

Developer(s)/publisher contacts

Developer(s): Ronald F. Ferguson
Publisher: Tripod Education Partners
101 Main Street, 14th Floor
Cambridge, MA 02142
info@tripoded.com
www.tripoded.com

Narrative

Description: The Tripod Teacher Survey was developed to measure a school's professional environment by asking teachers about their experiences as members of their school community. The survey includes teachers' self-report items and report on the principal and instructional leaders. It consists of 50 total items across six subscales: School Leadership, Schoolwide Academic Press, which the developers define as "the extent to which a school sets high standards for instructional quality and teachers believe they can and actually do collaborate to meet those standards" (Tripod Education Partners April 2019a, p.5), Organizational Effectiveness, Professional Development Quality, Quality of Professional Learning Community (PLC) Time Use, and Evaluation Quality. The School Leadership subscale includes content on "What leaders do" and "What leaders bring." The Schoolwide Academic Press subscale includes information on "Center culture, climate, and communication." The Quality of PLC Time Use and Observation Frequency subscales have content from the "Center practices" content area. The Professional Development Quality, Evaluation Quality, and Forms of Professional Support subscales have items relating to "Center structures and staff supports."

Uses of Information: The developer indicates that the Tripod Teacher Survey can be used by school administrators or district leaders to inform school development or improvement efforts. It can be used for monitoring purposes as two subscales directly relate to leadership accountability. It can also be used for research or evaluation purposes as the developer describes measuring associations with a site's professional environment.

Methods of Scoring: The Tripod Teacher Survey uses three 5-point scales. The most common scale includes: totally untrue (1), mostly untrue (2), somewhat (3), mostly true (4), and always true (5). The survey also includes items with a response scale that quantifies: none (1), slight (2), some (3), a lot (4), and N/A (5); and a frequency scale: seldom or never (1), several times this year (2), monthly (3), bi-weekly (4), and every week (5). A school-level average can be calculated for each item. Then, the school-level average for each item within a subscale can be averaged to determine the school's average for that subscale. The averages for the School Leadership subscale and Schoolwide Academic Press subscale can be averaged to calculate the school's organizational effectiveness score.

The developer also estimated the school-level reliability using a multilevel approach, which ranged from 0.45 (Evaluation Quality) to 0.69 (Schoolwide Academic Press) for the subscales. The developer noted that the low school-level reliability estimates for the Professional Development Quality, Quality of PLC Time Use, and Evaluation Quality subscales are likely due to the small number of teachers within a school and the limited number of items in the scales and caution calculating scores for these subscales for summative purposes unless large sample sizes are available. The developer recommends that at least 10 teachers are required for using the School Leadership and Schoolwide Academic Press scores for monitoring or accountability purposes.

Interpretability: Higher scores on the composite and subscales indicate a more positive professional environment in schools.

Reliability:

(1) Internal consistency reliability: The developer provided Cronbach's alpha coefficients for the subscales: 0.71 (Evaluation Quality), 0.84 (Quality of PLC Time Use), 0.96 (Schoolwide Academic Press), 0.97 (Professional Development Quality), and 0.98 (School Leadership).

(2) Test-retest reliability: The developer conducted test-retest reliability for school-level scores across consecutive years. Between year 1 and year 2, they found significant correlations ranging from 0.36 (Professional Development Quality) to 0.81 (Schoolwide Academic Press) (n = 48), and between year 2 and year 3, they found correlations ranging from 0.12 (Quality of PLC Time Use) to 0.61 (School Leadership) (n = 27) for the six subscales. The year 2 and year 3 correlations were significant with the exception of Quality of PLC Time Use. School Leadership (0.56 in year 1 to year 2, and 0.61 in year 2 to year 3) and Evaluation Quality (0.81 in year 1 to year 2, and 0.51 in year 2 to year 3) were stable across the years.[33]

(3) Alternate form reliability: No alternate form.

(4) Inter-rater reliability: Not applicable.

Validity:

(1) Content validity: The developer conducted a review of the literature on school, principal, and teacher effectiveness to inform survey development.

(2) Construct/concurrent validity:

Construct validity: The developer conducted multilevel confirmatory factor analysis (CFA) to determine whether the School Leadership and Schoolwide Academic Press subscales were empirically distinguishable. The results showed a significant Spearman rank correlation of r = 0.85 between the two subscales, supporting the use of a composite organizational effectiveness measure that includes these two subscales. The factor loadings ranged from 0.88 to 0.99 for School Leadership and 0.75 to 0.99 for Schoolwide Academic Press. The developer also conducted CFA on the Professional Development Quality, Quality of PLC Time Use, and Evaluation Quality subscales and found correlations ranging from –0.03 (Professional Development Quality and Quality of PLC Time Use; though not significant) to 0.45 (Professional Development Quality and Evaluation Quality; and Quality of PLC Time Use and Evaluation Quality), suggesting that these subscales do not form a composite. The factor loadings ranged from 0.93 to 1.00 for Professional Development Quality, 0.73 to 0.86 for Quality of PLC Time Use, and 0.59 to 1.00 for Evaluation Quality.[34] Both models had acceptable fit with most fit statistics (with the exception of the standardized root mean square residual).

Concurrent validity: The developer described a study conducted by Liu et al. (2014), which found that the Schoolwide Academic Press subscale significantly correlated with student perceptions of teaching effectiveness (r = 0.39) measured by Tripod's 7Cs student survey, and that the Schoolwide Academic Press subscale and School Leadership subscale significantly correlated with teacher

[33] Strong stability across years would not be expected, particularly given potential changes in staff, professional development opportunities, and curriculum.

[34] Per personal communication with the publisher, the factor loading of 1.0 was a result of the residual variance being fixed to 0 for the items that had a small negative residual variance in a prior model (residual variance is the error in an item that is not explained by the factor).

perceptions of their own effectiveness (r = 0.42 and 0.22). The Tripod's 7Cs student survey collects information regarding classroom teaching practices and student peer supports. Another study by Liu et al. (2014) found the School Leadership subscale and Cultural Press for Excellence (a subcomponent of the school-wide academic press subscale) were significantly positively associated with school-level value-added achievement gains in reading and math. Liu et al. also found in a third study that the Schoolwide Academic Press subscale was significantly positively associated with school-level proficiency rates. These studies provided evidence of concurrent validity.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: No information available.

Training Support: Tripod's website notes the developers can provide support with survey administration and reporting results.

Key Considerations for Early Care and Education (ECE): Some of the terminology within the 50 items will need to be adapted to be relevant to the ECE setting. For example, one set of items in the Tripod Teacher Survey references principals and schools. Such items would likely need to be revised to refer to administrator or director and centers.

Previous Version: The developer notes that the original Tripod Teacher Survey was created in 2001, and the current version was developed through data collection between spring 2012 and spring 2014, but no additional information on the original survey is provided.

References:

Tripod Education Partners. *Tripod Teacher Survey Technical Manual.* April 2019a.

Tripod Education Partners. "Survey Assessments." 2019b. Available at https://tripoded.com/surveys. Accessed August 13, 2019.

Vanderbilt Assessment of Leadership in Education (VAL-ED), 2009

Purpose and context

Purpose: Development/improvement, monitoring, research/evaluation

Field: K–12 education

Content

What leaders do (Promote quality practices, foster respect and learning, promote family/community partnerships, manage effective operations)

Administration characteristics

Respondent: Principal, teachers, other (principal's supervisor)

Level of measure: Site, individual

Data sources: Survey (mode–self-administered, report level–self-report and report of others)

Usability

Technology: Not required

Staff level of qualifications

Personnel for administration: Individual with basic clerical skills with some training

Training for administration: Basic self-training < 1 hour

Ease of administration and scoring: 5 (administered or scored by publisher)

Time/length: 30 to 45 min, 72 items

Administration interval: Annual

Languages available: English

Technical information

Development sample

Locale: United States

Setting: School

Sample: 235 principals, 253 supervisors, and 8,863 teachers from a sample of 309 schools—however, only 218 schools had complete participation (data gathered from all three response groups); 39% elementary schools, 32% middle schools, and 28% high schools; 23% from the West, 30% from the South, 22% from the Midwest, and 25% from the Northeast; 39% urban, 39% suburban, and 22% rural

Year of development: 2008

Measure performance

Reliability: 3 (meets minimum acceptability ratings—0.70)

Validity:

-Construct/concurrent validity: Available

-Predictive validity: Not Available

Availability

4-Permission required, with costs
Material, training, and scoring costs: No information available

Developer(s)/publisher contacts

Developer(s): Stephen N. Elliott, Ellen Goldring, Joseph Murphy, and Andrew Porter
Publisher: Resonant Education
301 Scott Ave.
Nashville, TN 37206
(877) 212-6458
Measure website: https://resonanteducation.com/valed/

Narrative[35]

Description: The Vanderbilt Assessment of Leadership in Education (VAL-ED) is a paper or web survey to assess leadership behaviors of school principals. It provides 360 degree feedback by including a self-report survey for principals to complete and similar surveys for teachers and the principal's supervisor to complete. The term "*teachers"* is used broadly by the VAL-ED developers to include school staff such as teachers, teacher aides, librarians, and counselors. Individuals rating principals should know the principal and have worked with him or her for at least two months. The VAL-ED should be administered by an objective evaluation coordinator, and the principal should not be present. It includes 72 items and can be completed in 30 to 45 minutes, but administration time is noted as one hour to include time for instructions. The VAL-ED is available as two parallel forms referred to as Form A and Form C, developed so the VAL-ED could be administered in consecutive years without respondents receiving the same items twice. The VAL-ED is intended to measure principal behaviors that directly influence teacher performance and, in turn, student achievement through the intersection of six core components and six key processes, resulting in a matrix of 36 cells. The core components relate to what principals must accomplish to improve student learning, and include six subscales: (1) High Standards for Student Learning, (2) Rigorous Curriculum (content), (3) Quality Instruction (pedagogy), (4) Culture of Learning and Professional Behavior, (5) Connections to External Communities, and (6) Performance Accountability. The key processes relate to how those components are achieved by the principal, and include six subscales: (1) Planning, (2) Implementing, (3) Supporting, (4) Advocating, (5) Communicating, and (6) Monitoring. Each cell of the matrix includes 2 items, resulting in 12 items in each core component or key process subscale. Developers refer to the grouping of two items in each cell as a *core component key process cluster*. The items in the VAL-ED relate to the "What leaders do" content area.

Uses of Information: The VAL-ED can be used for development or improvement in the form of performance feedback, professional development planning, and monitoring progress toward a goal. It can also be used for research or evaluation purposes.

Methods of Scoring: For each item in the VAL-ED, respondents are asked to indicate how effective the principal is at that particular action on a 5-point effectiveness scale with: ineffective (1), minimally effective (2), satisfactorily effective (3), highly effective (4), and outstandingly effective (5). Before marking the effectiveness scale rating, respondents first identify the source of evidence they are using to rate the principal for that item, choosing between five sources—reports from others, personal observations, school documents, school projects or activities, and other sources—or no evidence. Respondents are instructed that if they select "no evidence" for any item, they must also score that item as ineffective (1). All respondents (principals, teachers, supervisors) complete the same items of the VAL-ED with one exception—only the teacher and supervisor forms have the option of "don't know" for the effectiveness rating. The VAL-ED is scored by computer by the publisher. A mean score can be calculated for each subscale, and a total mean effectiveness score can be calculated using the mean subscale scores for each respondent group and using the equal-weight averages across the respondent groups. It is also possible to calculate the percentage of times respondents selected each type of evidence source.

Interpretability: Generally, the higher the effectiveness rating, the more a principal is considered to effectively exhibit the leadership behaviors identified in the subscales. Additionally, the developers viewed having a variety of "portfolios" of evidence as positive. The developers created norms for the subscales and total scores by respondent type and for the combination of all respondent groups based on

[35] The information summarized in this profile was taken from Elliott et al. 2009 unless otherwise noted.

the 2008 field test. The norms are based on a sample of elementary, middle, and high school principals across urban, suburban, and rural regions of the United States. Once mean scores are calculated, they can be translated to percentile rankings using these norms. The developers also defined four performance levels based on cut scores that are set on the overall equal-weight average across the three respondent groups: below basic, basic, proficient, and distinguished.

A principal report is developed for each individual principal. The principal report summarizes information about who participated in the principal's assessment by number of respondents and type and the evidence used to rate the principal's effectiveness. It shows how effective the principal is rated overall and for each of the 12 subscales (mean score, performance level, and percentile rank) across the three respondent groups, including variations in responses across the respondent groups. Finally, the report includes a matrix of strengths and areas for improvement across the subscales based on the performance levels, and a list of the six lowest-rated *core component key process clusters* (for example, the Connections to External Communities x Advocating cluster).

Reliability:

(1) Internal consistency reliability: The developers reported alpha coefficients for the total effectiveness score and for each of the 12 subscales (Porter et al. 2010) across respondent groups and for Form A and Form C. The total score coefficient was similar across the forms and respondent groups from 0.98 (principal Form A and Form C) to 0.99 (supervisor and teachers, Forms A and B). The subscale coefficients varied across the forms and respondent groups 0.87 (Advocating and Communicating subscales within the principal Form A) to 0.97 (Performance Accountability subscale within the supervisor Form A, and teacher Forms A and C; and the External Community subscale within the teachers Form A).

(2) Test-retest reliability: Covay Minor et al. (2017) describe a study the developers of the VAL-ED conducted to determine test-retest reliability. The VAL-ED was administered to 71 elementary and secondary schools in seven school districts across the United States. It was administered at two points in time within a period of 2 to 29 weeks. All the correlations were significant ($p \leq .001$). The correlation between the total effectiveness score between the two time periods was 0.64 for principal self-reports (n = 35), and the subscale correlations between the two time periods ranged from 0.55 (External Community) to 0.71 (High Standards). For teacher reports, the total effectiveness score correlation was 0.91 (n = 71), and the correlations for the subscales ranged from 0.87 (External Community) to 0.92 (Advocating). The developers did not conduct this analysis on supervisor respondents because of too few respondents. The developers also studied the mean differences between the two administrations of the VAL-ED. They found that principal self-reported scores were higher in time 2 than time 1 (n = 35), with an effect size ranging from 0.25 (Quality Instruction) to 0.52 (Planning). The effect size ranged from 0.07 (External Community) to 0.21 (Communicating) for teacher-reported scores for difference in scores between time 1 and 2 (n = 71).

(3) Alternate form reliability: Mean effectiveness rating scores, overall, and by respondent group varied by no more than 0.04 points between Form A and Form C, indicating that both forms were operating as parallels. The mean subscale scores varied by no more than 0.22 points between the two forms for principals and teachers (n = 106 [Form A], n = 130 [Form C] principal; n = 113 [Form A], n = 132 [Form C] teacher) whereas the difference between the means of the two forms was less than 0.15 for supervisors (n = 124 [Form A], n = 130 [Form C]). (Porter et al. 2008)

(4) Inter-rater reliability: See below.

(5) Generalizability study:[36] The developers conducted a generalizability study to examine the variance in scores that could be attributed to the component versus process subscales and the different respondent types (Porter et al. 2008). Most of the core components subscales demonstrated unique variance as a subscale from other information, with two exceptions: (1) the High Standards for Student Learning subscale was not significantly differentiated on Form A for principals or supervisors and (2) the Quality Instruction subscale did not differentiate on Form C for supervisors. In terms of the six key processes subscales, only Supporting and Advocating demonstrated differentiation (or unique variance) from subscales and the overall score across all respondent groups and for both forms. The Planning and Implementing subscales were not well differentiated on all forms for principal and supervisor respondents. The Communicating and Monitoring subscales were not well differentiated for supervisors using Form A. The technical manual (Table 4.13) presents which subscales, alternate forms, and respondents provide unique variance as compared to the overall score.

Validity:

(1) Content validity: The developers first conducted a review of the literature on school leadership effects on student achievement and developed a conceptual framework before writing the items in the VAL-ED measure. The developers then conducted a sorting study and two rounds of cognitive interviews to refine the measure to include 108 items. After a pilot test was conducted in nine schools (three elementary schools, three middle schools, and three high schools) in an urban district in the Midwest, the developers shortened the measure to 72 items and revised the benchmarks of the effectiveness scale. A third round of cognitive interviews helped the developers test the web prototype of the measure. Results from the interviews indicated that respondents had common understanding of the revised items and no difficulties with the web version of the measure. The developers conducted one final pilot study of the 72-item measure in 11 schools across four districts in the Midwest (Porter et al. 2008).

(2) Construct/concurrent validity:

Construct validity: The developers conducted exploratory factor analysis and tried a 6-factor, 8-factor, and 12-factor solution for Forms A and C. The results provided some initial support for the conceptual framework of the VAL-ED. The developers conducted confirmatory factor analysis on a core components model and a key processes model using data aggregated across the three respondent groups and higher-order factors for the subscales and overall score. Both models had acceptable fit across Forms A and C. The core components model had item factor loadings ranging from 0.63 to 0.98, whereas the key processes model had factor loadings within the range of 0.62 to 0.95 (Porter et al. 2010).

The developers examined the intercorrelations among subscales. The intercorrelations ranged from 0.78 (External Communities and High Standards; External Communities and Rigorous Curriculum) to 0.95 (Supporting and Quality Instruction; Supporting and Communicating) for supervisor, and 0.88 (External Communities and Rigorous Curriculum; External Communities and Quality Instruction) to 0.97 (High Standards and Planning; High Standards and Implementing; High Standards and Communicating; Communicating and Performance Accountability; Implementing and Planning; Implementing and Supporting) for teachers. The developers did not report intercorrelations for principal data (Porter et al. 2008).

Concurrent validity: No concurrent validity information was provided by the developers of this measure.

[36] This section is only included in profiles for measures that conducted a generalizability study.

(3) Predictive validity: No predictive validity information was provided by the developers of this measure.

Bias Analysis: The developers conducted bias analysis as they were developing the measure. Nine testing and rating scale experts were invited to participate in the panel. They were given an electronic version of the survey and interviewed after completing the survey. They conducted a fairness review of Forms A and C based on the Educational Testing Service (2000) test fairness guidelines. The results of the fairness review showed 27 total items that raised a fairness concern—13 items in Form A and 14 items in Form C. Four of these 27 items were identified as being of "serious concern" by the panelists. The developers discussed the concerns with the panelists and made appropriate revisions to those four items (Porter et al. 2008).

The developers conducted differential item functioning analysis and found that five items on Form C exhibited urbanicity differences, with responses varying between urban-rural respondents or urban-suburban respondents (Porter et al. 2010). The developers provided information about the magnitude of the differences (0.22 to 0.46 standard deviations), but no information about the direction. Four of these items were later revised based on bias panel recommendations.

Training Support: No information available.

Key Considerations for Early Care and Education (ECE): Some VAL-ED items would need to be tweaked to be applicable for ECE by changing terms such as faculty, school, and students to staff, center, and children.

Previous Version: None.

References:

Covay Minor, E., A.C. Porter, J. Murphy, E. Goldring, and S.N. Elliott. "A Test-Retest Analysis of the Vanderbilt Assessment for Leadership in Education in the USA." *Educational Assessment, Evaluation and Accountability,* vol. 29, 2017, pp. 211–224. doi:/10.1007/s11092-016-9254-9.

Elliott, S.N., E. Goldring, J. Murphy, and A.C. Porter. "VAL-ED Handbook Implementation and Interpretation. Vanderbilt Assessment of Leadership in Education." Charlotte, NC: Discovery Education, July 2009.

Porter, A.C., J. Murphy, E. Goldring, S.N. Elliott, M.S. Polikoff, and H.M. Vanderbilt. *Vanderbilt Assessment of Leadership in Education Technical Manual.* Version 1.0. Nashville, TN: Learning Sciences Institute, Vanderbilt University, 2008.

Porter, A.C., M.S. Polikoff, E.B. Goldring, J. Murphy, S.N. Elliott, and H. May. "Investigating the Validity and Reliability of the Vanderbilt Assessment of Leadership in Education." *The Elementary School Journal*, vol. 111, no. 2, 2010, pp. 282–313.

Appendix A

Glossary of Terms

This page has been left blank for double-sided copying.

Glossary of terms

Alternate form. Two or more versions of one measure that are considered interchangeable because they purportedly measure the same constructs in the same ways. The alternate forms are intended for the same purpose and administered with the same directions. Alternate forms is a generic term used to describe measures in any of three categories: (1) parallel forms have equal raw score means, standard deviations, error structures, and correlations with other measures for any given population; (2) equivalent forms do not demonstrate statistical similarity, but the differences in raw score statistics are compensated for in conversion to derived scores or in the forms' norm tables; or (3) comparable forms are similar in content but have no demonstrated statistical similarity. (See Alternate form reliability.)

Alternate form reliability. Publishers' provision of two or more versions of the same measure to permit several assessments of the same skills or behaviors (as in a pre-post or longitudinal study with the same group of staff). The use of alternate forms reduces concerns that scores may change solely as a consequence of "learning the test" from repeated administration of the same items. To demonstrate that both forms of the measure are essentially equivalent, a group of individuals takes both forms of the measure (the time between administrations may vary). Alternate form reliability is demonstrated if the scores on the two forms are highly correlated. (See Reliability.)

Bias analysis. Characteristics of a measure that unfairly favor one or more groups of individuals on the basis of factors such as agency type, job type or role, or reporter characteristics (for example, sex, race/ethnicity, or culture). In a statistical context, bias reflects a systematic error in scores that compromises the generalizability of the results to a broader population. One common statistical procedure for examining bias of assessment items is differential item functioning. (See Differential item functioning [DIF], Generalizability.)

Classical test theory. Theory used for most test development in the last century. Classical test theory states that the observed score is equal to the true score (the latent ability) plus error. The theory assumes that the error is distributed normally and uniformly among individuals, has an expected value of 0, and is not correlated with any variables. The item discrimination and difficulty in classical test theory (that is, point-biserial correlations and p values) are dependent upon the distribution of abilities in the sample so that large representative samples are needed to establish item properties. In classical test theory, these parameters are fixed and cannot be separated for an individual score. (See Latent trait).

Concurrent validity. Demonstration of the association (usually measured as a correlation) between a score on a given measure and performance on another measure of the same or similar construct obtained at approximately the same time (known specifically as convergent validity) or with a measure of a different construct (known as divergent validity). Concurrent validity of a measure also includes an estimate of the association between the measure and an outcome assessed at approximately the same time. (See Construct validity, Convergent validity, Criterion-related validity, Divergent validity).

Construct. The trait to be assessed (for example, interpersonal skills, fostering quality practices, or charisma). The construct is a concept or characteristic of an individual that a measure is supposed to measure.

Construct validity. Estimate of the degree to which a measure assesses the theoretical construct it claims to measure and to which inferences based on the measure are relevant to the construct. Different sources of evidence support estimates of construct validity including evidence of a positive relationship with other measures of that construct or a similar construct (convergent validity) and expected weak or negative relationships with other constructs (divergent or discriminant validity). Evidence of construct validity also includes criterion-related validity evidence that demonstrates a relationship between the score and an

independent measure of some something related to the construct, such as possessing a credential or qualification or demonstrating high observed quality. (See Convergent validity, Criterion-related validity, Divergent validity.)

Content validity. An indicator providing information about whether a measure includes items relevant to and representative of the construct it is supposed to assess. No statistics are associated with content validity. Instead, the indicator is based on the professional judgment of experts who review the items to verify that the measure represents the content that the developer intended and that the items provide variety and a range of difficulty. (See Construct.)

Convergent validity. A type of construct validity providing evidence of a positive relationship with other measures of that or a similar construct. This may be evaluated by looking at bivariate correlations between measures or the evidence may include the use of factor analysis that demonstrates that items in similar measures load on the same construct (while items in other measures load on different constructs demonstrating divergent validity). (See Construct validity, Divergent validity.)

Correlation. The degree to which two sets of scores or other data vary together, ranging from -1.0 (a perfect negative relationship) to 1.0 (a perfect positive relationship), with 0 indicating no association.

Criterion. A definition of acceptable performance levels or specific behaviors. Criteria may be used to develop scoring rubrics or to determine levels of practices for items on a measure such as "conducts child assessments three times a year".

Criterion-related validity. The extent to which scores on a measure are statistically related to a criterion (such as receiving a credential) or to scores on some other measure (preferably a well-respected or established measure) of the same objectives or criteria. It includes both concurrent validity (taken at same time) and predictive validity (the criterion measured in the future).

Cronbach's coefficient alpha. An estimate of internal consistency reliability that is, how well groups of items on a measure "hang together" or measure a particular trait or characteristic because of common factors among them. The greater the covariance among items, the higher the reliability is (and thus the higher the value of Cronbach's coefficient alpha). Values of the alpha can range from -1.0 to 1.0 with greater values indicating stronger internal consistency. The Cronbach's coefficient alpha is an extension of Kuder-Richardson Formula 20 (KR-20), a measure of internal consistency that is used when the items are dichotomous (right/wrong). (See KR-20 Kuder-Richardson Formula 20).

Differential item functioning (DIF). A statistical property of a test item. DIF is evident when different groups of individuals with the same overall ability or level on the trait being tested demonstrate differences in how they perform on an item according to their particular group membership (for example, male versus female, White versus Hispanic). An item does not show evidence of DIF when different groups of individuals who have roughly the same skill level (for example, pedagogical knowledge), regardless of group membership, perform similarly on the item. Typically, comparisons are based on sex, race/ethnicity, and education, but others are possible. Items demonstrating significant DIF often undergo review by content experts and may be removed from the measure if their inclusion unfairly favors one group over another.

Discriminant analysis. Statistical analysis to determine if a measure discriminates between individuals or programs with different expected levels on the measured trait (for example, an individual with less training would score lower on a knowledge assessment than individuals with relevant training).

Divergent validity (sometimes referred to as **Discriminant validity**). Evidence of a weaker or absent relationship between two measures intended to represent different constructs (for example, a lack of a

significant relationship between the individual's score on a leadership style measure and ratings of the student's social interaction). Divergent validity may also be demonstrated by a strong negative relationship between two constructs (for example, a measure of relational leadership that promotes people coming together to affect change may be negatively associated with a measure of transactional leadership that focuses on compliance). (See Construct, Convergent validity.)

Factor analysis. Statistical analysis that examines the pattern of relationships among items in related groups to measure underlying latent constructs (that is, unobservable abstract concepts), using correlations or a covariance matrix. Factor analysis may be exploratory (looking at how items group together in the data) or confirmatory (examining whether the relationships among items are consistent with a predetermined hypothesized factor structure). (See Correlation.)

Factor loadings estimated in factor analysis indicate the strength of the associations between the items and the latent constructs. Standardized factor loadings are typically in the range of -1 to 1. A factor loading close to zero indicates that the item is not contributing to the measurement of the latent construct. Items with factor loadings less than 0.40 suggest weaker associations with the latent constructs (Stevens 2012).

For confirmatory factor analysis (CFA), model fit can be assessed with various fit statistics. We focus on the Comparative Fit Index (CFI), Tucker Lewis Index (TLI; also referred to as nonnormed fit index or NNFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) as indicators. Exhibit A.1 shows the recommended fit statistics to assess the overall fit of the models. CFA generally requires 5 to 20 cases per parameter estimated (Kenny 2015). Chi-square is another indicator of model fit. However, for models with relatively large sample size ($n > 200$), chi-square statistics nearly always indicate poor model fit. Moreover, chi-square also indicates poorer model fit when the correlations are larger in the model. Therefore, we use other fit indices to assess model fit for these analyses.

Exhibit A.1 Recommended fit statistics for confirmatory factor analysis

Fit statistics	Recommended fit
Comparative Fit Index (CFI)	Brown (2015) suggests that a value of .90 or above is acceptable.
Normed Fit Index (NFI)	Bentler and Bonett (1980) suggests that a value of .90 or above is acceptable.
Tucker Lewis Index (TLI) or Nonnormed Fit Index (NNFI)	Brown (2015) suggests that a value of .90 or above is acceptable. The TLI is designed to correct for the complexity of the model, but is more sensitive to small sample sizes.
Root Mean Square Error of Approximation (RMSEA)	MacCallum et al. (1996) suggest that 0.01, 0.05, and 0.08 for RMSEA indicate excellent, good, and mediocre fit, respectively. Others have used 0.10 as the cutoff for poorly fitting models (Kenny 2015). RMSEA is sensitive to model complexity.
Standardized Root Mean Square Residual (SRMR)	A value of less than .08 for SRMR is considered a good fit (Hu and Bentler 1999).

Generalizability Theory (G-theory). A statistical theory for a more comprehensive estimate of the reliability and accuracy of results (Cronbach et al. 1972; Shavelson and Webb 1991). Traditional reliability estimates (such as Cronbach alpha and test-retest reliability) each provide a single measure of reliability. G-studies provide information about potential sources of error in measurement and what the reliability is under different conditions. Whenever a measure is administered, there are different factors (facets) that can influence and may add error in measurement, for example, raters, items, timing of measurement, number of observations. A G-study estimates the variance for the different facets in a measure and answers the question about how accurately the scores can be generalized to other administrations and conditions of measurement. The G-study is frequently followed by a decision study

(D-study). The D-study then uses the results from the G-study to estimate how the reliability of the measure might be improved, for example, increasing the number of observation cycles.

Internal consistency reliability. A measure of the reliability of a score derived from the relationship among items of a single measure and the extent to which they measure the same construct. Internal consistency reliability is presented as the correlation between groups of items or among all items. For example, split-half reliability refers to the correlation between the odd- and even-numbered items in an assessment. Another measure of internal consistency reliability is based on the correlations among all individual assessment items such as Cronbach's alpha or Kuder-Richardson Formula 20 (KR-20). (See Cronbach's coefficient alpha, KR-20, Split-half reliability.)

Interrater agreement (IRA). Provides information on the consensus of ratings across multiple individuals and can be used to determine the appropriateness to aggregate individual reports to a higher level (for example, when a teacher reports on the leadership behaviors of the principal and the information is reported at the school level). There are multiple statistics that can be used to assess the index of agreement (like a_{wg} or r_{wg}). IRA, for example, examines the agreement between teachers rating the same principal. There is not one standard "acceptable" metric as it would depends on the number of people surveyed and the specific statistic. Please see O'Neill 2017 for additional information on the various statistics and interpretation of them.

Inter-rater reliability. Extent to which different raters or observers obtain the same information; it can include agreement on scoring of items, administrative procedures, or observation of a given behavior. It is usually reported as either the correlation between the scores or ratings obtained by two observers or the percentage of items on which the two agree. Developers may also use an intra-class correlation (ICC) to compare the variance between raters to the total variance in the ratings. In research, inter-rater reliability is often a certification criterion for assessors/observers that must be met at the conclusion of training and during in-field data collection. (See Correlation, Reliability.)

Intraclass correlation (ICC). When used as a measure of inter-rater reliability, the ratio of the variance due to the independent variable (trait) divided by the explained variance plus the residual variance due to rater differences and measurement error. The ICC is sensitive to the sample of individuals surveyed, such that samples with more restricted variance will have lower reliability estimates than those with greater variance.

Intraclass correlations (ICC) can also be used to examine group reliability as justification for aggregating scores for reporting purposes. Higher between-group variance (for example, the differences between groups such as schools) indicates higher group reliability. Based on existing research (Bliese 2000) typical values range from 0.05 to 0.20. A score in this range or higher could be considered adequate.

IRT. See Item response theory model.

Item. A statement, question, exercise, or task on a measure.

Item response theory (IRT) model. A method of producing scale scores based on a set of principles of measurement that result in estimates not biased by the sample distribution of ability. IRT uses information from all of the items and all of the individuals to estimate the item difficulties and the person abilities on the same scale. IRT models use the responses of all individuals to all of the questions to estimate the item difficulties and the person abilities on the same scale. The individual's score on the measure is the estimate of the item difficulty at which the respondent has a 50 percent probability of answering the item correctly.

Kuder-Richardson Formula 20 (KR-20). A derivation of Cronbach's coefficient alpha that is used when items are dichotomous (right/wrong). KR-20 is often used as an indicator of internal consistency. Values can range from 0 to 1.0 with higher values indicating stronger internal consistency. The length of the measure, variance in scores and the difficulty of the test can influence the KR-20. (See Cronbach's coefficient alpha.)

Latent trait (latent ability). A construct that cannot be directly observed, for example, intelligence, leadership ability, empathy. A latent trait can be measured using observable behaviors or indicators that are related to the trait.

Likert scale. A type of rating scale that assesses varying levels of performance, behavior, or quality. It allows respondents to indicate the extent to which they agree or endorse a questionnaire statement. For example, a Likert scale may be used to assess teaching staff's perceptions of their center climate. Response categories may range in number (for example, a four-point scale may range from strongly agree, agree, or disagree to strongly disagree).

Measurement bias. See Bias analysis.

Meta-analysis. A statistical method that synthesizes the results from several independent studies of comparable phenomena to estimate the strength of the relationship between variables.

Normal curve equivalent (NCE). A standard score with a mean of 50 and a standard deviation of 21.06. The NCE ranges from 1 to 99 and is a conversion of percentile rank into an equal-interval scale, making the NCE more suitable than percentiles for comparisons relative to the sample or to a normative sample.

Norming sample. The group of individuals whose scores on a measure are used to establish the standardized scoring system, or norms for the measure. Norming samples are selected to be representative of the population of interest.

Norms. The distribution of expected scores obtained from the norming sample (see above) that describes performance on a particular measure relative to the average of those in the sample. Norms typically serve to represent a larger population.

Percentile rank. Indicates a score's relative ranking in units 0 to 100 to other scores in a sample, usually a nationally representative norming sample. Interpretation is based on the percentage of individuals in the norming sample that performed in a similar way. An individual whose score is at the 65th percentile has scored higher than 65 percent of the individuals in the norming sample (and higher than 65 percent of the individuals nationwide if the norming sample is nationally representative). However, caution should be taken in comparing percentiles to each other because the raw score difference between percentiles will vary depending on the percentiles' location and the distribution of scores. In other words, percentiles are not on an equal interval scale. Normal curve equivalents convert percentile ranks into equal interval scores for ease of comparison of performance over time and across assessments. (See Normed scores, Normal curve equivalent, Norming sample).

Predictive validity. Indicator of a type of criterion-related validity that demonstrates how accurately scores from a measure can predict scores on another measure or criteria assessed or gathered in the future. Researchers and measure developers determine whether the measure is correlated with later functioning. If the correlation of a measure with another measure obtained at a later time is high, evidence of predictive validity is established. If, for example, a measure of leadership in the fall is highly correlated with team effectiveness in the spring, the leadership measure could be said to have evidence of predictive validity. In some cases, researchers use other activities or events as the criterion, rather than another

measure. For example, researchers might show a positive correlation between leadership practices and program accreditation as evidence of predictive validity. (See Criterion-related validity).

Psychometrics. The study of psychological or educational measurement in areas such as knowledge, aptitude, attitude, skill, and the quality of the care and education environment. Psychometric properties document evidence that indicates how reliable and valid a measure is based on the purposes for which it was designed and used.

Rasch model (Rasch-based scores). A latent trait model, also considered a one-parameter item response theory (IRT) model. Rasch models assume single trait is measured and equal item discrimination. Rasch models estimate the student scores in relation to the difficulty of the items. Rasch-based scores are equal interval with both the student scores and the item difficulties estimated on the same scale. These scores are expressed in logits that have positive and negative values, and so are often transformed to have positive values.

Reliability. The extent to which scores obtained from a measure or group of measures are consistent over one or more possible sources of error, including time, raters, items, environment, and sample groups of a population. Indicators of reliability assess how dependable a measure is for the purpose it is used. Reliable measures are stable over time and include items that measure the same thing in different ways. Statistical measures of reliability are typically reported as coefficients, which range from 0 to 1.0, with a greater value reflecting greater reliability. Many researchers and assessment developers require that measures have reliability values of 0.7 or higher. Typical indicators of reliability include alternate form, internal consistency, inter-rater, and test-retest. An unreliable assessment cannot be valid. (See Alternate form reliability, Internal consistency, Inter-rater reliability, Test-retest reliability).

Sample. A selection of a specified number individuals from a larger set of people called the population.

Split-half reliability. A form of internal consistency reliability, obtained by splitting the items on a measure in half and obtaining two independent scores. The correlation between these two scores, usually adjusted using the Spearman-Brown formula (derived from classical test theory), provides an estimate of the reliability of the entire measure.

Standard error of measurement (SEM). The standard deviation of an individual's observed scores from repeated administrations of a measure under identical conditions. The SEM is typically estimated from group data (rather than from repeated measures from a single person) and can be interpreted as the precision or reliability of scores on the assessment. Every measure has a different SEM for a given sample of individuals.

Statistical significance. The finding that empirical data are inconsistent with a null hypothesis, usually that no difference exists between groups, at some specified probability level. A statistically significant finding shows that the probability of getting the finding (for example, two groups are different in skills) only by chance is low (for example, less than 5% of the time) even if the null hypothesis is true.

Subscale (also called **Subtest).** A set of items within a larger measure that assesses a particular aspect of the trait being measured. Subscales may be specified based on theoretical grounds (grouping items based on their content) or empirical evidence (factor analysis of items in a longer scale may reveal meaningful subscales).

Test-retest reliability. The stability of measure results over time. Evidence of test-retest reliability involves testing the same group of individuals at least twice, with a relatively short interval between administrations, usually no longer than a few days or weeks apart. The reliability coefficient is then obtained by correlating both sets of scores. The higher the test-retest reliability, the more stable the

measure is considered to be. Longer periods between administrations of the same measure will reduce the reliability, partly because the individual's situation (for example, skill) can be expected to change. Some also consider a measure to be test-retest reliability when an individual is tested on different forms of the same test. (See Alternate-form reliability.)

Validity. The degree to which an assessment accurately measures what it is designed to measure. Validity is often measured in comparison to other instruments established to measure the same or similar behavior/traits. Types of validity include content, construct, and predictive. An assessment cannot be valid if it is not reliable.

Sources

The definitions in this glossary were adapted from Malone et al. 2010 *Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions. Volume II: Technical Details, Measure Profiles, and Glossary (Appendices A-G)* with the following exceptions: recommended values of fit statistics for confirmatory factor analysis, definition of interrater agreement, and the use of intraclass correlations as a measure of group reliability.

Bentler, P.M., and D.G. Bonett. "Significance tests and goodness-of-fit in the analysis of covariance structures." *Psychological Bulletin, vol. 88,* 1980, pp. 588–600.

Bliese, P.D. "Within-Group Agreement, Non-Independence, and Reliability: Implications for Data Aggregation and Analysis." In K.J. Klein and S. W. J. Kozlowski (Editors), *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions.* San Francisco: Jossey-Bass, 2000.

Brown, T.A. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Publications, 2015.

Hu, L., and P.M. Bentler. "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives." *Structural Equation Modeling,* vol. 6, 1999, pp. 1–55.

Kenny, D.A. "Measuring Model Fit." 2015. Available at http://www.davidakenny.net/cm/fit.htm., Accessed on February 22, 2019.

MacCallum, R.C., M.W. Browne, and H.M. Sugawara. "Power Analysis and Determination of Sample Size for Covariance Structure Modeling." *Psychological Methods*, vol. 1, 1996, pp. 130–149.

Stevens, J. P. *Applied Multivariate Statistics for the Social Sciences*. New York: Routledge, 2012.

This page has been left blank for double-sided copying.

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

mathematica.org