Westat

Improving lives through research™

Center to Improve Project Performance

Evaluating Special Education Preservice Programs Resource Toolkit

Authors

Jill Lammert, Westat

Sarah Heinemeier, Compass Evaluation & Research

Jennifer M. Schaaf, Westat

Thomas A. Fiore, Westat

December 20, 2013

For easy reference, the Quick Start Guide on page iii lists a number of questions an evaluator may ask regarding the design or conduct of a PDP project evaluation, with links to the specific section in the Toolkit where related information may be found. A Table of Contents with hyperlinks to the different sections is on page v.

Updated 9/23/14 and 6/16/16

Prepared for:
Office of Special Education Programs
U.S. Department of Education
Washington, D.C.

Prepared by:
Westat
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

About this Toolkit

This Toolkit was developed as part of the Center to Improve Project Performance (CIPP) operated by Westat for the U.S. Department of Education, Office of Special Education Programs (OSEP). The authors thank the OSEP, Westat, and CEEDAR (Collaboration for Effective Educator Development, Accountability, and Reform) staff who provided input.

Suggested Citation:

Lammert, J. D., Heinemeier, S., Schaaf, J. M., & Fiore, T.A. (2016). *Evaluating special education preservice programs: Resource Toolkit*. Rockville, MD: Westat.

The Center to Improve Project Performance has been funded with Federal funds from the U.S. Department of Education, Office of Special Education Programs, under contract number ED-OSE-13-C-0049. The project officer is Dr. Patricia Gonzalez. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government.

IDEAs that Work
Office of Special Education Programs
U.S. Department of Education

Overview of the Center to Improve Project Performance

First formed in 2008, CIPP's overall mission is to advance the rigor and objectivity of evaluations conducted by or for OSEP-funded projects so that the results of these evaluations can be used by projects to improve their performance and used by OSEP for future funding decisions, strategic planning, and program performance measurement.

CIPP provides evaluation support, oversight, and technical assistance (TA) to OSEP projects. CIPP staff work with project and OSEP staff to refine project logic models and develop evaluations. Based on the evaluation design and plan, CIPP staff have overseen evaluation activities and provided technical assistance (TA), as needed, to the grantees including selecting samples; developing draft instruments; monitoring data collection and performing reliability checks; analyzing study data; providing accurate descriptions of the methods and valid interpretations of findings; and organizing, reviewing, and editing project evaluation reports. In addition to providing TA to OSEP-funded projects on request, CIPP staff prepare a variety of TA products focused on evaluation issues, and deliver presentations on evaluation through webinars and conferences.

Contact Information:

https://www.cippsite.org/SitePagesPublic/Home.aspx
1-888-843-4101

Thomas Fiore, CIPP Project Director
Westat
ThomasFiore@westat.com

Jill Lammert, CIPP Deputy Project Director
Westat
JillLammert@westat.com

Patricia Gonzalez, Project Officer
Office of Special Education Programs
U.S. Department of Education
Patricia.Gonzalez@ed.gov

Quick Start Guide

	Toolkit Section
Selecting Outcomes and Measures	
How do I select high-quality outcomes and measures?	2.2.1
Creating a Logic Model	
How do I develop a high-quality **logic model**?	2.1
Is there a **logic model template** that I can use?	Appendix A.1
Is there a **sample logic model** for an evaluation that includes graduate outcomes?	Appendix A.2
Is there a **sample logic model** for an evaluation that includes both graduate and student outcomes?	Appendix A.3
Creating an Evaluation Plan	
What elements should my **evaluation plan** contain?	2.2
Is there an **evaluation plan template** that I can use?	Appendix A.4
Is there a **sample evaluation plan** that includes both graduate and student outcomes?	Appendix A.5
Selecting a Sample & Choosing an Evaluation Design	
How do I **select my study sample**?	2.2.4
What **evaluation design** should I use?	2.2.2
How can I use **goal attainment scaling** to measure my project outcomes	4.3.5
How can I use **single case designs** in my data collection?	4.1.4
Planning for Data Collection	
What should I consider when **planning for data collection**?	2.2.5
Do I need to get approval from an **Institutional Review Board (IRB)**?	3.1.1
What do I need to consider when planning **data collection in schools**?	3.1.2
What do I need to consider when my evaluation involves **large/multiple school districts**?	3.1.2
Are there **sample district, school and graduate notification letters** that I can use?	Appendix A.6, A.7 & A.8
How do I get **consent** from participants?	3.2.2
Are there sample **consent forms** that I can use?	Appendix A.9 & A.10
Selecting Data Collection Measures	
What should I consider when **creating a new survey or adapting an existing one**?	4.3.1
How do I get a **good response rate to my surveys**?	4.3.1.4.1
What should I consider when I want to include **observations** in my data collection?	4.3.2
What should I consider when I want to include **interviews** in my data collection?	4.3.3
What should I consider when I want to include **focus groups** in my data collection?	4.3.4
Collecting Data	
What strategies can help me to **recruit study participants**?	3.2
How do I **keep track of PDP graduates**?	3.2.1
What elements should my **data tracking system** contain?	3.3.1
How do I obtain **access to student data** to measure the PDP graduates' performance?	3.1.3

Evaluating Special Education Preservice Programs
Resource Toolkit

Contents

Tables

Figures

Boxes

Introduction

This Toolkit is intended to support the design and conduct of high quality evaluations of the U.S. Department of Education's (Department) Office of Special Education Programs (OSEP) Personnel Development Program (PDP) projects. It is not intended to provide a lengthy discussion of every aspect of project evaluation. Rather, it is designed to highlight important issues to consider when planning and conducting a PDP project evaluation—and to offer a variety of resources related to evaluation, including sample templates and forms, links to professional standards and other educator resources, and recommended readings on research and evaluation.

The Toolkit is organized into four sections, plus an Appendix:

1. High-Quality Outcomes for PDP Project Evaluations
2. Planning the Evaluation
3. Conducting the Evaluation
4. Methodological Considerations

The Quick Start Guide on page iv lists a number of questions an evaluator may ask regarding the design or conduct of a PDP project evaluation, with links to the specific section in the Toolkit where related information may be found. A Table of Contents with hyperlinks is on page vi.

1. High-Quality Outcomes for PDP Project Evaluations

When selecting outcomes for evaluations, projects should select a mix of input-based, or process, measures and outcome-based measures. Input-based measures concern the activities of projects, while outcome-based measures concern the results of the activities. For example, an input-based measure might demonstrate how much time a PDP scholar spent as a student-teacher during their program, while an outcome-based measure might show how well that scholar performed in the classroom after graduation. Input-based measures show how a project is progressing during implementation, while outcome-based measures give an indication of to what extent and how well the project is achieving its goals. As such, when planning an evaluation, evaluators and project staff should strive to include the greatest number of high-quality outcome-based measures possible. The list that follows contains important categories of outcomes-based measures that are strong indicators of program effectiveness.

1. **Job placement and retention rates.** Personnel development programs (PDPs) are encouraged to work with states and districts to track shortage areas and report on whether or not the program graduates are (a) fully qualified, (b) filling these shortages, and (c) maintaining employment in shortage areas for at least three years.
2. **Graduate performance.** PDPs are also encouraged to measure the effectiveness of their program graduates. A common method for evaluating effectiveness is through surveys of the graduates and their supervisors. Surveys can include questions regarding the effectiveness of PDPs for developing

the skills graduates need in the earliest years of their careers. Other data, such as observations of PDP graduate performance, can also inform the question of preparation program effectiveness.

3. **Growth of students taught by program graduates.** Finally, PDPs are encouraged to make efforts to link the performance of their graduates with student outcomes. We recommend using "*multiple, valid measures*" of student outcomes and, if possible, measuring student growth over time. Some states, such as Colorado, Delaware, Florida, Georgia, Louisiana, North Carolina, Ohio and Tennessee, have developed and implemented systems that link student outcomes to teachers and teacher preparation programs.[1] Where these exist, PDPs are encouraged to work with states to obtain access to these data for evaluation purposes.

In the sections that follow we provide a framework for conceptualizing outcomes in these areas. Throughout, we pay specific attention to the context and constraints associated with evaluating outcomes for teachers and related-services providers working with students with disabilities.

[1] Doherty & Jacobs, 2013.

1.1. Job Placement and Retention

Almost all states have a shortage of special educators, including teachers and related-services professionals. In 2015-2016, 47 states and the District of Columbia reported shortages in special education or related fields.[2] Research has shown that special education positions can be among the hardest to fill.[3] Thus, states and PDPs are encouraged to track both shortages and placement of fully qualified teachers into shortage areas. Further, states and PDPs can track teacher retention within shortage areas.

Much has been written on the causes of personnel shortages in special education, including shortages of fully qualified special education teachers and related-services personnel in specific geographic locations and among certain age groups.[4] With regard to teachers, Tyler (2008) and Thornton, Peltier, and Medina (2007) described two factors that explain this ongoing problem: lack of qualified staff emerging from teacher preparation programs and relatively high rates of turnover among special education teachers. It is likely that in its application for OSEP funding, a PDP grantee already has described its project as an effort to respond to personnel shortages. Some grantees say that collaboration with schools, local education agencies, and the state education agency is particularly helpful in determining why shortages are occurring and describe how their PDP project contributes to alleviating shortages. In addition, resources that are available to help guide schools and districts in recruiting and retaining fully qualified, high quality personnel[5,6] also can serve as a guide in developing outcomes associated with job placement and retention for summative evaluation purposes.

1.2. Graduate Performance

Another category of outcome-based measures—graduate performance—focuses on the quality of PDP graduates. The assessment of teacher quality in particular has received much attention, as has the relation of personnel preparation programs to teacher quality.[7] However, there is widespread agreement that assessments of educator quality are better informed by the use of “multiple methods, multiple times, over multiple years.”[8] Methods that have been and continue to be debated include

[2] U.S. Department of Education (2015). New Mexico, New York, and Pennsylvania did not report teacher shortages, all US territories reported teacher shortages in special education.
[3] Tyler, 2008.
[4] See, for example, Billingsley, 2003; Billingsley, 2004; Gehrke & McCoy, 2007; McClesky et al., 2003; Nance & Calabrese, 2009; and Nichols et al., 2008.
[5] See, for example, Berry et al., 2012; Leko & Smith, 2010; Northeast Regional Resource Center, Learning Innovations at WestEd & University of the State of New York, The New York State Education Department, 2004; Sundeen & Wienke, 2009; and Thornton et al., 2007.
[6] The National Center to Improve Recruitment and Retention of Qualified Personnel for Children with Disabilities, for example, provides a framework for collaboration among state agencies, preparation programs (which most often are organized through Institutes of Higher Education), and local districts.
[7] The National Center for Teacher Quality, for example, intends to rate all teacher preparation programs and make its findings public.
[8] Ferguson, 2012, p. 25.

classroom observations, principal evaluations, teacher portfolios, teacher self-reports/self-evaluations, review of teacher products such as lesson plans.[9]

These approaches can be divided into assessments of educator practice and assessment of student outcomes.[10] These are briefly discussed below.

Assessments of Graduate Practice

- **Observations.** Tools for conducting observations of graduate performance range from informal case notes to standardized observation protocols. Multiple individuals might implement the observation, including mentors, assistant administrators, principals/supervisors, or subject matter chairpersons. Observations can be implemented for formative or summative purposes; formative observations are used to provide a graduate with feedback and guide their practice and summative observations are used to assess or grade graduate performance. Concerns regarding use of observations in evaluating graduate quality are often methodological: choice of instrument and whether or not its ratings are linked to graduate or student outcomes of importance, validity of an instrument for formative or summative purposes, rater and inter-rating training, and conditions under which an observation can or should be conducted.[11] See Section 4.3.2 for more information about planning and conducting observations.

- **Principal/supervisor evaluations.** Similar to observations, principal or supervisor evaluations can vary with regard to the formality of the process (e.g., Is there a protocol for the evaluation? Is there a standard process for collecting data?), the nature of the data used to inform the evaluation, and the use of the evaluation for formative or summative purposes.[12] An advantage of such evaluations is that principals or supervisors can observe graduates in various applied contexts and collect data on graduate performance over time. These evaluations also take place within the local school and district educational context, perhaps with a school or district strategic or improvement plan in mind.

- **Educator portfolios.** Portfolios rely upon a collection of artifacts. Unlike the analyses already described, however, graduates develop the portfolios for use in the evaluation and thus are given the task of collecting the raw materials for their assessment.[13] As with the other methods discussed in this section, the best data arguably are collected when there are specific instructions for graduates on how to construct their portfolios (e.g., what types of artifacts to include, what range of performance to represent, how frequently to update the portfolio, the minimum level of information that must be included).[14] Similarly, the reliability of this approach increases when there are specific instructions for the assessment of the portfolio by the assessor—in essence the creation and implementation of a rubric for assessing the portfolio, as

[9] See, for example, Ferguson, 2012; Goe et al., 2008; Goe et al., 2011; Goe & Holdheide, 2011; Tyler, 2011.
[10] Tyler, 2011.
[11] Goe et al., 2008.
[12] Ibid.
[13] Goe et al., 2008.
[14] Ibid.

well as the correlation of findings from the review with outcomes of interest. Training assessors to achieve high levels of reliability also increases the overall quality of data received from this approach (see Section 4.3.2.6 for a discussion of calculating inter-rater reliability, or inter-observer agreement).

- **Graduate self-reports/self-evaluations.** Graduates can collect and report data about their own practices using tools such as surveys, interviews, case notes, and logs. This approach can prove a very rich source of data as graduates can make the connection between planning, implementation, and outcome of the techniques used in their practices.[15] However, this approach also is difficult to standardize and is susceptible to bias, as graduates may be very subjective in their responses. As with other methods, the quality of the data increases when graduates are trained in the data collection methods and, where possible, trained to high reliability in how to use data collection instruments.

- **Review of graduate products.** In this approach, the quality of products such as intervention plans, teacher-made lesson plans, assignments, and tests is assessed. The value of this approach depends upon the rigor of the review process and the quality of the instruments guiding the review. Evaluators can take steps to improve the quality of the review process such as developing a protocol or guide for conducting the assessment, pilot testing the protocol to assess its validity and reliability, training all assessors, and correlating findings from the review with outcomes of interest to validate the use of this approach for evaluating product quality.

- **Student surveys.** Students, arguably, may be the richest source of data regarding teacher skill and technique. In fact, a study conducted by the Bill and Melinda Gates Foundation (2011) found that student perceptions of a teacher's strengths and weaknesses were consistent across different groups of students, and that students seemed to know effective teaching when they experienced it. However, Goe et al. (2008) cautioned that student data may be skewed in some cases by age, grade level (elementary, middle, or high school), the expected or actual grade received from a teacher in a course, or the course's level of difficulty or the nature of the requirements for course completion. As with any data collection method, the quality of student survey or other self-reported student data will depend upon the validity and reliability of data collection instruments and the rigor of the data collection methods.

Assessment of Student Outcomes

- **Analysis of student products.** In this approach, the quality of student products such as classwork, student portfolios, or other non-standardized tests is assessed. This approach can be valuable when a student does not participate in the school's standardized tests. The analysis of student products can follow standardized procedures such as using a standard protocol to rate the products and training raters (e.g., teachers or administrators) in the use of the protocol until they achieve high reliability/inter-rater reliability (see Section 4.3.2.6 for a discussion of inter-rater reliability, or inter-observer agreement). Much of the effort in implementing this type of

[15] Ibid.

analysis, in fact, may be allocated to the development and validation of a protocol (or instrument) and the training to reliability of the assessors.[16]

- **Standardized student performance measures.** There is increasing interest in the use of student performance on standardized measures to inform teacher evaluations. However, it is quite controversial. As Goe stated "*it holds teachers solely accountable for achievement, rather than including others who also contribute to student outcomes. Using a single score for a teacher as a measure of his or her effectiveness suggests that all, or nearly all, of the student learning in a particular subject or classroom in a given year was the product of a single teacher's efforts*."[17] Goe presented the weaknesses of this general approach as including (a) lack of alignment between assessment method and the nature and content of instruction (i.e., the test or assessment is not well aligned with what was actually taught in the classroom), (b) presence of classroom effects, (c) range in the validity of techniques used to assess student performance, and (d) variation in student ability upon entering the classroom.

1.3. Student Growth

In recent years there has been a movement to incorporate student achievement and outcomes into teacher evaluation systems. Between 2009 and 2015 the number of states requiring student growth and achievement to be considered in teacher evaluations grew from 15 to 43.[18] However, the field is currently "at a crossroads in implementing measures of educator effectiveness."[19] For example, there are concerns about the validity and reliability of the models themselves[20] as well as their intended and unintended effects.[21] In addition to debates about how best to measure teacher performance, concerns about the [lack of] effectiveness of teacher preparation programs[22] have helped to fuel debate about the best way to measure these programs.[23] The use of these types of evaluation systems can be particularly problematic for professionals within special education. At this time, "*little is known—in terms of research and practice—about whether student growth can be adequately measured for students with disabilities and appropriately attributed to teachers for the purpose of teacher evaluation*."[24]

---

[16] Ibid.
[17] 2008, p. 45.
[18] Doherty & Jacobs, 2015.
[19] Doherty & Jacobs, 2015.
[20] See Loeb & Candelaria (2015) for a study that assesses the stability of VAM scores and Polikoff & Porter (2014) for information about how model scores align with instructional standards and observational and student survey measures of teacher quality.
[21] Harris & Herrington, 2015.
[22] Feuer, Floden, Chudowsky, & Ahn, 2013.
[23] For example, the National Center for Analysis of Longitudinal Data in Education Research (CALDER) has produced four working papers examining the conceptual and technical challenges facing efforts to evaluate preservice teacher training programs. For more information see http://www.caldercenter.org/calder-conversation/calder-conversations-evaluating-teacher-training-programs
[24] Brownell & Jones, 2015; Holdheide, Browder, Warren, Buzick, & Jones, 2012.

The challenges and possible approaches to using student growth have been addressed in literature reviews[25] and in a recent report on the challenges that appeared in the rollout of large-scale assessments to measure college-and career-ready (CCR) standards.[26] Specific approaches to using student growth were elaborated in a forum led by a group of researchers and practitioners held in September 2011.[27] The forum included four models for measuring growth for students with disabilities and identified ways to use such data for teacher evaluations.[28] The models included value-added models, the use of student learning objectives, classroom-based measures, and alternate assessments based on alternate achievement standards (AA-AAS). These are briefly discussed below.

- *Value-added models* (VAM) depend upon statistical techniques to isolate individual teacher or school effects on academic performance. These techniques control for student socioeconomic and demographic factors, a student's prior academic achievement, and school or community influences. Value-added models are considered rigorous and relatively objective, yet there are concerns about the application of VAM, particularly to evaluate the performance of special educators.[29] These concerns include the limited participation of students with disabilities in standardized assessments of academic achievement and the (typically) small sample sizes of students with disabilities in individual classrooms—both of which affect the reliability of analyses of individual educators' contributions to student growth. Further, if students are not consistently tested because of disability (or grade or subject), the VAM cannot be applied. The Council for Exceptional Children released a position statement that cautions against the use of VAM to evaluate special education teachers until there is a general consensus among researchers about the validity of the models, and other national associations have also expressed concerns.[30] In addition to the concerns about using VAM to evaluate teachers, researchers have expressed reservations about using VAM to evaluate PDPs. For example, Haertel explicitly discouraged the use of VAM to evaluate *individual* teacher preparation programs, but he argues that VAM can be useful for comparing the effects of particular training approaches across multiple training programs.[31] Goldhaber has argued that VAM can be informative about program effectiveness but cannot distinguish between effects of the program's training compared to the already existing talents of those admitted to the program.[32]

[25] Noell, Brownell, Buzick, & Jones, 2014.
[26] Lazarus & Heritage, 2016.
[27] Sponsored by the National Comprehensive Center for Teacher Quality, the Council of Chief State School Officers, Assessing Special Education Students State Collaborative on Assessment and Student Standards, and Educational Testing Service.
[28] Holdheide et.al., 2012.
[29] McCaffrey & Buzick, 2014.
[30] Council for Exceptional Children, 2012. The American Educational Research Association (AERA) and the American Statistical Association (ASA) have also cautioned against the use of VAM. For more information, see AERA Statement of Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html and ASA Statement on Using Value-Added Models for Educational Assessment http://www.amstat.org/policy/pdfs/asa_vam_statement.pdf
[31] Haertel, 2013.
[32] Goldhaber, 2013.

For more information on using value-added modeling to assess educator outcomes see:

- *Incorporating Student Performance Measures into Teacher Evaluation Systems.* Rand Education. Available online at: http://www.rand.org/content/dam/rand/pubs/technical_reports/2010/RAND_TR917.pdf
- *Value-Added Modeling 101* *http://www.rand.org/education/projects/measuring-teacher-effectiveness/value-added-modeling.html*
- McCaffrey, D. F., & Buzick, H. (2014). Is value-added accurate for teachers of students with disabilities? *Carnegie Knowledge Network*. Stanford, CA: Carnegie Knowledge Network.

- *Student-learning objectives* (SLOs) are an alternative means of assessing student growth, when standardized test scores are not available. SLOs typically are aligned with district or state achievement goals, constructed through team discussion and with the participation of key administrators, and applicable to more than one student (e.g., a classroom, a set of classrooms, a school, or a district). However, it is possible to individualize SLOs for each student, so that a student's current proficiency and past performance can be used to create reasonable and achievable goals for that student. It is not possible to control for other variables (e.g., socioeconomic and demographic factors) that may affect student performance in the measurement of SLOs.

For more information on student-learning objectives see:

- *Methods of Evaluating Teacher Effectiveness.* National Comprehensive Center for Teacher Quality. Available online at: http://files.eric.ed.gov/fulltext/ED543666.pdf
- *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis.* National Comprehensive Center for Teacher Quality. Available online at: http://files.eric.ed.gov/fulltext/ED521228.pdf

- *Classroom-based measures* are another means of assessing student growth when standardized test scores are not available. These measures typically are developed and implemented by teachers and correspond to state learning standards.[33] Many classroom-based measures are formative in nature, providing data that can guide instruction. These measures allow educators to track individual student progress over shorter timeframes than most standardized tests, and depend upon teachers administering assessments during class time. However, as with SLOs, this approach brings with it the need for teacher training and oversight, and does not allow for the standardized inclusion and control of other variables or past academic performance. Also of concern is the standardization of measures across classrooms (and possibly across subjects) to ensure that the same levels of proficiency and achievement are assessed for all students.

[33] An example is the subject matter test used by teachers to determine the extent to which students have mastered a unit of study.

For more information on classroom-based measures see:

- *Methods of Evaluating Teacher Effectiveness.* National Comprehensive Center for Teacher Quality. Available online at: http://files.eric.ed.gov/fulltext/ED543666.pdf
- *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis.* National Comprehensive Center for Teacher Quality. Available online at: http://files.eric.ed.gov/fulltext/ED521228.pdf

- *Alternate assessments/alternate achievement standards* (AA-AAS) are approaches specifically developed for the assessment of the performance of students with significant cognitive disabilities. Generally speaking, these approaches take student learning needs into consideration and are aligned with state learning standards and goals. The assessments accommodate student learning needs by varying the range or complexity of test items or by relying on student products or rating scales.[34] AA-AAS also take student testing needs into consideration, allowing for one-on-one test administration or testing environments that accommodate specific learning or testing needs.

For more information on alternate assessments or alternate achievement standards see:

- *Scoring Alternate Assessments Based on Alternate Achievement Standards: A Proposed Typology of AA-AAS Scoring Practices.* National Alternate Assessment Center. Available online at: http://www.naacpartners.org/publications/ScoringTypology.pdf
- *Methods of Evaluating Teacher Effectiveness.* National Comprehensive Center for Teacher Quality. Available online at: http://files.eric.ed.gov/fulltext/ED543666.pdf
- *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis.* National Comprehensive Center for Teacher Quality. Available online at: http://files.eric.ed.gov/fulltext/ED521228.pdf
- *A Proposed Typology for Characterizing States' Alternate Assessments Based on Alternate Achievement Standards: Developing a Common Vocabulary to Describe these Assessments.* National Alternate Assessment Center. Available online at: http://www.naacpartners.org/publications/ApproachTypology.pdf

The types of methods and measures used to evaluate educator quality in summative evaluations will naturally depend upon the evaluation questions and resources available. Construction and measurement of outcomes related to personnel quality can reflect diverse thoughts on quality and the supports (e.g., preparation, mentoring, induction periods) and time necessary for new personnel to reach (or be expected to reach) a high-quality level of performance.[35] Construction and measurement of outcomes also must attend to the quality of data collection instruments and the rigor of the data collection methods. Collecting a lot of data poorly will yield findings with relatively little value. While research supports evaluating teacher quality and performance with multiple methods and from multiple

[34] See, for example, Quenemon et al.,2010.
[35] See, for example, Goldrick et al., 2012.

sources, the evaluator also must consider the overall quality of the data collected and must expend resources efficiently so as to receive the highest quality data possible.

The next sections provide guidance and information on establishing outcomes, creating a logic model and developing a comprehensive evaluation plan. We hope these sections will assist project managers and evaluators in designing and implementing an evaluation that includes high-quality outcomes based measures to the greatest extent possible. Additionally, the appendices contain sample templates, forms, and other resources that can be used in the design and implementation of evaluations.

2. Planning the Evaluation

In this section we present a framework for planning an evaluation of a PDP project that can provide evidence of the project's progress toward achieving its outcomes, starting with the creation of a logic model and continuing with development of a comprehensive evaluation plan. This is followed in Section 3 by information on some of the technical aspects of conducting an evaluation, including data collection, management, aggregation and analysis, and reporting—with a particular focus on summative aspects.

2.1. Creating a Logic Model

A logic model is an essential starting point for developing an effective evaluation plan. A logic model is a good way to visualize the inputs, strategies, activities, and outputs necessary to respond to a need and make progress towards a desired outcome. The logic model helps (1) define outcomes that are meaningfully connected to project activities and (2) support evaluations so that the process will improve a project's overall performance. In general, a logic model should be precise and include features and content that support and promote the project's evaluation, but there are many different variations on logic models (see, for example, the W.K. Kellogg Foundation Logic Model Development Guide, http://www.wkkf.org/resource-directory/resource/2006/02/wk-kellogg-foundation-logic-model-development-guide, and Frechtling, 2007).

The CIPP Logic Model Template provided in this Toolkit (Appendix A.1) is designed to portray a project's overall plan and clarify the relationships among a project's goals, strategies and activities, outputs, and projected outcomes. Inputs and external factors are also included. These elements of the logic model are briefly described below.

- **Goals/Objectives**—Goals capture the overarching purposes of the project. Goals make clear the anticipated impact on systems or individuals. Goals imply gaps or deficits that will be remedied when the project produces its long-term outcomes. Objectives, if used in a logic model, are targeted sub-goals.

- **Inputs**—Inputs include the resources that are available to the project. This includes external funding, internal resources, and intangibles such as experience and the state of the knowledge in the field.

- **External Factors**—External factors relate to the context in which the project is being implemented. This may include other federal initiatives, the OSEP policy environment, an institution's accumulated experience and visibility, and public demand for or receptiveness to the project.

- **Strategies/Activities**—Strategies are the broad approaches to addressing the goals and generally include multiple activities. Activities are the specific actions funded by the grant or supported by other resources under the umbrella of the project.

- **Outputs**—Outputs are the direct results of the project activities, including project products and programs. Most outputs will be quantifiable, including tallies of the number of products and programs or counts of the customer contacts with those products and programs.

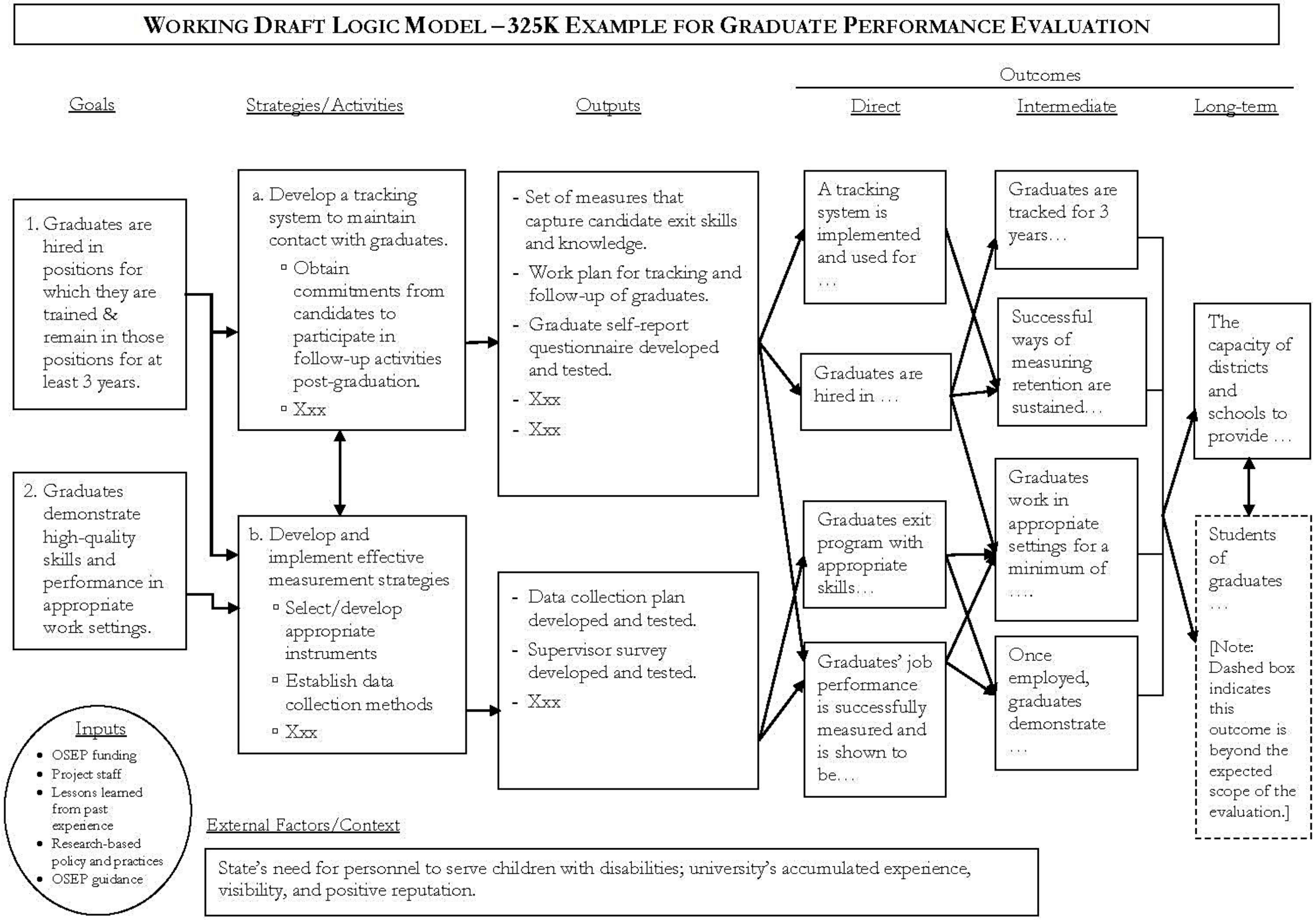
- **Direct/Intermediate Outcomes**—Direct outcomes are what customers do or become as a result of outputs. Usually, direct outcomes are changes in the customers' actions or behaviors based on knowledge or skills acquired through project outputs. Intermediate outcomes result either directly from outputs or indirectly through direct outcomes. Often, intermediate outcomes are changes in the knowledge, skills, or behavior of persons touched by the project's direct customers. They generally come later in time than direct outcomes and often represent a step between direct outcomes and long-term outcomes.

- **Long-term Outcomes**—Long-term outcomes are the broadest project outcomes and follow logically from the direct and intermediate outcomes. They are the results that fulfill the project's goals. However, they may not be assessed during the evaluation due to time or resource constraints. Outputs, direct outcomes, and intermediate outcomes all contribute to the achievement of the long-term outcomes. Although the long-term outcomes represent fulfillment of the purpose of the project, they may or may not represent the achievement of a desired larger project impact. That is, the project may have an anticipated impact that is beyond the immediate scope of the project, either temporally or conceptually, and thus beyond the scope of the logic model.

A completed logic model depicts how the project is expected to work and thus represents the program logic or the program's theory of change. As such, the overall logic model can be thought of as defining a hypothesis or a series of hypotheses along these lines: If we use these resources to do these activities we will get these outcomes. Evaluation is then the test of those hypotheses, and a logic model can be used by evaluators and project managers to refine and guide data collection and analyses for assessing both process and performance.

In developing a logic model, we find it helpful to begin with a summary chart that contains the information that will populate the logic model. The summary chart in Appendix A.1 outlines the OSEP priority, assumptions, external factors/context, and inputs. It then displays, in table format from left to right, the project's goals and objectives, strategies and activities, outputs, and outcomes. The logic model then can be created based on the content in the summary chart. The content of the logic model is less comprehensive than that in the summary chart, using lines and arrows to depict the temporal and causal connections among the various project elements. Not surprisingly, multiple lines or arrows may come to or from multiple boxes, indicating the complexity of the expected relationships. Both the summary chart and the logic model can be updated continuously as the evaluation progresses, such as when planned activities are revised or when unintended outcomes occur. The logic model also may change as the relationships among the components develop over time, most likely by becoming more complex and interactive. Figure 1 presents an example of a hypothetical PDP project logic model that has outcomes related to both graduate and student performance (see Appendix A.2 for an example of another logic model that only focuses on graduate outcomes).

Figure 1. Sample Logic Model



2.2. Developing an Evaluation Plan

Once the logic model is complete, it is time to develop the more comprehensive evaluation plan. The evaluation plan ties together the project's approach to collecting, managing, aggregating, analyzing, and reporting data on outcomes. The evaluation plan is a step-by-step guide to the "*who, what, where, when, how, and why*" of an evaluation. Evaluation plans typically have the following sections:

Introduction. The introduction generally incorporates an overview of the need that the project is responding to and a short discussion of why the need exists. The introduction also often contains background information on the project and the project's theory of change, which is a summary of the activities and strategies that are incorporated into a project model so as to bring about a desired change or outcome. The project logic model often is included here. It is common for evaluators to document the research or empirical evidence that supports the project model, thereby establishing that the model has a good or better likelihood of success in achieving the desired outcomes. The introduction also can present any unique or confounding factors that can or should be considered when evaluating the project.

Identification of Evaluation Questions. This section of the evaluation plan presents the questions that guide the evaluation. The evaluation questions reflect the goals of the evaluation; some (if not most) of the evaluation's goals may be established in the request for applications or in the application for funding. If the questions were not defined in advance, they should be developed through discussions with the project staff and key stakeholders, including the funder, as appropriate. Further, the questions should be based on a thorough understanding of the project's overarching objectives and program theory; consequently, it is not uncommon for evaluation questions developed prior to the start of an evaluation to change once the team becomes more aware of the specific context and focus of the project.

There generally are two types of evaluation questions: formative and summative. Formative evaluation questions focus on the project's processes and address the extent to which (and how well) the project is being implemented. This includes questions related to whether the project is complying with established rules or policies, proceeding as planned, and producing the expected outputs—sometimes called monitoring questions—as well as questions addressing the fidelity of implementation of the project model. Formative types of questions include:

- Are key inputs being utilized as expected? As designed?
- How are activities being received by participants?
- Are key outputs occurring at the desired level?
- What do participants and stakeholders like about the current program?
- What changes do participants and stakeholders suggest should be made?

Summative evaluation questions target the extent to which a project achieves its expected outcomes. These questions typically require an investigation of the extent to which a change has occurred, the factors associated with a change, or the measurement of change among different populations.

Summative questions are best addressed using comparison data (e.g., for treatment and control groups) to give the evaluator an idea of the counterfactual—that is, what would have happened if the project had not been implemented. Examples of summative types of questions include:

- What outcomes (expected and unexpected) have occurred?
- What expected outcomes have not occurred?
- Where is change the greatest?
- What impacts (expected and unexpected) have occurred?
- What expected impacts have not occurred?
- To what degree have outcomes occurred?
- What is the unique contribution of the program to the observed change?
- What is the cost/benefit of these outcomes?

When developing evaluation questions it is important to be as specific as possible. For example, a summative question for a PDP project evaluation might be "*Do increased numbers of program graduates become fully qualified, find positions in fields appropriate to their training, and retain those positions for three years?*" or "*Do higher numbers of students receiving special education services from the PDP program graduates, and who are eligible to participate in standardizing testing, make adequate yearly progress on standardized state assessments of reading and mathematics than students of graduates from other institutions?*"

It is common for formative evaluations not to include a summative component. It is difficult, however, to complete a comprehensive summative evaluation without paying any attention to formative details, as formative evaluation provides data to test fidelity to the project model and explain why the desired changes may or may not be occurring. However, since the purpose of this Toolkit is to provide resources related to the design and conduct of PDP project evaluations, we limit our focus throughout primarily to summative aspects of evaluation.

Methodology. The methodology section of the evaluation plan is where the evaluator details the specific approach to data collection, data management, data aggregation and analysis, and reporting. This can be done through the development of a data collection plan (see Section 2.2.5), a data analysis plan (see Section 2.2.3) and a data aggregation and analysis plan (see Section 3.4.2).

Timeframe and Responsibilities. Finally, the evaluation plan should contain a timeframe for all of the evaluation activities and a clear identification of the party or parties responsible for each step (see Section 2.2.5.6 for information on creating a timeline of evaluation activities and Appendix A.4 for the CIPP Evaluation Plan Template and Appendix A.5 for a sample completed Template).

We recommend that evaluators prepare a comprehensive evaluation plan early in the project period (ideally within 3 months of startup) and then conduct their evaluations in a manner that is consistent with their evaluation plans. Completing an evaluation plan early on will give the evaluator a roadmap for conducting the evaluation and will enable the evaluator to incorporate the plan into subsequent reports, thereby reducing the amount of time needed to document how the evaluation was carried out at a time when resources typically are limited and time is of the essence. Another benefit of following a pre-established plan is that it will help evaluators to avoid the appearance of "fishing" for positive results during data analysis while obscuring results that may not show the project in a positive light. Of course, there are myriad reasons why an evaluator may need to make changes to an evaluation plan during the course of an evaluation, but whenever possible evaluators are urged to follow their plans for data collection, analysis, and reporting. When this is not possible, evaluators should document any changes and the reasons for them in their reports.

2.2.1. Identifying High-Quality Outcomes and Measures

The quality of an evaluation depends in large part on the quality of the outcomes used to demonstrate project effects. As mentioned above, there are three levels of outcomes: direct, intermediate, and long-term. Direct outcomes are among the first changes that can be recognized, and they usually are occurring as an immediate result of program activities and investments. Direct outcomes are the first signs that later outcomes are achievable. Intermediate outcomes typically represent the cumulative effect or the somewhat more distal effect of direct outcomes, while long-term outcomes speak to alleviation (or elimination) of the originating need.

The high-quality measurement of outcomes depends in part on the nature of the outcomes themselves and in part on the rigor of the data collection and analysis procedures used to evaluate the outcomes. The best outcomes are rigorous, have a high degree of utility, and are informed by high-quality data. Rigorous outcomes are those that are considered valid and that have been measured with a great degree of methodological precision and accuracy. An outcome with a high degree of utility is useful in informing decisions and planning. Finally, an outcome should be informed by data that have been collected using high-quality measures (see, for example, Boller et al., 2010 and Mallone et al., 2010).

In the section below we outline one approach to determining the quality of an outcome. This is followed by a brief discussion of the characteristics of high-quality outcome measures.

2.2.1.1. The SMART Approach to Determining Outcome Quality

One way to assess an outcome's rigor and utility is to determine if the outcome is "SMART." This acronym was coined by George Doran[36] for management purposes, but has since been used by multiple authors to reflect varied terms to describe project objectives or outcomes. With an adaptation to better apply to evaluation of PDPs, the acronym describes objectives that are "*Specific and Clearly Stated, Measurable and Based on Data, Attainable and Realistic, Relevant, and Time-Bound*."

- ***Specific (and Clearly Stated)***. Outcome statements should reflect the need or problem the program is responding to, in sufficient detail, such that the reader can determine if the originating need or problem has been addressed. Thus, a well-specified and documented need

[36] Doran, 1981.

or problem statement can be an important first step in creating a well-specified outcome statement. The examples presented in Table 1 illustrate sample need and outcome statements.

Table 1. Examples of Non-Specific versus Specific Need and Outcome Statements

Sample Need/ Outcome	Non-Specific	Specific
Need #1	Rural districts have a hard time retaining special education teachers.	30% of fully qualified special education teachers in the rural districts in which program graduates are employed leave their position within three years.
Outcome #1	Retention rates in rural districts.	The number or percent of fully qualified special education teachers graduating from our program and employed in rural districts who remain in their position for more than three years.
Need #2	Program graduates are not finding jobs within their training area.	35% of degree/certification recipients who are unable to find teaching positions in the area for which they were trained upon graduation.
Outcome #2	Hiring rate in appropriate positions.	The number or percent of degree/certification recipients who have obtained teaching positions in the area for which they were trained upon graduation.
Need #3	Children are not arriving at school ready to learn.	30% of entering kindergarten students are below expectations in one or more developmental domains.
Outcome #3	Children arrive at school ready to learn.	The number or percent of entering kindergarten students who start kindergarten at or above developmental expectations in all domains.
Need #4	Special education students are falling behind on standardized assessments.	45% of students receiving special education services, and who are eligible to participate in standardizing testing and attend schools in districts that we serve, fail to make adequate yearly progress on standardized state assessments of reading and mathematics.
Outcome #4	Special education student achievement on standardized assessments.	The number or percent of students (who are eligible to participate in standardized testing and attend schools in districts that we serve) receiving special education services from our graduates who make adequate yearly progress on standardized state assessments of reading and mathematics.

- ***Measureable (and Based on Data).*** All outcomes must be assessed using standardized procedures for the collection, aggregation, and analysis of data that are relevant to the outcome. Creation of a well-crafted need or problem statement can be a very important step in determining which data are best for outcome assessment. Table 2 presents examples of good and poor data matches for a set of sample outcomes. Another necessary step is to determine *a priori* if the data are, or can be, available for use in outcome assessment. If high-quality, outcome-specific data are not available through existing data collection and management systems, the evaluator must determine if and how the data can be collected. Ultimately, if high-quality, outcome-specific data are not available and cannot be collected for a specific outcome (within the limits of the available time and resources), it is better to select another outcome for which such data *are* available than to report results for an outcome based on poor-quality data.

Table 2. Examples of Poor versus Good Outcome-Data Matches

Sample Outcome	Poor Data Match	Good Data Match
Outcome #1: The number or percent of degree/certification recipients who have obtained teaching positions in the area for which they were trained upon graduation	For graduates of the program, over time: - Number of graduates - Number of graduates who have positions by graduation	For graduates of the program, over time: - Number of graduates - Number of graduates who have positions by graduation - Area of training for all graduates - Position title and duties for all positions obtained by graduates
Outcome #2: The number or percent of fully qualified special education teachers graduating from our program and employed in rural districts who remain in their position for more than three years.	For districts served by the program, over time: - District name - Number of special education positions - Annual turnover rate	For districts served by the program, over time: - Rural identifier (is the district considered rural?) - Number of special education staff who are graduates of our program who meet the definition of fully qualified - Dates that our program graduates begin and end employment in the district
Outcome #3: The number or percent of entering kindergarten students who start kindergarten at or above developmental expectations in all domains.	For districts served by the program, over time: - Number of entering kindergarteners - Number of kindergarteners considered ready by kindergarten teachers	For districts served by the program, over time: - Number of entering kindergarteners - Number of entering kindergarteners who received developmental assessments (for one or more developmental domains) Of the students who received assessments: - The number and percent who were assessed as at or above developmental expectations, for each domain assessed.
Outcome #4: The number or percent of students (who are eligible to participate in standardized testing and attend schools in districts that we serve) receiving special education services from our graduates who make adequate yearly progress on standardized state assessments of reading and mathematics.	For districts served by the program, over time: - Standardized reading and mathematics scores - Number of special education students	For districts served by the program, over time: - Number of students taught by program graduates - Number of students taught by program graduates who participated in standardized testing - Standardized reading and mathematics scores - Definition, per student, of annual expectations (e.g., adequate yearly progress)

- ***Attainable (and Realistic).*** Good outcomes reflect changes that are achievable within a given timeframe. It may be helpful to have content- and problem-area experts on the evaluation team to help decide the goals and outcomes that are achievable during the project period. This discussion should address the resources that will be necessary to achieve the goal—in other words, are there sufficient financial, tangible, and other resources in place for the program to achieve its goals and outcomes within the time frame allotted? The process of answering these questions should take place during the development of the project logic model. The discussion also may include the likelihood that each program graduate will experience success in the program and enter employment in special education, as well as the current educational and policy climates that may affect employment (and attrition) decisions, or state and district policies regarding assessment of graduate performance.

- ***Relevant (to Objectives).*** Relevant outcomes address the degree to which the underlying need or problem has been alleviated, reflect needs and problems of consequence to communities and schools, and generate information for future decision-making. In the context of a PDP project evaluation, outcome relevance may best be defined broadly, to include educator placement and retention, educator quality, and student growth. With this in mind, the concept of "relevant" outcomes captures the desired changes that can occur after services are deployed to achieve these outcomes.

- ***Time-Bound.*** Good outcomes are achievable within a defined period of time. Our experience with the PDP projects indicates that direct outcomes generally can be observed within one program year, whereas long-term outcomes can take the entire grant period or more to assess and achieve—it is not uncommon or inappropriate for evaluators and program staff to identify long-term outcomes that are expected to occur well beyond the period of a grant. Intermediate-outcomes generally represent changes that occur between the direct and long-term outcomes and, because they are likely to fall within the time boundaries of the grant period, may be the most distal outcomes on which the evaluation can realistically focus.

As can be seen, developing outcomes that have all these criteria requires the evaluator to have at least some background knowledge of why needs or issues exist, the steps required to address these needs or issues, and the length of time that will be required to complete these steps. In many situations a framework for addressing these issues will have been developed as part of the application for OSEP funding. In cases when this information is not available, existing research from similar or existing programs may be available. In other cases, an evaluator may use information from related fields or service areas to construct a model for how, when, and under what circumstances change may occur.

Of course, having a high-quality outcome means nothing if the data related to that outcome are collected using low-quality measures. In the next section we discuss characteristics of high-quality outcome measures.

2.2.1.2. Characteristics of High-Quality Outcome Measures

ED's Institute of Education Sciences' What Works Clearinghouse (WWC) Procedures and Standards Handbook[37] (http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19) has identified four characteristics high-quality outcome measures possess:

- **Face validity**—The measure must appear to be a valid measure of the outcome (e.g., a reading fluency test should not be used to measure mathematics outcomes).

- **Adequate reliability**—This depends on the type of outcome measure (test score, scale, observation measure) and whether or not the measure is based on a standardized test or state-required achievement test. The WWC recommends following these minimum standards: (a) internal consistency (such as Cronbach's alpha) of 0.50 or higher, (b) temporal stability/ test-retest reliability of 0.40 or higher, or (c) inter-rater reliability (such as percentage agreement, correlation, or kappa) of 0.50 or higher.

- **Lack of over-alignment with the study intervention**—The measure must not be designed or administered in ways that are specifically aligned to an intervention so that the individuals receiving the intervention are being taught directly the content included in the outcome measure (e.g., a student should not be tested for reading fluency using the 50 words that she practiced reading aloud repeatedly during an intervention).

- **Consistent data collection across groups**—The outcome data must be collected using the same rules or procedures across groups of participants in the study (e.g., student outcome data should not be collected by special education teachers as part of their regular classroom activities in one school and by graduate research assistants in a pull-out activity in another school). If data are not collected consistently, the evaluator should try to ascertain to what degree the differences in data collection may contribute to differences in outcomes across groups.

Another characteristic of a high-quality measure is that it is **consistently defined across groups**. For example, if an evaluation is looking at the performance of transition specialists and wants to use rates of college enrollment among students who are deaf/hard of hearing (DHH) as an outcome, it is important to know whether the college enrollment rate for DHH students is calculated in similar ways across schools and school districts. If the outcome is not consistently defined across groups it is difficult for an evaluator to know to what extent those differences in definition may account for variation in outcomes. Consequently, when differences in definition exist, the evaluator should try to identify ways that those differences may influence the measurement of outcomes.

Of course, not all outcome measures are going to meet every criteria for quality presented above. However, evaluators should select outcome measures that fulfill as many of the quality criteria as possible. Table 3 presents a checklist developed for this Toolkit containing questions that evaluators can

[37] U.S. Department of Education, 2014.

ask when constructing outcomes and identifying measures. Evaluators are encouraged to carefully consider the impact of any items that receive a “no” or “somewhat” response on the evaluation’s ability to generate meaningful data regarding the project’s progress and results. These considerations can be summarized and reported as study limitations in the evaluation report. See Section 3.5.1 for a discussion of study limitations.

Table 3. Checklist for Constructing Outcomes

	Yes	No	Somewhat
1. Is the outcome informed by a need or problem statement?	☐	☐	☐
a. If yes, does the outcome reflect the specific nature of the need or problem?	☐	☐	☐
b. If no, has the same or a similar outcome been used in related programs or research?	☐	☐	☐
c. If no, are you using a recommended or required program outcome?	☐	☐	☐
2. Is the outcome relevant and of interest for program stakeholders and decision-makers?	☐	☐	☐
3. Is the outcome supported in related research or in evidence-based practices?	☐	☐	☐
4. Has or will a Program or Advisory Team reviewed the outcome for clarity and relevance?	☐	☐	☐
5. Have you identified the data variables that will be needed to generate a finding for the outcome statement?	☐	☐	☐
6. If using data collected by schools, districts, or the state, do you have access to the data variables that are needed to generate a finding for the outcome statement?	☐	☐	☐
a. If yes, are sufficient data available to generate a dataset that responds to the outcome in full?	☐	☐	☐
7. If using data collected by the project team (i.e., a new data collection), have you identified all of the steps and personnel that will be necessary for data collection and entry?	☐	☐	☐
8. If using data collected by the project team, do data collection methods and instruments rely on validated and reliable tools and techniques, for the outcomes or constructs of interest?	☐	☐	☐
9. Do you or will you have the specific expertise or assistance you will need for data entry and/or coding?	☐	☐	☐
a. If no, do you or will you have the resources to obtain the expertise or assistance you will need?	☐	☐	☐
10. Do you or will you have the specific expertise or assistance you will need for data analysis?	☐	☐	☐
a. If no, do you or will you have the resources to obtain the expertise or assistance you will need?	☐	☐	☐
11. Are there sufficient time and resources to collect and aggregate data for			
a. Direct outcomes?	☐	☐	☐
b. Intermediate outcomes?	☐	☐	☐
c. Long-term outcomes?	☐	☐	☐
12. Do you have required permissions or Memoranda of Understanding (MOU) to collect data?	☐	☐	☐
a. If no, do you have a process for obtaining permission or MOUs to collect data?	☐	☐	☐

2.2.2. Selecting an Evaluation Design

The design of the evaluation has bearing on the ability to link the training provided by the PDP project to the performance of the program graduates. Evaluation designs fall into one of three categories: non-experimental, experimental, and quasi-experimental. The selection of the evaluation design depends on (a) the questions that the evaluation is trying to answer; (b) the resources available for data collection, management, and analysis; (c) the availability and feasibility of control groups; and (d) the availability of data to measure outcomes. [38]

When considering which evaluation design to choose, it is important to think about issues related to the validity of the study, since design choices have multiple consequences for validity. It is beyond the scope of this Toolkit to go into detail about the concept of validity, so we briefly mention two basic types: internal and external.[39]

Internal validity refers to whether the study results are due only to the manipulation of the independent variables in the study, or whether there are other confounding variables that might influence the study outcomes and which were not taken into consideration in the study. **External validity** refers to the extent to which a study's results can be generalized to persons, settings, treatments, and outcomes not directly included in the study.[40] Each type of validity is subject to different threats (see Appendix E) which may call into question the study results. Consequently, when designing and conducting the evaluation, to the extent possible, evaluators should do the following:

- Identify and study *plausible* threats to validity;
- Use design elements (e.g., additional pretest observations, additional control/comparison groups) to control for possible validity threats; and
- Rule out plausible alternative causal explanations for an effect.[41]

When identifying threats to validity in a study, evaluators should ask three critical questions:

- How would the threat apply in this case?
- Is there evidence that the threat is plausible rather than just possible?
- Does the threat operate in the same direction as the observed effect, so that it could partially or totally explain the observed findings?[42]

It is better if evaluators can anticipate validity threats before beginning implementation of the study. If the study team knows what the threats may be but cannot use design controls in the study, an option would be to try measuring the threat directly in the study to see if it is actually operating. If this is not

[38] By *control* we mean to include both experimental control groups and quasi-experimental comparison groups.
[39] For more information about validity see Kane (2001), Messick (1989) and Shadish, Cook & Campbell (2002).
[40] Dimitrov, 2010; Shadish, Cook & Campbell, 2002.
[41] Shadish et al., 2002.
[42] Shadish et al., 2002, p. 40.

possible, at a minimum, evaluators should outline the possible validity threats in the study limitations section of the evaluation report (see Section 3.5.1 for a discussion of study limitations).

In the next sections we briefly discuss the three categories of evaluation designs. Some methodological considerations associated with the different designs are presented in Section 4.1. For more information about research design, Shadish, Cook and Campbell (2002) is an excellent resource.

2.2.2.1. Randomized Experimental Designs

The distinguishing feature of **randomized experimental designs, or the randomized controlled trial (RCT)**, is that the researcher has control of treatment—control that can take many forms.[43] The main characteristics of RCT studies are:

- Manipulation of the independent variable;
- Randomized assignment of participants to treatment groups; and
- Controlling for possible confounding variables.[44]

Randomization can take two forms: **random selection** of individuals to participate in the study and **random assignment** of participants to treatment and control groups. A fully randomized study includes both random selection *and* random assignment.

Confounding variables are those that are correlated (either positively or negatively) with both the dependent and independent variable, thereby affecting the study's ability to clearly associate an intervention or project with an observed outcome. Two common types of confounds include:

- **When there is only one unit in one or both conditions (also called the n=1 problem)**—In this situation, for example, the treatment group may include PDP graduates placed into one district, while the control group includes PDP graduates from the same program placed into another district. The n=1 confound makes it impossible for the study team to know whether the outcomes associated with each group are related to the skills and knowledge of the graduates themselves or unobserved characteristics of the districts in which they were placed.

- **When the characteristics of the units in each group differ in systematic ways that are associated with the outcomes**—In this situation, for example, the academic performance of students with low incidence disabilities might be compared with the performance of students with high-incidence disabilities, making it impossible for the study team to know whether some characteristic of the students themselves is associated with better (or worse) outcomes, rather than the instructional strategies being used in the classroom.

In general, the RCT is considered the most rigorous and best suited design for making causal claims about the effects of a project, since it provides information related to the counterfactual (i.e., what

[43] Shadish et al., 2002.
[44] Dimitrov, 2010.

would have happened if the project had not been implemented) and includes a variety of controls for threats to validity.[45] RCTs also generate data that can inform the unique contribution of project activities or services to desired outcomes. However, the very controlled nature of RCTs raises doubts about the ability of researchers to generalize the results.[46] That is, RCTs have high levels of internal validity, but low levels of external validity.

Additionally, RCTs are quite difficult to implement in the social sciences, and particularly in education. First, it is often difficult for researchers to randomly assign individuals (or classrooms, schools, etc.) to treatment conditions, let alone to be able to randomly select individuals, classrooms, or schools to participate in a study. Second, even when there is no opposition to the conduct of an RCT, the time and resources required to successfully implement RCTs in a school setting often limit researchers' ability to carry out such studies.[47] The requirements for conducting an RCT are even more difficult to achieve in the context of special education because of the low number of students and the heterogeneity of this population. For these reasons, we anticipate that few PDP project evaluations will feature randomized experimental designs.

It is not necessary for a study to be fully randomized for it to benefit from some of the characteristics of experimental studies. That is, even if selection to participate in the study is not random, incorporating random assignment into the study design will help to control for some confounding variables. We recommend that PDP evaluators try to incorporate experiments into their evaluation designs whenever possible. Shadish, Cook and Campbell (2002) is an excellent resources for evaluators who are interested in utilizing experiments in their evaluation designs. In Section 4.1.1 we present additional considerations for evaluators conducting RCTs.

2.2.2.2. Quasi-Experimental Designs

The main difference between randomized experiments and **quasi-experimental designs (QEDs)** is that quasi-experiments do not feature random assignment of study participants to treatment conditions.[48] Instead, assignment is done by self-selection or by non-random assignment to treatment conditions. Nevertheless, in QEDs researchers still may have control over the following study elements:

- Selecting and scheduling measures;
- Execution of non-random assignment;
- Selection of the comparison group; and
- Treatment schedule.

[45] There are those who would the claim that RCTs represent the "gold standard" of research design in education, including Conrad & Conrad, 1994; Scriven, 2008; and Sullivan, 2011.
[46] Shadish et al., 2002.
[47] Dimitrov, 2010; Shadish et al., 2002.
[48] Shadish et al., 2002.

QEDs generally are easier to conduct than experiments, while still providing a measure of methodological rigor. However, QEDs provide less support for counterfactual inferences—that is, making inferences about what would have happened if the intervention or project had not been implemented—than RCTs since the lack of random assignment to groups means that the treatment groups may differ in systematic ways that may affect the outcome.[49, 50] Consequently, researchers conducting QEDs should outline as many *plausible* alternative explanations for the study results as possible "*and then use logic, design, and measurement to assess whether each one is operating in a way that might explain any observed effect.*"[51] Of course, this has an impact on the complexity of the study design and, by extension, the difficulty of study implementation, so the study team will need to decide whether ruling out a plausible alternative explanation is worth the time, money, and effort required. Any plausible alternative explanations that were not accounted for in the study design should be discussed in the study limitations section of the evaluation report (see Section 3.5.1 for more information about limitations).

When planning an evaluation with a QED, **evaluators should consider the following important points**:

- If the design involves a **comparison group**,
  - Evaluators should consider whether it would be possible to use stratification or **matching** to improve comparability among groups. In stratification, units are grouped in homogeneous sets (e.g., gender) that contain more units than the study has conditions. When matching, units with similar scores on a matching variable (e.g., school size, ethnicity) are grouped, so that treatment and comparison groups both have units with the same (or very similar) characteristics on the matching variable. See Section 4.1.2.7 for more information about matching.
  - Evaluators should try to determine whether the treatment and comparison groups are similar at baseline, also known as **calculating baseline equivalence** (see Section 4.1.2.8).

- If the evaluation involves comparing data from graduates or students across multiple schools or districts—meaning **that data will be collected at *multiple levels*** (student, school and/or district)—we recommend that evaluators consider accounting for school or districts effects in the data analysis. This is known as **multilevel analysis** and is briefly discussed in Section 4.4.2.3.

In Section 4.1.2 we highlight a few quasi-experimental designs that we believe can be readily applied in PDP project evaluations.[52]

[49] Dimitrov, 2010; Shadish et al., 2002.
[50] It is important to point out that even with random assignment it is possible for the treatment groups to differ in systematic ways—a phenomenon referred to as "unhappy randomization"—but it is less likely when group assignment is fully random. Further, it is possible to minimize the risk of unhappy randomization by matching study participants on key characteristics and then conducting randomization (Dimitrov, personal communication, 2009).
[51] Shadish et al., 2002, p. 14.
[52] See Shadish et al., 2002, for additional discussion of these designs.

2.2.2.3. Non-Experimental Designs

Non-experimental designs include case studies, descriptive studies or surveys, correlational studies, and ex post facto studies (i.e., studies that take place *after the fact* using secondary data). It is possible to investigate a presumed cause and effect (e.g., which factors in a PDP project contribute to improved graduate performance) in a non-experimental study, but the structural features of experiments that help to rule out possible alternative explanations and identify the counterfactual—that is, information about what would have happened if a particular intervention or project had not been implemented—are often missing. Non-experimental designs generally are considered to be more appropriate for formative evaluations or monitoring the progress of (or fidelity to) project implementation than for summative evaluations. However, sometimes non-experimental studies are the only viable option for evaluators—especially if the evaluation was not planned prior to beginning implementation of the project or if the evaluator has little control over events during a study. Section 4.1.3 includes information on the different non-experimental designs that might be used in a PDP project evaluation. Nevertheless, whenever possible, we recommend that evaluators of PDP projects use either experimental or quasi-experimental designs to answer their summative evaluation questions.

2.2.2.4. Mixed-Method Designs

At the most basic level, mixed-method designs are those that combine quantitative and qualitative data collection and analysis approaches.[53] An entire field of literature has developed related to the nature of mixed methods versus mixed methodology[54] (that is, in general, mixed methods studies combine different data collection and analysis methods while mixed methodology studies combine different theoretical approaches as well as data collection and analysis methods), but it is beyond the scope of this Toolkit to discuss it. For our purposes here, we focus on mixed method studies, in which "the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative approaches or methods in a single study or a program of inquiry."[55] Specifically, we list four decisions that relate to the selection of a mixed methods approach to a study:

1. What is the implementation sequence of the quantitative and qualitative data collection in the proposed study?
2. What priority will be given to the quantitative and qualitative data collection and analysis?
3. At what stage in the [study] will the quantitative and qualitative data and findings be integrated?

[53] See, for example, Brewer & Hunter 2006; Greene, 2006; Morgan, 2007; Tashakkori & Creswell, 2007; and Tashakkori & Teddlie, 1998, 2003.
[54] Tashakkori & Creswell, 2007, p. 4.
[55] Creswell, 2003, p. 211.

4. Will an overall theoretical perspective (e.g., gender, race/ethnicity, lifestyle, class) be used in the study?[56]

For more information on the design and conduct of a mixed method study, see Creswell, 2003; Creswell, Plano Clark, Gutmann & Hanson, 2003; Brewer & Hunter, 2006; and Tashakkori, & Teddlie, 1998.

The following sections provide a brief overview of the steps involved in further developing the evaluation plan. First, we present considerations for creating a data analysis plan, followed by issues and considerations for using and selecting a sample and planning for data collection.

2.2.3. Creating a Data Analysis Plan

An essential part of developing the evaluation plan is creating a data analysis plan. Creating a data analysis plan prior to the data collection period will help ensure that the evaluator is able to collect and analyze data that will respond to the evaluation questions in the most rigorous way possible. Moreover, it ensures that (a) the instrumentation chosen or developed for the evaluation will gather the needed data in the correct format or scale, and (b) sufficient numbers and types of respondents or data sources will be included in data collection. For example, if one of the evaluation's analytic goals is to describe how much time the PDP graduates spend on a particular classroom activity, the evaluator will need to consider whether to report results using a checklist (e.g., Daily, Weekly, Monthly) or in units of time (e.g., minutes/day, hours/week, days/month). If, however, an analytic goal is to meaningfully compare the amount of time different types of program graduates spend on classroom activities, the evaluator not only needs to consider the type of scale to use and the format of the data but also the types of graduates that will be compared and, potentially, the minimum number of respondents for each type (i.e., total sample size and sampling strategy). This will help the evaluator to capture high-quality data, a sufficient quantity of data, and data that have sufficient generalizability.

Answering these questions requires the evaluator to identify the *unit* that will be subject to data analysis. Identification of the *unit of analysis*[57] is informed by the language of the outcome measure and evaluation question as well as the availability of data for analysis. Examples of different units are provided by Patton, who states "*each unit of analysis* [e.g., preparation program, graduation cohort, special education classroom, individual child] *implies a different kind of data collection, a different focus for the analysis of data, and a different level at which statements about findings and conclusions would be made.*"[58] Decisions about the unit of the analysis and the corresponding data collection and analysis strategy must be made during the development of the evaluation plan.

[56] Creswell, Plano Clark, Gutmann & Hanson, 2003, cited in Creswell, 2003, p. 211.
[57] The unit of analysis (i.e., the level at which outcomes will be analyzed) is different from the unit of selection (i.e., the level at which units were put into treatment or comparison groups).
[58] 2002, p. 228.

Table 4, on the next page, presents a data analysis plan template created for this Toolkit. As can be seen in the table, the data analysis plan includes information related to the:

- Study design (see Section 2.2.2 and Section 4.1)
- Treatment and control (or comparison) groups
- Type of data analysis (see Section 4.4 for a general discussion of data analysis)
- Variables to be used for quantitative analyses (see Section 4.4.2 for a discussion of quantitative analysis)
- Instruments and data collection techniques (see Section 4.3 for a discussion of data collection methods)
- Sample (see Section 4.2 for a discussion of sampling)
- Minimum number of responses and/or response rate

Since it allows the evaluator to specify a different analysis approach for each evaluation question, the data analysis plan template presented below can be used either for a single method study (e.g., a quantitative study), or a mixed-method study.

Table 4. Data Analysis Plan Template

Evaluation Question	Design		Data Analysis	Necessary Variables for Quantitative Analyses		Variable Sources (instruments or data collection techniques)	Data Sources	Minimum number of responses and/or response rate
	Design Type	If experimental or quasi-experimental, who constitutes the…		Statistical Tests	Descriptive statistics			
(1)	□Experimental □Quasi-experimental □Non-experimental	Treatment group: Control or comparison group:	□Statistical tests □Descriptive statistics □Qualitative analysis	Dependent: Independent: Covariates:	□Frequency □Mean		□Census □Sample (indicate sampling framework)	
(2)	□Experimental □Quasi-experimental □Non-experimental	Treatment group: Control or comparison group:	□Statistical tests □Descriptive statistics □Qualitative analysis	Dependent: Independent: Covariates:	□Frequency □Mean		□Census □Sample (indicate sampling framework)	
(3)	□Experimental □Quasi-experimental □Non-experimental	Treatment group: Control or comparison group:	□Statistical tests □Descriptive statistics □Qualitative analysis	Dependent: Independent: Covariates:	□Frequency □Mean		□Census □Sample (indicate sampling framework)	

When developing a data analysis plan it also may be helpful to create table shells (i.e., empty tables that illustrate how results will be displayed) to help identify the specific variables that will be needed and visualize how to format the data for reporting. Examples of table shells follow.

Table 5. Table Shell Example 1. Amount of time spent on the targeted classroom activity among program graduates

The purpose of this table is to describe how much time graduates spend on a particular classroom activity, reported in units of time (e.g., minutes/day; hours/week; days/month)

Amount of time spent on the targeted classroom activity among program graduates (n=)
Average amount of time in minutes per day
Standard Deviation
Range
Average amount of time in hours per week
Standard Deviation
Range
Average amount of time in days per month
Standard Deviation
Range

Table 6. Table Shell Example 2. Amount of time spent on the targeted classroom activity, by type of program graduate

The purpose of this table is to descriptively compare the amount of time, in different *units of time*, that different types of graduates spend on classroom activities

Amount of time spent on the targeted classroom activity, by type of program graduate	
Treatment Group (n=)	**Control Group (n=)**
Average amount of time in minutes per day	
Standard Deviation	
Range	
Average amount of time in hours per week	
Standard Deviation	
Range	
Average amount of time in days per month	
Standard Deviation	
Range	

Table 7. Table Shell Example 3. Testing the Statistical Significant of Differences between Treatment and Control Groups.

The purpose of this table is to display the results of *t*-tests of differences between means of the treatment and control groups.

	Treatment n =		Control n =				95% CI	
	M	*SD*	*M*	*SD*	*t(df)*	*p*	*LL*	*UL*
Minutes per day								
Hours per week								
Days per month								

Note. CI = confidence Interval; LL = lower limit; UL = upper limit

As can be seen, in the first example (Table 5) the table shell shows that the study will need to collect or derive data from evaluation participants on three types of variables: minutes per day, hours per week, and days per month that participants spent on the target activity over the period of time that was specified. Thus, in example one, if the study wants to report on all three types of variables, evaluators will need to collect data for at least one month. Additional months may be necessary to ensure the data are not sensitive to seasonal differences in classroom schedules (e.g., holiday breaks) and may help to create a more stable estimate of the average time spent on the activities.

The second example, presented in Table 6, posits that program graduates can be grouped according to whether or not they received a “treatment,” perhaps specialized training in the targeted classroom activity. Table 6 illustrates that in order to fulfill the data needs for this example it will be necessary to collect the same data that were presented in Table 5 for both the treatment and control groups. Further, if the study uses sampling to compose the treatment and control groups, it will be necessary to select sufficient numbers of both treatment and control group members if a goal is to generalize the study findings to the larger group of all program graduates. Finally, Table 7 illustrates the results of independent samples *t*-tests comparing the graduates in the treatment and control groups.

2.2.4. Identifying a Study Sample

While conducting a census of all graduates (or all students of graduates) of a PDP project is technically possible because the entire population is potentially known to the evaluation team, it is sometimes not feasible or even particularly desirable. Imagine, for example, an evaluation of a PDP project with 200 or more graduates—conducting a survey of all 200+ graduates would be time-consuming, costly, and likely to result in a low response rate. Selecting a sample enables the study team to save time and money, and, if done correctly, can help to improve the quality and accuracy of the data collected.[59] This becomes even more relevant if the evaluation plan includes use of a comparison group, or when the evaluation seeks to collect data on the performance of students of graduates where the numbers may reach 1,000 or more.

[59] Fiore et al., 2012.

Some considerations when making decisions about how to select a sample include:

- **Available resources**—How much time and money can be spent? How many people are available to work on the study (e.g., to recruit, collect data, analyze data)? Is a census cost-prohibitive?
- **Desired precision of estimates** –What is the minimum sample size needed in order to reach conclusions with a pre-specified level of confidence (see Section 4.2.1 for a brief discussion of power analysis)? Will results be used to make comparisons between groups?

Sampling is one aspect of an evaluation where a team member with specific training, expertise, or experience is needed, as there are multiple technical details that must be accommodated when the evaluation uses a sample, including how to analyze the data that is collected using sample weights. There also are many strategies for designing a sampling framework and selecting a sample. Section 4.2 presents additional information about two basic types of sampling—random sampling and purposeful sampling—as well as a brief discussion on power analysis. For additional resources on sampling methodology, see the recommended readings in Appendix C.

2.2.5. Preparing a Data Collection Plan

The final step in preparing the evaluation plan is developing the overall data collection plan and timeline. This entails identifying the specific tasks that will need to be completed to ensure success in collecting the data needed for the evaluation. The CIPP Evaluation Plan Template in Appendix A.4 presents a number of tables that can be used as part of a data collection plan. Rather than limiting our discussion to those specific tables, however, here we have outlined a series of questions to guide evaluators' thinking while developing a data collection plan. Then evaluators can choose the format for the plan that best suits their needs.

2.2.5.1. What instruments or data collection techniques will supply the variables that are needed?

It may be possible to collect some of the needed data using existing data sources, or "secondary source" data. Some publicly available databases are readily accessible online. These databases may allow users to download data files simply with the click of a mouse, or they may require users to sign a Data Usage Agreement. For example, researchers can easily download Excel files of student test results on the Massachusetts Department of Elementary and Secondary Education website (for data on the participation of students with IEPs in 2012 MCAS ELA and Mathematics Tests in MA, see http://www.doe.mass.edu/mcas/2011/results/IEPparticipation.xls). The U.S. Department of Education Common Core of Data (http://nces.ed.gov/ccd/) is an example of a publicly available database that requires users to sign a Data Usage Agreement before downloading the data files.

Secondary data also may be available through state or local education agencies. In some cases it will be necessary for the evaluators to complete a formal data request; in others they will need to work with the local research department or Institutional Review Board (IRB) to obtain access to the data (see Section 3.1.1 for more information).

If the data that are needed are not available as secondary data, and the study team decides that these data are essential to the evaluation, it will be necessary to conduct a unique data collection. In this case, it may be possible to use an existing data collection instrument that another individual, group, or publisher has developed and is making available. However, it is possible that existing instruments will not supply the specific variables that are needed to respond to the specific evaluation questions. In these cases, the evaluator will need to develop a data collection instrument or instruments to capture the specific data variables that are identified in the data analysis plan. We address this briefly in the next section and in more detail in Section 3.3 and Section 4.3.

2.2.5.2. What types of instrumentation or forms need to be identified or developed?

Data collection instruments come in many formats or types. Some of the most popular are surveys, observation checklists and rubrics, multiple choice tests, and standardized assessments of knowledge and skills. Other types include individual interview and focus group protocols, case notes and case management logs, and tracking logs. Data also may be collected through audio or video data recorders. For some of these, the evaluator will have a choice between instruments that already have been developed and, ideally, tested, so as to establish the validity and reliability of the instrument. In other cases, it will be necessary to develop the instrument, pilot test it, and finalize it for the evaluation's unique data collection activities. Instrument choice or development is guided by

- the nature of data and individual variables that are needed to respond to evaluation questions;
- the opportunities the study team will have to collect data; and
- the cost of data collection, including costs related to either buying or developing the instrument.

Some online resources on selection and development of data collection instruments include:

- Web Center for Social Research Methods *Research Methods Knowledge Base* (http://www.socialresearchmethods.net/kb/measure.php)
- The National Center for Education Statistics *Standards for Education Data Collection and Reporting* (http://nces.ed.gov/pubs92/92022.pdf)
- United States Census Bureau guidelines on developing data collection instruments and supporting materials (http://www.census.gov/about/policies/quality/standards/standarda2.html)
- Compendium of Student, Teacher, and Classroom Measures Used in NCEE Evaluations of Educational Interventions (http://ies.ed.gov/ncee/pubs/20104012/)

The CIPP Evaluation Plan Template in Appendix A.4 includes an Instrument Information table that can be used to monitor the status of the different instruments that will be used in the evaluation. Later in this Toolkit we provide information about methodological considerations for the development of different data collection instruments, such as surveys (Section 4.3.1), observation protocols (Section 4.3.2), and interview protocols (Section 4.3.3). Additionally, Appendix C includes good resources for learning more about instrument development.

2.2.5.3. How will data collectors and coders be trained? What materials need to be developed to document and support that training?

All data collectors and coders that will be supporting the evaluation must be trained in the data collection and coding protocols. The number of individuals needed for data collection and coding and the nature of the training will depend upon the evaluation design and on the amount, timing, type, and frequency of data collections. For example, for an evaluation supported entirely with quantitative secondary data, it may be necessary to train only one individual in the retrieval of all data. For a study that includes classroom observations or open-ended surveys of program graduates, on the other hand, multiple observers and multiple individuals may need to be trained to code the qualitative data that emerge from the surveys. A project in which the data need to be collected from multiple individuals over a very short time frame may necessitate multiple data collectors, to ensure data are collected within the prescribed time limit. In contrast, a project in which data can be collected over a long period of time may require that relatively few data collectors or coders operate concurrently. Training can take many forms, including in-person training, online training, or development of training modules or manuals. The time associated with training data collectors and coders should be factored into the data collection plan. We discuss training further in Section 4.3.2.2.

In those cases in which data collection is limited to gathering secondary data, training may entail an introduction to the source database and the instruments, techniques used to populate the database, and the methods for retrieving the data. In those cases where unique data are collected, training should entail an introduction to the study and evaluation questions; a thorough review of the instruments; a review of data collection rules, procedures, and timelines; and, if possible, practice runs to establish familiarity with data collection. Further, when there are multiple data collectors, it is important to establish the inter-rater reliability of data collections (see Section 4.3.2.6 for a discussion of this), so as to ensure that the complete dataset represents standardized and uniform data collection procedures.

When data are collected through interviews, focus groups, or audio or video recordings, the study team also may work with data coders who review the data and categorize or code the text, audio, or video files, but are separate from the actual data collection. In these cases, background materials and training should be developed for coders, and the evaluation team should establish reliability and inter-rater reliability across coders.

Data collectors also should receive training on human subject protections, such as those codified in the "Common Rule", or the "Federal Policy for the Protection of Human Research Subjects" (1991). The complete set of policies and regulations associated with the Common Rule can be reviewed at http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html. The Education Department's policies and regulations are available at http://www2.ed.gov/about/offices/list/ocfo/humansub.html. Other policies that commonly are referenced include the Health Insurance Portability and Accountability Act of 1996[60] (HIPAA) and the Family Educational Rights and Privacy Act (FERPA).[61] Most universities and colleges maintain resources related to human subject protections; training may be required for all university faculty and staff that conduct data collections. Some private organizations will provide training and support for human subject protections (usually for a fee) if it is not available to the study team otherwise.[62]

Some instrument developers require that all data collectors receive formal training in the use of their measure; the School Observation Measure (SOM) is an example of such an instrument (http://edit.educ.ttu.edu/site/jcheon/manual/SOM_manual.pdf). To become certified in the use of the SOM, all data collectors must participate in the developer's one-day training session. This can take place in-person or over a video conference. More commonly, instrument developers provide information or training manuals about how to use an instrument, but do not require users to obtain formal training (e.g., the Reformed Teaching Observation Protocol, http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Chemistry%20Assessments/RTOP%20Reference%20Manual.pdf).

2.2.5.4. When will data be collected? How frequently will data be collected?

Large evaluations that consist of multiple evaluation questions likely will be active throughout the year conducting data collection, data entry, analysis, or reporting activities. Small evaluations may have shorter, targeted data collection periods. Regardless, the key is to ensure that all necessary activities are conducted on schedule and in the right order. Being well organized is of particular importance for data collection activities, as they often are time-specific; once a window for data collection closes, the study team may not be able to capture the necessary data, depending on the specific constraints or parameters of the study. Failure to meet schedules can, at a minimum, affect data analysis and the resulting interpretations and lead to limitations in the study findings. It also is possible that failure to collect data on schedule may result in an inability to draw a conclusion or respond to one or more evaluation questions (Section 3.3 discusses managing data collection in more detail).

2.2.5.5. How will data be entered and verified for accuracy? Where will data be stored?

Once data have been collected, it will need to be reviewed and made available for analysis. Generally speaking, the raw data that are collected are not "analysis ready." Rather, data need be checked for quality and often need to be entered into an electronic database, spreadsheet, or statistical package for data analysis. Data entry serves the secondary purpose of maintaining an electronic file of all data collected—as a supplement to the original raw data files.

[60] http://www.hhs.gov/ocr/privacy/
[61] http://www.ed.gov/policy/gen/guid/fpco/ferpa/index.html
[62] One such company is Ethical and Independent Review Services, http://www.eandireview.com/ (NOTE: By providing a link to this website, Westat is not advocating the use of this company.)

Data entry requires data technicians or support staff, all of whom should receive training and periodic "spot checks" to ensure accuracy. A senior-level staff person often will develop the data entry files and coding procedures and then work with data technicians to train them and review their data entry.

A persistent topic in data analysis is what to do about missing data and about responses that are not what the study team were expecting (e.g., outliers; see Section 4.4.1 for a discussion of missing data and Section 3.3.2 for a definition of outliers). A senior-level staff person often will work with data technicians on how to code or flag these fields for later treatment.

2.2.5.6. Timeline of Evaluation Activities

After answering the questions outlined above, the next step is to prepare a timeline for the evaluation that will enable the study team to monitor evaluation activities and ensure that the evaluation is on schedule. As mentioned above, the CIPP Evaluation Plan Template in Appendix A.4 includes tables that can be used as part of a data collection plan. The Gantt chart is another commonly-used organizational tool for planning evaluation projects. Gantt charts provide a general timeline for each of the evaluation's major activities. Table 8 presents an example of a Gantt chart of a hypothetical data collection schedule.

Table 8. Gantt Chart of the Project's Data Collection Schedule

Task	Time Units (Days, Weeks, Months, Years, etc.)											
	1	2	3	4	5	6	7	8	9	10	11	12
Develop logic model												
Develop evaluation plan												
Develop analysis plan												
Prepare data collection instruments												
Complete IRB process												
Secure district participation												
Prepare training materials												
Conduct data collector training												
Conduct data collection												
Enter and clean data												
Conduct coder training												
Code data												
Analysis and reporting												

Now that we have discussed the steps involved in planning the evaluation, it is time to turn to the different aspects of conducting the evaluation.

3. Conducting the Evaluation

In this section we discuss various steps involved in conducting the evaluation. We would like to point out that even though these steps are presented in a specific order in this section, many of the steps actually will take place concurrently and others may occur in a different order than presented here. We have simply ordered them in a way that seems to make sense within the structure of this Toolkit.

3.1. Obtaining Permission to Carry Out Evaluation Activities

Prior to beginning the evaluation, the study team must obtain permission to carry out the evaluation activities in the participating districts and schools. Evaluators affiliated with a university or college likely must obtain approval from the institution's Institutional Review Board (IRB)—the group responsible for reviewing research to assure the protection of the rights and welfare of the human subjects. Large research organizations have their own IRB responsible for approving evaluation activities. Smaller evaluation companies often do not have their own IRB, but if the evaluation design involves collecting data from students, their parents/guardians, or school personnel, at a minimum the study team will need to obtain permission from the local school district. Furthermore, if secondary data will be used in the evaluation it may be necessary to obtain permission to access the data. These various types of permissions are briefly discussed below.

3.1.1. Getting IRB Approval

It is reasonable to expect that the evaluation will require collection of data on individuals involved in or affected by the PDP project (i.e., the "treatment" population). Further, experimental or quasi-experimental evaluations also call for data collection from a control or comparison group (see Sections 2.2.2, 4.1.1 and 4.1.2 for more details on those types of evaluation designs). Therefore, it is likely that evaluators will have to seek Institutional Review Board (IRB) approval for data collection, since human subjects are granted protections from data collections that may be harmful either at the time of data collection or, if foreseeable, at a future point in time (see http://www2.ed.gov/about/offices/list/ocfo/humansub.html for more information about the requirements for protection of human subjects in research).

Obtaining IRB approval involves describing the evaluation's approach to data collection and detailing any and all circumstances in which there is contact with a human subject (e.g., the student, parent, or teacher that is the subject of the data collection). Requesting and receiving IRB approval guarantees the study team has taken all of the necessary steps to ensure that human subjects are protected and that the research protocol discloses any risks associated with participating in the study. It is common for state and local education agencies to require IRB approval before allowing any unique data collections. Evaluators that are affiliated with an institution of higher education may have access to an IRB panel through their institution. Private, for-profit IRB panels may also be contracted to perform the review.

Typical fields in IRB applications include:

- General information
- Objectives of proposed project
- Description of human participants
- Summary of research and data gathering procedures
- Location of project
- Confidentially safeguards
- Assessment of risk
- Consent procedures
- Potential benefits

In some cases the evaluation may be exempted from a full IRB review. The IRB may determine a research activity to be exempt from the need for IRB review when the only involvement of human subjects will be in one or more of the following categories:

A. Research conducted in established or commonly accepted educational settings, involving normal educational practices, such as:
    1. research on regular and special education instructional strategies, or
    2. research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.
B. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless:
    1. information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and
    2. any disclosure of the human subjects responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.
C. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior that is not exempt under paragraph (B) of this policy, if:
    1. the human subjects are elected or appointed public officials or candidates for public office; or
    2. Federal statutes require without exception that the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter.

Even if it seems the evaluation may be exempt from a full review, most IRBs still require the evaluator to submit the information listed above to receive approval for the exemption. Evaluators should always refer to their institution's (or the district's) specific rules and guidelines for obtaining IRB approval.

3.1.2. Securing District and School Approval

Once the study team has identified the graduates who will be participating in the study (see Section 3.2.1 and Appendix D for more information on keeping track of graduates), it will be necessary to obtain approval from the corresponding school districts and schools, informing them of the details of the evaluation and obtaining all necessary permissions. Many districts have their own research departments and established protocols for conducting research in the district, in which case the evaluator should contact the district research office to review its requirements. Often this information is available online (see for example the Fairfax County, VA, Office of Program Evaluation: http://www.fcps.edu/pla/ope/research_a/research_approval_process.shtml). Typical procedures require the submission of a research proposal that is reviewed by the district's research board. Additionally, some districts have requirements about what times of the academic year data can be collected (e.g., data collection may not be permitted during the district's mandatory testing periods). If a formal research approval application is not required, the study team still should notify the district administration of the evaluation and what it will entail. Appendix A.6 presents an example of a District Notification Form that might be used for a PDP project evaluation.

In general, once a district has given approval for a study to take place, school approval will follow. This is not always the case, however. School principals always must be informed of any data collection activities that will take place in their schools (e.g., surveys, observations, or interviews). Appendix A.7 presents an example of a School Notification Form. If district approval included an assigned study number, this should be included in the School Notification Form. To avoid potential problems related to questions about whether the study team obtained the appropriate permissions, it is good practice to keep records of all contacts with districts and schools, including the date, time, and person contacted. This information can be maintained in the evaluation's data tracking system (see Section 3.3.1 and Appendix D for more information on data tracking systems).

3.1.3. Obtaining Access to Secondary Data

If the evaluation intends to use secondary data it may be necessary for the evaluator to complete a Data Usage Agreement. This agreement simply may require the evaluator to agree to follow the terms of use of the data, or it may provide more specific data usage instructions, such as the specific variables the evaluator will have access to, how the data will be transmitted to the study team (e.g., spreadsheet file, database, secure file sharing protocol), and any limits that are dictated by state or local district policy on confidentiality and protection of human subjects. A data sharing agreement generally is formed after the evaluator has (successfully) applied to a state or local education agency to conduct research (see above for a discussion of obtaining district permission to conduct the study).

3.2. Recruiting Study Participants

Once permission to conduct the evaluation has been obtained, the next step is to recruit participants. Since evaluation activities almost always are voluntary—even for the population that is receiving the service or program being evaluated—the study team will need to develop strategies for engaging and recruiting participants for the study. Two essential aspects of recruiting participants include keeping track of the graduates and obtaining consent from participants. We discuss these aspects below, following a brief presentation of some considerations for communicating with participants.

Listed below are some guidelines evaluators should follow when communicating with participants:

- Clearly state what the evaluation will be asking the participant to do, preferable at the start of the communication (e.g., "We are asking you to complete the enclosed survey…").
- Provide a deadline for a response (e.g., "Please complete the survey no later than DATE").
- Explain the overall purpose of the study. A pamphlet or short document for the participant to review and keep may be helpful.
- Use clear, concise language. Compose written materials that will be sent to parents (e.g., consent forms) at an 8th grade reading level (for information on how to check the reading level of text in Microsoft Word, see https://support.office.com/en-us/article/Test-your-document-s-readability-0adc0e9a-b3fb-4bde-85f4-c9e88926c6aa).
- Always provide a telephone number and email address for a contact person on the study team so participants can obtain additional information.
- Include self-addressed return envelopes if participants must return hard copy materials.
- For online surveys, include instructions for obtaining usernames and passwords. If providing a username and password to participants, and the information to be collected is sensitive, do not include the username and password in the same communication. This is necessary for ensuring confidentiality and protection of participant data.
- When parents or children who are not native English speakers are involved, communication should occur in the participant's primary language, whenever possible. It may be helpful to ask a project staff person with whom the target participants are familiar and comfortable to help communicate the evaluation's plans and importance. In some cases, the evaluator may need to enlist the assistance of an interpreter or translator. For instance, if recruitment letters are being sent to a population with a significant percentage of non-native English speakers, it could be

helpful to have the materials translated and send each potential participant materials in two languages.[63]

Another strategy for recruiting participants to the study may be the provision of incentives or awards in return for participation in data collection. While this is a common practice, it is important to consider whether local regulations would allow the offering of such an incentive (e.g., some school districts prohibit offering of incentives to district staff). If an incentive is acceptable, the evaluator should determine the level of incentive or award that will be necessary to recruit a sufficient sample size. Of course, the number and types of participants (e.g., graduates, graduates' supervisors, students) to be recruited will depend on the evaluation design and the type of sample chosen (see Section 4.1 for information on different types of evaluation designs and Section 4.2 for information on sampling). For example, if the evaluation calls for a longitudinal study (i.e., an evaluation in which there will be multiple data collections from the same individuals over a period of time), the evaluator will need to consider the level of incentive or award necessary to maintain participants over the entire period of the study. Also, in the case of a longitudinal study, it is reasonable to expect attrition of participants over time (i.e., participants discontinuing their participation in the study), and researchers often develop larger-than-necessary samples to ensure that a sufficient number of participants will be available at the last planned data collection.

3.2.1. Keeping Track of PDP Graduates

Maintaining contact with PDP project graduates is essential to the success of the evaluation. The best way to achieve this is to incorporate into the PDP project strategies to maintain on-going contact with graduates (e.g., by having faculty mentors for graduates during their first post-program year, or by using social media sites to continuously engage graduates). Whether or not these strategies are successful, the evaluator will need to work with project staff and administrators to gather as much data as possible about the graduates, before program completion, from administrative records and then reach out to the graduates to request their participation in the evaluation.

Some steps that the study team can take to keep track of graduates include:

- collecting alternate contact information (e.g., parent, sibling) from scholars while they are in the program and immediately prior to the graduation date;
- establishing a web-based portal (e.g., Moodle or SharePoint site) for scholars and graduates to update their information and to communicate in other ways with faculty or each other;
- working with the institution's alumni relations office to reach out to graduates; and

[63] Interpreters are used when the communication will be in-person, while translators are used when the communication will be in written form.

- maintaining a secure database with contact information and ongoing records of contacts (e.g., emails, telephone calls) with graduates and alternate contacts, including the initial contact and any follow-up contacts.

Section 3.3.1 and Appendix D present additional information about the types of information that should be collected as part of the evaluation's data tracking system. Since it may be difficult to select a random sample of graduates for the evaluation or to keep up with all PDP graduates, in Section 4.2.3 we present a number of purposeful sampling options. However, it is important to consider the effects that selection bias can have on the evaluation, especially if a convenient sample of graduates who already are keeping in contact with the funded institution is selected.

3.2.2. Obtaining Participants' Consent

Any evaluation that plans to interact with human subjects must obtain consent from the participants prior to beginning the data collection activities (see http://www2.ed.gov/about/offices/list/ocfo/humansub.html for more information about the protection of human subjects in research). Studies that include participants who are considered children must obtain consent from the children's parents or legal guardians.[64] In general, there are two forms of consent:

- **Passive consent:** Assumes that the participant has given consent unless action, such as a written statement, is taken to indicate otherwise. Passive consent often involves the distribution of a letter explaining the study to parents and informing the parent that they should return the signed letter to the school or study team if they do *NOT* want their child to participate in the study (see Appendix A.9 for an example of a passive consent form). In many study contexts, passive consent is assumed when an adult agrees to participate in data collection activities such as completing an online survey or telephone interview.

- **Active consent:** Requires the participant to provide a written signature for consent. This is often required when children are participants in a study. In this case, the child cannot be involved in data collection unless a written signature from the parent is provided to the study team (see Appendix A.10 for an example of an active consent form).

Originating in the health services and medical fields, the term "consent" represents the concept of providing information to an individual who is capable of understanding the information and making an informed and judicious decision to either participate or not participate in the study. A similar standard applies in educational research: the consent giver must be provided with sufficient information in a way that allows the consent giver to understand the study, be aware of any risks participation entails, and make an informed and judicious decision about participating. In practice, this means that consent forms must be written in clear and simple language. The evaluator also bears the burden of using means of communication that are accessible by the participant or consent giver and answering any questions the

[64] For simplicity, we will only use the term "parents" in this section.

participant may have regarding the evaluation and his or her participation in data collection. Finally, it is important to stress that participation is voluntary; the participant can withdraw from p out punishment (see Section 3.2 above for a brief discussion of considerations for communicating participation at any time with participants).

In addition to obtaining informed consent from parents, researchers might need to obtain verbal or written assent from child participants. Assent differs from informed consent because it does not imply an understanding of the purposes, risks and benefits of research, but merely indicates a willingness to participate in the research activities. An IRB will make a decision about whether to judge participating children capable of providing assent by considering the age, maturity, and psychological state of the children involved. If participating children are judged to be capable of providing assent, the IRB will usually require the research protocol to "make adequate provision to seek assent from child participants".[65] The assent process can be as simple as asking a young child whether they are willing to "play a game" with the researcher. With older children, the assent process might involve an explanation of the research and a signature on a brief assent form.

3.3. Managing Data Collection

With the instruments chosen or designed, the data collection staff identified and trained, the participants recruited, and participants' consent obtained, the evaluation team is now ready to collect data. Rather than going into detail about the specific steps involved in data collection, we choose to highlight here two important strategies to manage the data collected during the evaluation: the creation of a data tracking system and the assessment of data quality. These are briefly discussed below. For more information about the methodological considerations associated with the different data collection methods, see Section 4.3.

3.3.1. Creating a Data Tracking System

Data tracking systems can be sophisticated web-based databases accessible to multiple users spread out across many sites or they can be basic spreadsheets operated and maintained by a single person. Whichever type of tracking system the evaluator chooses will depend on the size and complexity of the evaluation and the available resources. Appendix D includes the outline of a PowerPoint presentation that CIPP staff made at the 2012 OSEP Project Directors' Conference. Here, we briefly highlight some important features of that presentation.

In general, when making decisions about which type of tracking system the evaluation needs, the evaluator should ask four questions:

---

[65] See U.S. Department of Health and Human Services information on the definition of assent at http://www.hhs.gov/ohrp/policy/faq/children-research/child-assent-age-requirements.html and requirement for assent at http://www.hhs.gov/ohrp/policy/faq/informed-consent/requirements-for-assent-in-research-with-children.html

- What kind of database do I want to use? (e.g., a simple database such as Microsoft Excel, or a relational database such as Microsoft Access);
- How do I want to input the data? (e.g., manual entry by a project team member, or a web form linked to the database);
- How do I want to update the data? (e.g., point-in-time through manual entry, or real-time through web forms); and
- How do I want to analyze the data? (e.g., from within the database using queries and reports to generate descriptive analyses, or by exporting the data to a statistical analysis program such as SPSS/SAS/STATA to conduct descriptive and inferential analyses)

The answers to the questions outlined above will depend in part on the types of data that need to be included in the tracking system. For the evaluation of a PDP project, we recommend including the following types of data in the tracking system:

- Preparation program characteristics (e.g., institution name, program size, certifications offered, course requirements);
- Graduate characteristics (e.g., a unique ID number for each graduate, demographic info, contact information, and any other characteristics that might be important to include in the analysis);
- Graduate performance data (e.g., placement and retention, quality of services provided, and student achievement growth, if available);
- Comparable data for a graduate comparison group (if applicable); and
- Timelines for data collection activities (e.g., contacts with participants, administration of surveys or tests, conduct of observations)
- Data collection responsibilities (e.g., preparing the IRB application, developing data collection instruments, contacting graduates, contacting graduates' schools, obtaining consent, collecting data, inputting data into the database, updating data, preparing data for analysis, conducting data analysis, writing the report).

If the evaluation also includes measures of the performance of the program graduates' students—something ED and OSEP would encourage, whenever possible—the data system also should include characteristics of the graduates' students, student performance data, and data for a student comparison group (if applicable). In order to link the performance of the students with the program graduates, the evaluator will need to ensure that the data system includes a unique ID number for each student that can be linked to a graduate. Some districts and states have their own data systems that include these unique ID numbers; if they are not available, the evaluator can work with the district data manager to generate unique ID numbers, or can generate study ID numbers themselves. As previously mentioned, when obtaining access to student data, evaluators will need to obtain approval from the district and school, as well as from the students' parents (see Section 3.1.2 for more information). Further, when working out the data sharing agreement with the state or district (see Section 3.1.3), evaluators will need to be able to demonstrate that they have a secure way to transfer and store student data.

When creating a data tracking system, evaluators should keep in mind the following tips:

- Automatic reminders can facilitate timely data collection (e.g., the database generates emails to graduates after a certain time if they haven't entered their student data);
- Pull-down boxes within the database reduce risk of incorrect data entry (see http://office.microsoft.com/en-us/excel-help/create-or-remove-a-drop-down-list-HP010342357.aspx for more information on how to do this in Microsoft Excel); and
- Consider creating "checkpoints" to control data entry (e.g., person inputting data must check a box for "obtained consent" before being able to enter student data).

Another step in the management of data collection is the assessment of data quality, discussed below.

3.3.2. Assessing Data Quality

Before moving on to the analysis and reporting of the data that are collected, we recommend that evaluators review the quality of the data. This section relates primarily to quantitative data but the basic principle of assessing data quality applies to qualitative data as well (see Section 4.4.4 for more information on qualitative data analysis). To begin assessing data quality we encourage evaluators to think about each variable in the dataset and to examine the values entered for each variable for accuracy and precision. In short, there are several data characteristics to review in assessing data quality:

- **Completeness**—Whether or not there is a full response to the data question or prompt. For example, if the question asks for a respondent's complete date of birth, does the respondent provide the month, day, and year?
- **Consistency**—Whether or not data that measure similar or related constructs are in agreement. For example, if there is a question on a graduate survey that asks for a graduate's specific area of training and another that asks for his or her coursework, do the responses agree or make sense together?
- **Legitimacy**—Whether or not a data value is reasonable for a specific question or prompt. That is, are the data values logical given the context of the program and the specific question or data prompt? For example, if the question or prompt asks for respondent's gender, does the respondent indicate either male or female?[66] If a question or prompt requests the length of

[66] This assumes that the survey does not provide an option for an alternate way for respondents to identify their gender (e.g., transgender).

time a teacher has been working, do mathematical calculations confirm that a teacher hasn't been working longer than he or she has been alive?[67]

These types of edit checks help to verify the accuracy and quality of the data. We offer the following additional suggestions related to assessing the quality of the data.

- **Identify opportunities to check for data consistency and legitimacy and deploy the checks.**
  - Identify questions that ask for related data in order to check for data consistency. Data skip patterns should be included in these checks. For example, if a teacher indicates in Question 7 of a graduate survey that she does not teach students with autism, and Questions 8 and 9 ask further questions about teaching children with autism, there should be no data for Questions 8 and 9 for this teacher.
  - Identify the set of possible responses for each question. Review data frequencies to ensure all reported values fall within the set of possible responses. In some cases, the set of possible responses is readily apparent. For example, if the data item reports a date, evaluators should verify that the values for the date fields fall within acceptable limits (e.g., values for the "month" field have two digits between 01 and 12). In other cases, the evaluator's knowledge and expertise will inform decisions about the set of possible values and the identification of outliers. Note: in this latter case, a value that falls outside what would normally be considered a possible value should not be discounted or determined to be an outlier without further investigation.
- Decide how to handle missing data.
  - Is the amount of missing data extensive? If one respondent is associated with multiple missing values, does this warrant exclusion of all of the respondent's data? Will statistical techniques for imputing missing values be employed? (See Section 4.4.1 for more information on dealing with missing data).
- Calculate basic descriptive statistics for each variable and review data frequencies for outliers (see Section 4.4.2).
  - **Outliers** are data values that fall outside of the bulk of the data that were received. Outliers may reflect data entry errors or an extreme response reported by a respondent. In general, the study team should discuss rules for identifying and reacting to outliers. For example, evaluators may consider a value that is more than 20% away from the next closest data value to be an outlier. Will the evaluator keep the outlier in the dataset as a valid value? Will the evaluator choose statistical techniques that are relatively insensitive to the presence of outliers?

---

[67] This assumes that the evaluator has obtained the respondent's date of birth or age in order to conduct the calculations.

In assessing data quality, evaluators may need to identify a software package to help implement this plan. A number of options are available, including:

- Microsoft Excel: A product that includes a number of data validation options including being able to easily sort variable columns and establishing logic rules that highlight cells that do not meet specified criteria.

- Open Refine: A free online application that provides a user friendly interface with data editing features. The application is downloaded to a computer desktop so evaluators do not need to worry about sharing confidential data. The advantage to Open Refine is that it allows easy editing of data and keeps a log of edits. Tutorial videos are available to explain many of the features available: https://github.com/OpenRefine/OpenRefine

Finally, evaluators should establish rules that will be followed in cleaning the data and fully document those procedures. Data cleaning refers to the procedures that will be followed to make the data ready for analysis. For example, evaluators may want to code raw data into a defined set of values. Evaluators also may find that they need to transform some variables into new variables for the purpose of analysis. It is important to document all of these decisions and data transformations (and coding). Further, evaluators may find that data cleaning raises additional questions about specific data values or responses. In such cases, evaluators may be able to contact respondents to clarify responses or may elect to code some inconsistent data as missing.

3.4. Analyzing the Data

It is beyond the scope of this Toolkit to discuss the myriad methods of quantitative and qualitative data analysis. Section 4.4 outlines some methodological considerations that should be kept in mind when conducting data analysis. In this section we highlight some important aspects related to preparing the data and briefly discuss how to aggregate the data and report the results.

3.4.1. Preparing the Data for Analysis

In connection with the data collection and data quality review activities, evaluators will have entered the data into a database and begun preparing them for analysis. This process includes the following:

- Checking for duplicate records

- Identifying outliers

- Identifying the different types of measurement scales that are present in the quantitative data (e.g., nominal, ordinal, interval, ratio)

- Determining what types of variables the data represent (e.g., categorical, continuous)

- Assigning a numeric score to each response category for each close-ended question in a survey, item in a structured observation protocol, or question in a structured interview (e.g., 0 = Poor, 1 = Fair, 2 = Good, 3 = Excellent or 0 = No, 1 = Yes) (Note: Data obtained in unstructured observations or interviews will need to be analyzed qualitatively and codes may be developed; if desired, numbers can be assigned to the codes but generally they should be treated as nominal data.)

- Determining how to code missing data (e.g., assigning a code of 99 or 999 to missing data will cause a problem when calculating means)

- Recoding variables such as negatively worded survey items so they will be consistent with the positively worded item coding

- Reviewing qualitative data gathered during open-ended survey questions, qualitative observations and unstructured interviews for later analysis.(see Section 4.4.4 for information on qualitative data analysis)

Since data analysis is a complex topic, we do not cover it here (see Section 4.4 for information on data analysis). Instead, we limit our focus to creating the final variables for analysis, discussed next.

3.4.2. Aggregating Data and Reporting Results

The data analysis plan (see Section 2.2.3) will guide most of the work during the analysis and reporting phases of the evaluation. In particular, this plan will help the study team to organize the output into tables and charts to answer the evaluation questions. Remember, the analysis plan identifies the specific variables and analysis techniques that will be used to respond to each evaluation question.

Data aggregation is the set of procedures that evaluators will use to combine data from multiple respondents or multiple items in order to report study findings. The data aggregation steps will vary by each evaluation question and analysis approach.

3.4.2.1. Aggregating and Reporting Percentages

We generally report percentages if we are working with nominal or ordinal data. For example, if an evaluation question requires reporting the percent of a target population that achieves a specific benchmark or achievement, evaluators should complete the following aggregation steps:

1. Identify which responses are eligible for inclusion in the numerator and denominator. In particular, consider the following questions:
    - Are only those data values that pass data quality assessment checks to be included?
    - How will missing data be handled in the calculation?

2. Calculate the numerator.
   - Identify the range or set of values that qualify a data value for inclusion in the numerator, then calculate the numerator by *counting* the number of eligible responses in which the qualified data values exist.

3. Calculate the denominator.
   - Identify the range or set of values that qualify a data value for inclusion in the denominator, then calculate the denominator by *counting* the number of eligible responses.

4. Conduct the calculation by dividing the numerator by the denominator and multiplying by 100.

In reporting the results, evaluators should be sure to not only report the required percent, but also the total number included in the calculation as well as the number of responses that are *not* included because of the data quality assessment and data eligibility checks. If the results are limited to a subset of the target population, evaluators should include language that describes the true scope of the findings.

3.4.2.2. Aggregating and Reporting Means

We generally report means if we are working with interval or ratio data. For example, if an evaluation question requires reporting a mean value for a target population or benchmark, evaluators should complete the following aggregation:

1. Identify which responses are eligible for inclusion in the numerator and denominator. In particular, consider the following questions:
   - Are only those data values that pass data quality assessment checks to be included?
   - How will missing data be handled in the calculation?

2. Calculate the numerator.
   - Identify the range or set of values that qualify a data value for inclusion in the numerator, then calculate the numerator by *summing* the total eligible responses in which the qualified values exist.

3. Calculate the denominator.
   - Identify the range or set of values that qualify a data value for inclusion in the denominator, then calculate the denominator by *counting* the number of eligible responses.

4. Conduct the calculation by dividing the numerator by the denominator.

In reporting the results, evaluators should be sure to not only report the required mean, but also the range, the standard deviation, and the number of responses that are not included because of the data quality assessment and data eligibility checks. If the results are limited to a subset of the target population, evaluators should include language that describes the true scope of the findings.

3.4.2.3. Conducting and Reporting Results of Inferential Statistical Analyses

It is beyond the scope of this Toolkit to discuss the various methods of conducting and reporting inferential statistical analyses. We recommend that evaluators without statistical training consult a statistician to help with these analyses. Section 4.4.2.2 presents a brief discussion of inferential statistical analysis. Hinkle, Wiersma, and Jurs (2003) and Dimitrov (2010) are good reference books and Rice University, the University of Houston Clear Lake, and Tufts University have developed an online statistics book that is free and available to the public at http://onlinestatbook.com/2/index.html. Evaluators also can find information about statistical analysis online at the Web Center for Social Research Methods (http://www.socialresearchmethods.net/).

3.5. Preparing the Final Report

The last step in the evaluation is the preparation of the final report. The content of the report may be established in the request for applications to which the study team responded or in the request for proposals specifically for the evaluation. If not, we recommend including the following sections in the report to allow readers to fully review and interpret the findings and to allow for replication.

- Table of Contents
- List of abbreviations and acronyms
- Executive Summary—Usually contains a basic description of the program, the evaluation questions, and key findings.
- Introduction—Provides a more detailed background of the program and identifies the evaluation questions.
- Methodology—Describes in detail the evaluation plan, data collection activities, sample, and data analysis and data quality assessment plan for each component of the evaluation.
- Analysis and Results—Presents the findings for each evaluation question in narrative form, supplemented by tables and graphs, and summary statistics.
- Conclusions and Implications—Summarizes overarching findings and evaluation results.
- Study Limitations—Identifies limits of the data collection and analysis techniques that affect the generalizability and interpretation of findings. This is discussed in more detail below.

3.5.1. Outlining Study Limitations

It is important for evaluators to identify the limitations associated with their evaluation. Limitations are characteristics of the design or implementation that affect how the evaluation's findings should be interpreted and the extent to which the findings can be generalized to other programs or contexts. Identifying limitations should take place in both the planning and reporting phases of the evaluation. In the planning phase, evaluators should consider the implications of the *a priori* methodological choices.[68] In the reporting phase, evaluators should reflect on any unanticipated limitations that arose during the course of the evaluation. This process will alert the reader to the need to exercise caution when interpreting the results. Further, it will give information that may help the study team to improve the design and data collection and analysis techniques in future evaluations.

The brief survey in Box 1 below is designed to guide analysis of each phase of the evaluation to identify potential limitations. In discussing the findings of the evaluation, evaluators should describe these limitations and how they may affect interpretations of the study results. In the sections that follow we discuss a number of methodological considerations that PDP project evaluators should keep in mind throughout the course of the evaluation.

[68] See for example Reybold, Lammert & Stribling, 2012.

Box 1. Study Limitations Survey

Planning and Design

1) Consider the overall evaluation design. Were other evaluation designs possible, if any, given the time and resources you had available for the evaluation? If so, why did you choose the design that you did? Would other study designs have generated findings with fewer limitations?
2) Consider who was included in data collection. Was there a control or comparison group? If no, why not? How does this affect the strength of your findings? Did you draw a sample or samples? What was the basis for your samples? Were there specific groups intentionally or unintentionally excluded? What other samples could have been drawn? What impact would these different samples have had on your evaluation?
3) Consider the instrumentation. Did your instruments provide the exact data you needed to respond to your evaluation questions? If not, where were the data less than exact? How does this affect your ability to respond to the evaluation question? Do the results from your evaluation suggest potential problems in the instrumentation? For example, did survey items or scale scores show enough variability? Would changing the response options or rubrics offer better data analysis options in the future?

Data Collection

4) Examine response rates. What were your response rates? Is there a pattern to response or non-response? Did participation in your survey vary based on any participant characteristics? For example, were principals at higher performing schools more likely to return surveys? If response rates were poor in general or among a sub-group, how well can you generalize the results of the evaluation?

Data Management and Data Quality

5) Consider your coding and data cleaning procedures. Did you check for errors in coding and data entry? If no, what potential errors could have been made in coding and entering the data?
6) If more than one rater was involved in collecting data, examine inter-rater reliability. What is your inter-rater reliability? Is there a lot of variation among raters? Did one or more raters provide consistently high ratings while one or more raters provided consistently low or lower ratings? Did you make any adjustments for rater-driven variation in ratings? Are there other patterns in your data that appear correlated with rater or data collector? How do such patterns affect your analyses and findings?
7) Consider missing data. Was there a high amount of missing data? How did you address missing data? What are the implications attached to your handling of missing data?

Data Analysis

8) Consider your analysis options. Did you use the most rigorous analytic technique possible to respond to the evaluation question? If not, why not? How did the type and amount of data collected affect your ability to analyze the data and choice of analysis technique? What other variables need to be explored and collected to further your analysis of the evaluation data?

4. Methodological Considerations

This section provides additional information about different methodological aspects of planning and conducting a summative evaluation of a PDP project. Throughout the section our goal is not to present a comprehensive discussion of each topic, but rather to give a brief overview and highlight important points or potential issues that should be taken into consideration. Where possible, we provide links to online resources and throughout we offer suggestions on good print or electronic resources that evaluators can look to for additional information.

4.1 Evaluation Design

As mentioned in Section 2.2.2, the type of design chosen for a PDP project evaluation will depend upon, among other things, the goals of the study and the outcomes that are identified in the project logic model, since different study methods are required to measure different types of outcomes. We anticipate that in many cases a PDP project evaluation will feature a mixed-method design that incorporates multiple study methods, such as non-experimental studies (e.g., surveys or qualitative case studies) of graduates' perceptions of the quality of their PDP-funded program and quasi-experimental or experimental studies of graduates' performance or their students' performance.[69]

In writing this section we do not intend to prescribe a particular course of action for evaluators to follow. Rather, we aim to highlight different design alternatives that we believe can be implemented successfully during a PDP project evaluation, taking into consideration the varying contexts and diverse constraints PDP project evaluators face.

4.1.1 Randomized Experimental Designs

Since we believe that it will be difficult for evaluators of PDP projects to utilize RCTs in their studies, we will not focus on the different randomized designs here and instead refer interested readers to Shadish, Cook and Campbell (2002) for details on the different types of designs. However, we do want to highlight an important consideration for evaluators who plan to conduct RCTs as part of their study—the need to calculate attrition.

[69] See Section 2.2.2.4 for a discussion of mixed-method designs.

4.1.1.1 Calculating Attrition

Attrition is the loss of response from participants that takes place after the participants have been assigned to conditions. This might occur if a respondent fails to respond to a particular survey question or if the respondent refuses to participate in the study after the project begins. If the evaluator drops a respondent from the study for one reason or another it also should be counted as attrition, *if the drop could have been caused by the treatment*.

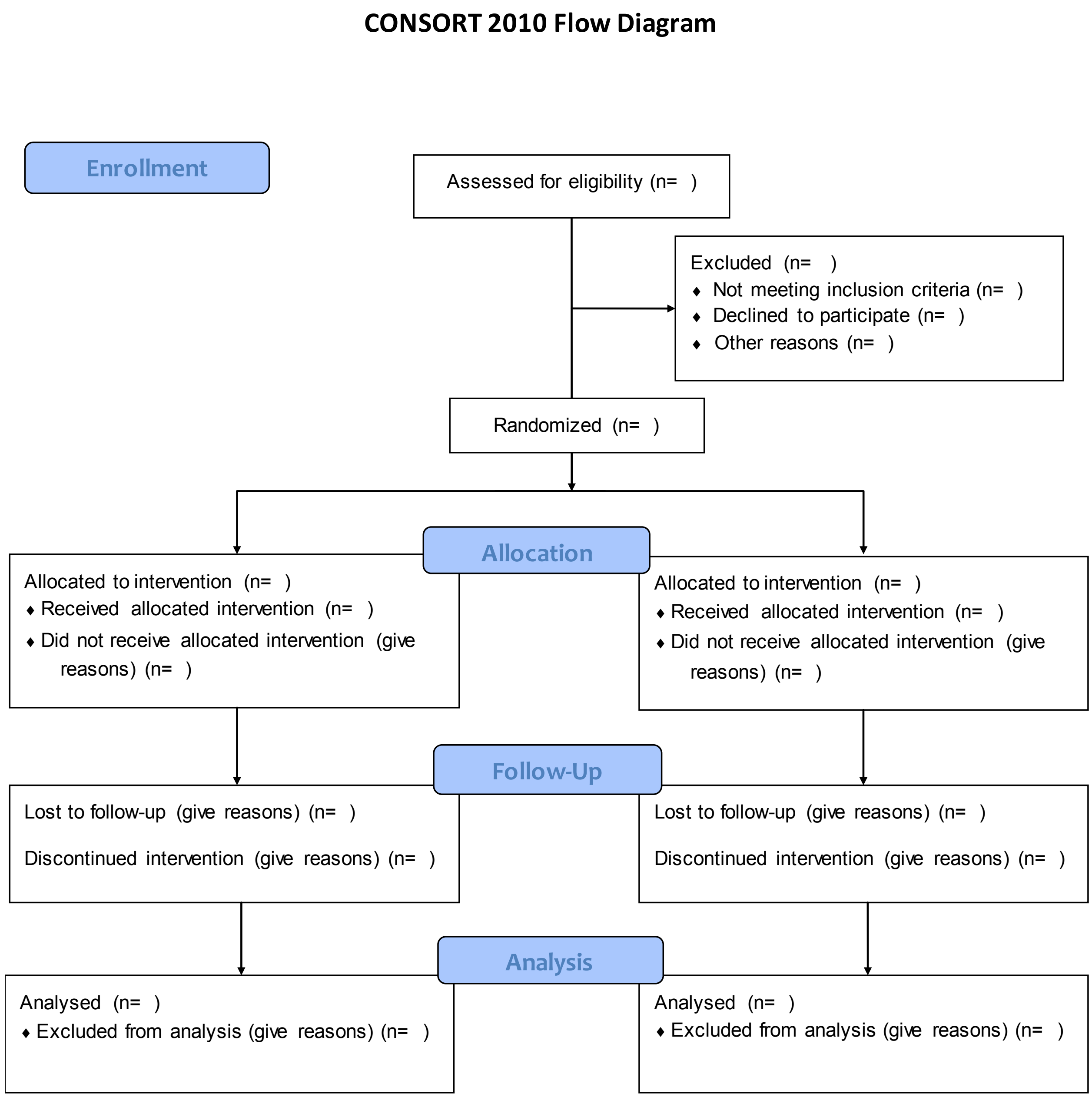
The primary problems with attrition are that it:

- lowers statistical power to detect effects, and
- any attrition from specific treatment conditions that appears to have been associated with the treatment (also called *differential attrition*) can threaten the internal validity of the study. When overall and differential attrition are high, the integrity of the randomized experiment is lost, thereby invalidating the assumption of equivalence among treatment and control groups.

The best strategy for dealing with attrition is to prevent it in the first place. Shadish, Cook and Campbell outline a number of strategies that can be used to retain and track participants in a study.[70] In most cases, however, it is not possible to completely prevent all attrition from occurring. For this reason, we recommend that evaluators use a tool such as the Consort Diagram presented in Figure 2 to track the participants in their experimental studies (we also recommend tracking participants in other study designs, but the problem of attrition relates specifically to randomized designs).

[70] 2002, pp. 225-334.

Figure 2. Sample Consort Diagram



Source: CONSORT (http://www.consort-statement.org/consort-statement/flow-diagram0/)

In the next section we talk about quasi-experimental designs, followed by a discussion of non-experimental designs.

4.1.2 Quasi-Experimental Designs

In the following sections we highlight a few quasi-experimental designs that we believe can be applied in PDP project evaluations.[71] This is not to say that evaluators should not try to use other types of designs, especially if those designs are more rigorous, but we simply are recognizing the constraints that evaluators of PDP projects may face when making design decisions. For each design we highlight validity threats (see Appendix E for a list of validity threats) and any available strategies to improve the strength of the design for making causal claims that were identified by Shadish et al. (2002). To describe these designs we will use the notations provided by Campbell and Stanley (1963)[72]:

- X represents an exposure of a unit to an experimental variable or event [i.e., the "treatment"], the effects of which are to be measured.
- O represents an observation or measurement recorded on an instrument [such as a standardized assessment, a teacher-made test, a survey, or a psychological scale]. The subscript following an O indicates a different time period.
- X's and O's in a given row are applied to the same specific units. X's and O's in the same column, or placed vertically relative to each other, are simultaneous [i.e., taking place at the same moment in time].
- The left-to-right dimension indicates the temporal order of procedures in the experiment (sometimes indicated with an arrow).
- Separation of parallel rows by a dashed horizontal line indicates that comparison groups are not equal (or equated) by random assignment. No dashed horizontal line between the groups displays random assignment of individuals to treatment groups.

4.1.2.1 One-Group Posttest-Only Design

In this design the study collects one posttest observation on respondents who experienced a treatment.

$$X \qquad O_1$$

This design can be acceptable for summative purposes in cases in which there is significant specific background knowledge about how the dependent variable might behave following a treatment. For example, if a child receives a well-established speech and language therapy intervention, and it is known that the child likely would not otherwise have developed the skills taught during the intervention, this design could be appropriate.

---
[71] See Shadish et al., 2002, for additional discussion of these designs.
[72] Cited in Creswell, 2003, pp. 168-9.

However, in most cases, with this design it is not possible to determine if a change has occurred following the implementation of a treatment and it is not possible to identify the counterfactual. Almost all threats to internal validity are possible with this design. However, "*for valid descriptive causal inferences to result [from this design], the effect must be large enough to stand out clearly, and either the possible alternative causes must be known and be clearly implausible or there should be no known alternatives that could operate in the study context.*"[73]

4.1.2.2 One-Group Pretest-Posttest Design

The one-group pretest-posttest design is a slight improvement over the one-group posttest-only design because it provides some information about the effect of the treatment on the outcome.

$$O_1 \quad X \quad O_2$$

There are multiple threats to internal validity with this design, including maturation, history, testing, and attrition. Researchers will "*rarely be able to construct confident causal knowledge with this design unless the outcomes are particularly well behaved and the interval between pretest and posttest is short.*"[74]However, adding another pretest to the design can reduce the plausibility of some of the validity threats, since it would give some idea of the trend of performance on the outcome measure prior to the implementation of the intervention.

$$O_1 \quad O_2 \quad X \quad O_3$$

Another way to improve this design is to add a non-equivalent dependent variable, diagrammed below.

$$O_{1A,}\, O_{1B} \quad X \quad O_{2A,}\, O_{2B}$$

In this design, measures A and B assess similar constructs, but measure A (the outcome) is expected to change because of the treatment, while measure B (the nonequivalent dependent variable) is not. Measure B is expected to respond to plausible internal validity threats in the same way as measure A would, so changes in measure B would illustrate whether these validity threats are actually operating within the study. For example, if the students' math scores (measure B) rose at the same rate as reading scores (measure A) when the treatment was solely focused on reading, something other than the treatment might be causing the reading score increases. While this design reduces the plausibility of many threats to internal validity, history remains a threat in this study.

[73] Shadish et al., 2002, p. 107.
[74] Ibid., p. 110.

4.1.2.3 Removed Treatment Design

In this modification of the one-group pretest-posttest design, a second posttest is added (O_3), then the treatment is removed and a third posttest (O_4) is administered.

$$O_1 \quad X \quad O_2 \qquad O_3 \quad \cancel{X} \quad O_4$$

The objective of this design is to show that the outcome changes with the presence or absence of the treatment. It is assumed that it would be difficult for many internal validity threats to operate in the same way, but the presence of outliers in the data can affect the results. Making the observations at equally-spaced intervals allows the examination of linear changes over time.

4.1.2.4 Repeated Treatment Design

In this design the researcher introduces a treatment, administers a posttest (O_2), then removes the treatment, administers another posttest (O_3), and then introduces the treatment again, followed by another posttest (O_4). This design and the one immediately above are commonly used by psychologists conducting behavioral research, usually with a single-case design (see Section 4.1.4).

$$O_1 \quad X \quad O_2 \quad \cancel{X} \quad O_3 \quad X \quad O_4$$

Again it is assumed that few threats to internal validity could explain a relationship that holds over this pattern of treatment introductions and removals. This design will not be able to show a causal relationship, however, if the treatment effects are not transient, or if the treatment creates a ceiling effect.

4.1.2.5 Posttest Only With Non-Equivalent Groups Design

This design is often used when the treatment begins before the start of a study. In this design, the evaluator selects a comparison group and administers a posttest to both the treatment group and the comparison group to see how their performance differs.

Group A	X	O
Group B		O

In general this design provides weak evidence of a causal relationship between the treatment and the outcome, since it is hard to determine if pre-existing group differences affect the outcome on the posttest. Nevertheless, it does provide some measure of a counterfactual.

If it is not possible to collect pretest and posttest data for the same group, it is sometimes possible to collect pretest data for a randomly formed independent sample that is drawn from the same population as the posttest sample. In the diagram below, the dashed vertical line indicates that the two observations are taken from two different samples over time.

O_1	¦	X	O_2
O_1	¦		O_2

It is important to randomly select the pretest and posttest samples from the same population (e.g., the total number of graduates of a PDP-funded program) or this design will introduce considerable selection bias into the results. This design should be used only when there is a strong need to have independent pretest and posttest groups or when there are significant problems collecting pretest and posttest data from the same groups over time.[75]

Some other ways to improve this design include:

- Matching or stratifying (see Section 4.1.2.7);
- Using internal control groups (e.g., individuals who registered for a class too late to be accepted into the study, but who are otherwise likely to be similar to students who are in the study); and
- Using multiple (non-equivalent) control groups (e.g., a group that is expected to underperform the treatment group if the treatment has no effect and a group that is expected to outperform the treatment group if the treatment has no effect).[76]

4.1.2.6 Non-Equivalent Pretest-Posttest Control-Group Design

This is probably the **most common of all quasi-experimental designs**. In this design, the treatment and control groups are selected without random assignment and the treatment is administered only to one group (Group A). Both the treatment and control groups are administered a pretest and a posttest.

Group A	O_1	X	O_2
Group B	O_1		O_2

With this design there is still a possibility of selection bias, since the groups are nonequivalent, but the pretest allows the researcher to explore the possible size and direction of that bias. We recommend that evaluators **calculate baseline equivalence** of the treatment and comparison groups using the pretest data, discussed briefly in Section 4.1.2.8.

This design also allows for an examination of attrition, which gives researchers an opportunity to see if there are differences among units that stay in and leave the study. The plausibility of different validity

[75] Shadish et al., 2002.
[76] See Shadish et al. (2002) for a full discussion of these options.

threats depends in part on the observed pattern of outcomes. This design can be improved by adding additional pretests, using switching replications, or using a reversed treatment control group.[77]

4.1.2.7 Matching Treatment and Comparison Groups

To decrease the odds of selection biases, evaluators can use matching to form comparison groups. This involves grouping units with similar scores on a matching variable (e.g., school size, ethnicity) so that treatment and comparison groups both have units with the same (or very similar) characteristics on the matching variable.[78] Different types of methods for matching include:

- Exact matching—when units have exactly the same score within a match (in practical terms this type of matching is not common, since it requires units to have the exact same scores);
- Caliper matching—when units have scores within a pre-defined distance of each other;
- Index matching—when multiple comparison units above and below a treatment unit are selected;
- Cluster group matching—when cluster analysis is used to embed the treatment group in a cluster of similar control units;
- Benchmark group matching—when control units that fall close to the treatment unit on a multivariate distance measure are selected;
- Optimal matching—when each treatment unit has multiple control units and vice versa;
- Cohort matching—when successive groups go through a particular treatment, such as in a PDP. Cohort matching is particularly useful *if*
  - one cohort experiences a treatment and earlier or later cohorts do not;
  - cohorts differ in only minor ways from their contiguous cohorts;
  - organizations insist that a treatment be given to everybody, thus precluding simultaneous controls and making possible only historical controls;
  - an organization's archival records can be used for constructing and then comparing cohorts;[79] and
- Propensity score matching—when treatment and comparison units are matched based on the conditional probability of receiving the treatment given a set of observed covariates. **Propensity**

[77] See Shadish et al., 2002 for additional discussion of this.
[78] Shadish et al., 2002.
[79] Ibid., p. 149.

score matching is gaining popularity among researchers and evaluators and is considered by many to be the preferred method.[80]

While matching generally improves the similarity of the treatment and comparison groups, there is still the possibility that the groups may differ on some variable not included in the matching. Additionally, in some situations it is possible for matching to produce a result that is actually further away from the correct answer than if no matching had been used.

Some general principals to improve matching include,

- Identifying a possible comparison group that appears to be very similar to the treatment group before conducting the matching of individuals; and
- Using matching variables that are stable and reliable, and that are correlated with the outcome variable.[81]

It is important to remember that even though the treatment and comparison groups may be similar at the time of matching, evaluators still must determine whether the groups remain similar on important characteristics at the end of the study, after some of the participants have left the study. This can be achieved by calculating baseline equivalence, discussed next.

4.1.2.8 Calculating Baseline Equivalence

Whenever a QED features a comparison group and pretest data are available, it is important to determine whether the treatment and comparison groups are similar in important ways at baseline. At its most basic, the calculation of baseline equivalence refers to calculating differences in means on a pretest measure between the treatment and comparison groups. If it is not possible to administer a pretest during the evaluation period, it might be possible to calculate baseline equivalence using another baseline measure that is correlated with the outcome measure (such as the prior year's state assessment). A key point is that evaluators should calculate baseline equivalence for the *analysis sample*—that is, the sample of individuals in the treatment and comparison groups that remain in the study for the entire study period. The What Works Clearinghouse has developed standards for establishing baseline equivalence in experimental studies with high attrition and quasi-experimental studies.[82]

In the next section we talk about non-experimental designs.

[80] For more information on propensity score matching, see, for example, Barth, Guo, & McCrae, 2008; Heinrich, Maffioli, & Vazquez, 2010; and Luellen, Shadish, & Clark, 2005.
[81] Shadish et al., 2002.
[82] U.S. Department of Education, 2014.

4.1.3 Non-Experimental Designs

As mentioned in Section 2.2.2, we recommend that PDP evaluators use experimental or quasi-experimental designs to answer summative evaluation questions whenever possible; however, we anticipate in some cases evaluators may need to use a non-experimental design, either alone or as part of a mixed-methods study. In the sections that follow we briefly discuss the different types of non-experimental designs that might be used for a PDP project evaluation.

4.1.3.1 Case Studies

Case study research can include single- and multiple-case studies and can incorporate a mix of quantitative and qualitative data. Multiple case studies may yield some nominal or categorical data. Yin (1994) outlined five different applications of case study research:

- To explain the causal links in real-life interventions that are too complex for the survey or experimental strategies (e.g., to link program implementation to program effects);[83]
- To describe an intervention in a real-life context in which it occurred;
- To illustrate certain topics within an evaluation in a descriptive mode;
- To explore those situations in which the intervention being evaluated has no clear, single set of outcomes; and
- To serve as a *meta-evaluation*—a study of an evaluation study (p. 15).

According to Yin (1994), case studies are preferable to other types of study designs "*when 'how' or 'why' questions are being posed, when the investigator has little control over events, and when the focus is on a contemporary phenomenon within some real-life context*" (p. 1). Shadish et al. (2002)—authors of one of the most well-known books about experimental and quasi-experimental research design—recognized that intensive qualitative case studies can be good non-experimental alternatives for generating causal conclusions, although they preferred the use of experimental or quasi-experimental designs whenever possible. Examples of case studies that may be possible in PDP project evaluations include studies of individuals who received training in very small disciplines or fields of study. In these cases, the overall

[83] Shadish, Cook and Campbell (2002) acknowledged that intensive qualitative case studies can be good non-experimental alternatives to generating causal conclusions for three reasons: (1) "Journalists, historians, ethnographers and lay persons regularly make valid causal inferences using a qualitative process that combines reasoning, observation, and falsificationist procedures in order to rule out threats to internal validity" (p. 500); (2) qualitative methods can explore causation in a way that incorporates much more complexity than can be achieved with an experiment; and (3) intensive case studies can yield much more information to the researcher and to policy makers than experiments can. However, Shadish et al. also pointed out that many qualitative studies do not provide sufficient information about the counterfactual to sufficiently reduce uncertainty about cause. For more information about using qualitative research to make causal claims, see Donmoyer, 2012a, 2012b; and Maxwell 2004a, 2004b, 2011, 2012.

sample size for the evaluation may be very small, which affects the generalizability of findings to larger populations. In addition, the participants may have received such specialized training that a survey or other quantitative assessment also lacks generalizability.

See Yin (1994) for an extended discussion of the design and conduct of case studies.

4.1.3.2 Descriptive Studies and Surveys

Descriptive or survey research aims to provide systematic and accurate descriptions of selected characteristics for a population under study. This can be done by conducting surveys with samples drawn from a population (see Section 4.2 for information on sampling) or by surveying all of the members of a certain population (e.g., all of the graduates of a special education teacher certification program at a Midwestern university during the period 2012-2015). Surveys most often generate nominal or ordinal data, although the responses to survey questions are often treated as interval data and analyzed accordingly. Descriptive studies are generally classified in terms of **how many times a sample is surveyed** (e.g., cross-sectional or longitudinal surveys) and **how the data are collected** (e.g., self-report surveys or observations). For example:

- **Cross-sectional surveys** involve one-time data collection with groups of individuals to compare groups at a single point in time. An example would be surveying all of the speech and language pathologists (SLPs) who are 1 year past graduation from a program and simultaneously surveying the group of SLPs who are 2, 3, 4, and 5 years past graduation. This study would be conducted all at one time and would provide a snapshot for each group of graduates.

- **Longitudinal surveys** involve collecting data multiple times from the same individuals, such as surveying the SLPs each year after graduation, obtaining data on the same individuals at 1, 2, 3, 4 and 5 years past graduation. This study would take 5 years to conduct.

It also is possible to combine cross-sectional and longitudinal surveys by starting a new longitudinal cohort at different points in time over the course of the evaluation.[84]

This type of research is well suited to answering “who,” “where,” “how much,” and “how many” questions. Further, surveys can be particularly useful when the goal of the study is to “*describe the incidence or prevalence of a phenomenon or when it is to be predictive about certain outcomes.*”[85] Finally, surveys might be preferable when multiple questions must be answered yet limited resources are available. As Cronbach put it, in such situations “*a survey might be preferred because it has a wider* ***bandwidth*** *that permits answering a broader array of questions even if the causal question is answered less well than it would be with an experiment*.”[86]

[84] Dimitrov, 2010.
[85] Yin, 1994, p. 6.
[86] 1982, cited in Shadish et al., 2002, p. 98, emphasis in original.

4.1.3.3 Correlational Studies

In general, correlational studies aim to **investigate relationships between variables** or use such relationships to make predictions about a variable of interest.[87] Correlational studies are commonly used in education—for example, to answer questions such as, "*Do graduates with higher grade point averages obtain employment in the area for which they are qualified in less time than graduates with lower grade point averages?"* or "*Do the students of graduates of personnel programs who report high self-efficacy have higher levels of academic achievement than students of graduates who report low self-efficacy*?" Correlational studies do not allow the researcher to determine which of the two variables being correlated came first, so it is not possible to identify a causal relationship in such studies. Further, correlational studies do not test alternative explanations for a presumed effect, thereby leaving the possibility that some other variable not under study (often called a confound) might actually be responsible for the observed relationship.[88] This may also be referred to as a "spurious correlation." Nevertheless, the existence of a strong correlation between two variables may point to hypotheses about causal effects that can be explored in subsequent, more rigorous, studies.

4.1.3.4 Ex Post Facto Studies

Ex post facto studies—studies that are conducted *after the fact* using secondary data—investigate cause-and-effect relationships by analyzing data on events that have taken place in the past. Ex post facto studies can be useful when (a) the independent variables cannot be controlled by the evaluator or (b) manipulation of independent variables is possible, but it may be unethical, impractical, or costly.[89] While this type of study cannot prove cause-and-effect relationships, they can lead to hypotheses that can be tested with other, more rigorous designs.

We now turn our discussion to single-case/single-subject designs, which are commonly used by special educators and related service providers to demonstrate changes in the performance of their students.

4.1.4 Single-Case/Single Subject Designs

Single-case designs involve in-depth study of a single "case," which can be a person, a group, an institution, or even a culture. Also known as single-subject designs, single-case studies can be either (a) a single-case study in a natural (uncontrolled) environment, or (b) a single-case experiment in a more controlled environment.[90] These designs are frequently used by psychologists, counselors, social workers, and special educators.

[87] Dimitrov, 2010.
[88] Shadish et al., 2002.
[89] Dimitrov, 2010, p. 43.
[90] Dimitrov, 2010; Kennedy, 2005.

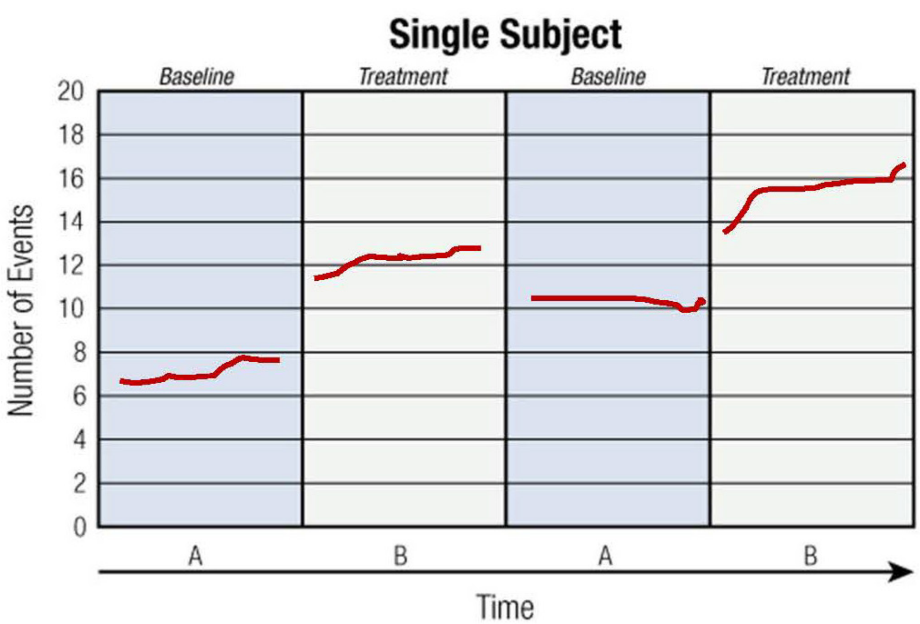
The two most common single-case designs are the A-B-A-B design and the multiple baseline design; these are presented briefly below. The interested evaluator is encouraged to consult Kennedy (2005) and Todman and Dugard (2001) for additional information about the different types of designs and how to conduct single-case experiments.

4.1.4.1 A-B-A-B design

An A-B-A design (where A = baseline and B = intervention) is the minimal type of experimental design that can establish experimental control in single case research, but researchers tend to prefer the A-B-A-B design for at least two reasons. First, the A-B-A-B design allows for two different replications of changes in pattern (i.e., when the baseline is reintroduced and when the intervention is reintroduced). Second, many researchers prefer to end a study in a way that is the most beneficial to the participant; if the intervention is working, the A-B-A-B design allows the participant to continue receiving the intervention as the study comes to a close.[91]

An important issue when using A-B-A-B designs is whether behavior will return to baseline levels after the intervention is withdrawn. If this does not occur, experimental control may be lost and no functional relation between the intervention and a change in behavior is demonstrated. Figure 3 demonstrates what data collected in an A-B-A-B design might look like.[92]

Figure 3. A-B-A-B Single-Case Design



Source: Adapted from http://www.cehd.umn.edu/nceo/Onlinepubs/Technical26.htm

[91] Kennedy, 2005.
[92] Ibid.

As can be seen, there are clear differences among the number of events that take place during the baseline phases compared to the treatment phases. However, the data for the second baseline show that the participant has not completely reverted to baseline levels. This might occur if the intervention introduces a new skill to the person; once the skill is learned it is difficult to reverse the effects of instruction. When this happens, it might be better to add additional baseline and intervention phases to see if the pattern of behavior change during the different phases continues.[93]

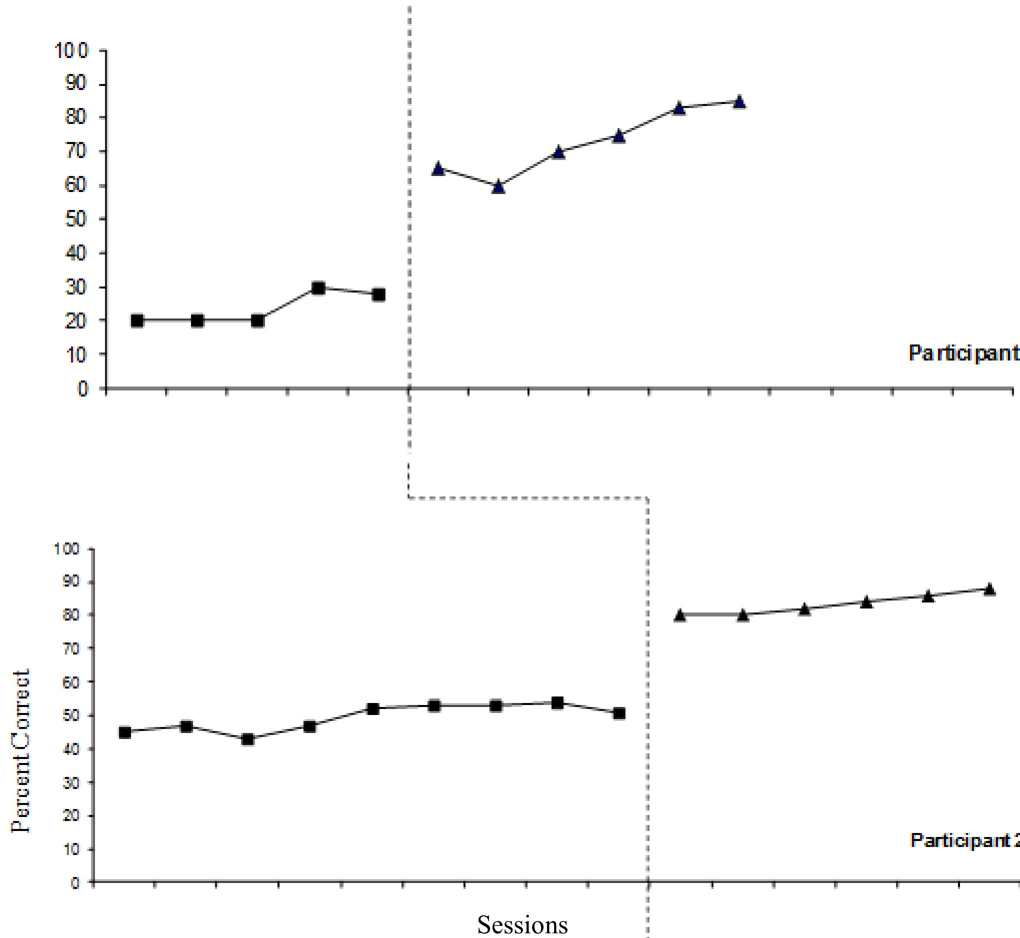
It is common in educational settings for a researcher to conduct an experiment when an intervention is already in place. In such cases, it is possible to use a B-A-B (or some extension of it). As with the A-B-A-B design, the establishment of experimental control in the B-A-B design depends on the variable of interest being sensitive to the withdrawal of the intervention. If behavior does not change during the withdrawal or reversal phase, a functional relation has not been established.[94]

4.1.4.2 Multiple Baseline Design

The advantage of the multiple baseline design is that it does not require withdrawal, reversal, or alternation of treatment conditions (not discussed here). Rather, two or more baselines are concurrently established and the intervention is introduced sequentially for each participant. This is an important alternative in situations in which it is not possible to remove the effects of the intervention once it has been introduced (e.g., when a student learns how to read it is unlikely that he or she will lose this ability).[95] Figure 4 illustrates what data collected during a multiple baseline study might look like.

[93] Kennedy, 2005.
[94] Ibid.
[95] Kennedy, 2005.

Figure 4. Multiple Baseline Single-Case Design



Source: Adapted from Lammert, 2007.

The logic behind the multiple baseline design is that individual baselines are established for each participant in the study (or group, classroom, etc.), consistent response patterns are observed, and then the researcher introduces the intervention to one participant at a time. The researcher should wait until a clear pattern is observed for the first participant post-intervention before introducing the intervention to the second participant, and so on. A functional relation is demonstrated when changes in the dependent variable occur only when the independent variable (intervention) is introduced.[96]

One of the drawbacks of the multiple baseline design is the fact that some participants must continue without the intervention for longer periods of time. This type of design may not be well suited to situations in which the behavior being observed is particularly undesirable (e.g., extremely disruptive behavior in class); instead, an A-B-A-B or other design may be more appropriate.[97]

[96] Kennedy, 2005.
[97] Ibid.

Morrison (2012) used a single-case approach to assess the performance of interns enrolled in a preparation program in Ohio. Some of the advantages or strengths of this approach to evaluating interns' performance included the ability of the single-case design to

- demonstrate practitioner's competencies and impact over time and across settings,
- assess a wide-range of professional competencies aligned with professional practice standards, and
- reflect naturally occurring work practices.

However, Morrison pointed out some disadvantages or limitations to this approach, including the following:

- Case studies will depend largely on opportunities practitioners have, which may vary by setting.
- Case studies are understood to be exemplars of professional competence, not necessarily representative of typical practice.
- A process for authenticating the product must be developed.

In the next section we briefly discuss different sampling strategies.

4.2 Sampling/Participant Selection

The size and type of sample to be used in a PDP project evaluation will ultimately depend on the evaluation questions and the type of analysis that the study team would like to conduct. Will the evaluation be comparing groups (e.g., comparing the graduates of the preparation program to graduates of other programs)? If so, how many? The more groups that need to be compared, the larger and more complex the required sample. Likewise, does the study team plan to conduct descriptive statistical analysis, inferential statistical analysis, or qualitative analysis of interview/case study data? The answers to these questions will help evaluators to decide which type of sample should be selected for the study. Of course, it is likely that there will be different samples for the different outcomes that will be measured in the evaluation, since different types of participants (e.g., program graduates or their students) will be involved in the intervention and data collection activities.

In experimental and quasi-experimental studies, the first step in selecting a sample often involves conducting a power analysis, which can be used to help identify the minimum sample needed in order to achieve a certain level of precision for statistical estimates. This is discussed briefly in the next section, followed by a discussion of two basic types of sampling—random sampling and purposeful sampling. As we mentioned before, sampling is an area where a team member with specific training, expertise, or experience is needed. For additional resources on sampling methodology, see the recommended readings in Appendix C.

4.2.1 Power Analysis

The term "statistical power" refers to the probability that a study will detect the effects of an intervention or treatment when there is indeed an effect (thereby helping the study to avoid Type II error). Statistical power analysis (a) is a way to determine the probability that a proposed research design will detect the anticipated effects of a treatment and (b) helps researchers determine whether to modify a proposed research design in order to achieve adequate power for detecting effects.[98]

For simple research designs with simple random samples (e.g., in which individuals are randomly selected and assigned to participate in a particular intervention), statistical power depends on three things:

- The desired significance level of the statistical test (e.g., α = .05)
- The expected size of the intervention effect (the effect size; e.g., ES = .3), and
- The sample size.[99]

[98] Hedges & Rhoads, 2010.
[99] Ibid.

In research designs that are considered *multilevel*—that is, for example, when students are clustered within classrooms or within schools—two other factors influence statistical power:

- The sample size at each level (e.g., for a three-level design, this might be the number of students in a classroom, the number of teachers in a school, and the number of schools in a district); and
- The extent of the clustering effects—that is, the amount of variation among clusters (e.g., classrooms) relative to the total variation in student outcomes for a school.[100]

Taking these factors into account will help the study team to determine the statistical power to detect effects of a particular size or larger—known as the Minimum Detectable Effect Sizes (MDES). Some good software programs have been developed to help conduct power analysis for individual and group-randomized experiments as well as quasi-experimental studies.

Power Analysis Software Tools

- Optimal Design—for randomized experiments (http://hlmsoft.net/od/)
- CRT Power – for simple and cluster-randomized experiments (http://crt-power.com/)
- PowerUp! – for randomized experiments and QEDs (http://web.missouri.edu/~dongn/PowerUp.htm)

It is beyond the scope of this Toolkit to go into detail about statistical power analysis. We recommend evaluators without training in this area consult a sampling statistician. Hedges and Rhoads (2010) and Dong and Maynard (2013) and http://www.causalevaluation.org/ are good resources for evaluators interested in conducting power analyses for their experimental or quasi-experimental studies.[101]

We now turn our discussion to two forms of sampling: random and purposeful sampling.

4.2.2 Random Sampling

Random sampling can be applied at two different levels in a study: **selection of units** for the study and **assignment of units** to treatment conditions. When sampling is random at both of these levels, a study is considered “fully randomized.” Random sampling aims for representativeness and is particularly good for minimizing threats to internal validity in a study. There are various types of random sampling, including simple random sampling, stratified random sampling, systematic random sampling, and cluster random sampling.

[100] Hedges & Rhoads, 2010.
[101] For additional information on power analysis see Raudenbush, Martinez & Spybrook, 2007; Raudenbush, et al., 2011; Raudenbush & Liu, 2000; Raudenbush, 1997; Schochet, 2008; and Spybrook, Raudenbush, Congdon & Martinez, 2009.

In a **simple random sample** all units in the population of interest have the same probability of being selected to participate in the study or the same probability of being assigned to one treatment condition or another. Simple random samples are relatively easy to select (assuming the population is known) and they permit generalization of results back to the population. However, when using simple random sampling to assign units to treatment groups, it is possible that the groups may be quite different from each other—a phenomenon known as "unhappy randomization." To reduce the risk of this, **stratified random sampling** involves dividing the sample into groups on selected variables (e.g., gender or ethnicity) and then selecting a simple random sample from within each group, thus not relying on a single sampling process. Stratified random sampling also allows the evaluator to be sure that specific groups are adequately represented in the study in order to conduct subgroup analyses.

Cluster (area) random sampling is similar to stratified random sampling, but instead of sampling subgroups within a population, the researcher divides the population into clusters (e.g., schools or school districts), randomly selects among the clusters, and then measures all units (e.g. classrooms or individual students) within the selected clusters.

Systematic random sampling involves numbering the units in the population from 1 to *N*, deciding on the sample size that is wanted or needed, determining an interval size (k = N/n), randomly selecting an integer between 1 to k, and then selecting every k^{th} unit for the sample. For this to work correctly the population must be listed in random order with respect to the specific characteristics being measured. Systematic random sampling is generally easy to do, and, under certain circumstances, can be more precise than simple random sampling. It may be a viable option for researchers who do not have the time or resources needed to conduct another type of random sampling. It is also possible to combine the various sampling strategies within the PDP project evaluation (http://www.socialresearchmethods.net/kb/sampprob.php).

We anticipate that it might be difficult for PDP evaluators to use the different random sampling techniques outlined above in their studies. Instead, the evaluations will likely feature some variant of purposeful sampling (also known as purposive sampling), discussed next.

4.2.3 Purposeful Sampling

In purposeful sampling the units in a study are not randomly selected or assigned to treatment conditions. This does not mean that the purposeful sample might not be a good representation of a population, or that it cannot serve the purposes of the specific study, but with such a sample researchers cannot know the statistical probability of the extent to which the sample is considered "representative"—thereby limiting generalization of the study results to other populations. In many situations, however, this is not required. Indeed, as Patton pointed out, "*the logic and power of purposeful sampling lie in selecting information-rich cases for study in depth. Information-rich cases are* *those from which one can learn a great deal about issues of central importance to the purpose of the inquiry."*[102]

Purposeful sampling selects information-rich cases strategically and purposefully; the specific type and number of cases selected depends on the study purpose and resources. The different types of purposeful sampling are outlined below.[103]

- **Stratified purposeful sampling:** Illustrate characteristics of particular subgroups of interest; facilitate comparisons.
- **Purposeful random sampling (still small sample size):** Add credibility to the study when the potential purposeful sample is larger than the evaluation can handle. Reduces bias within a purposeful category. (Not for generalizations or representativeness.)
- **Extreme or deviant case sampling:** Learning from unusual manifestations of the phenomenon of interest, for example, outstanding successes/notable failures; top of the class/dropouts; exotic events; crises.
- **Intensity sampling:** Information-rich cases that manifest the phenomenon intensely, but not extremely; for example, good students/poor students; above average/below average
- **Maximum variation sampling:** Document unique or diverse variations that have emerged in adapting to different conditions. Identify important common patterns that cut across variations (cut through the noise of variation).
- **Homogeneous sampling:** Focus; reduce variation; simplify analysis; facilitate group interviewing.
- **Typical case sampling:** Illustrate or highlight what is typical, normal, average.
- **Critical case sampling:** Permits logical generalization and maximum application of information to other cases because if it's true of this one case, it's likely to be true of all other cases.
- **Criterion sampling:** Picking all cases that meet some criterion; for example, all program graduates who remained in the state where the Personnel Development Project is located.
- **Confirming and disconfirming case sampling:** Elaborating and deepening initial analysis; seeking exceptions; testing variation.
- **Combination or mixed purposeful sampling:** Triangulation; flexibility; meet multiple interests and needs.
- **Convenience sampling:** Do what's easy to save time, money, and effort. Poorest rationale; lowest credibility. Yields information-poor cases.

[102] 2002, p. 230.
[103] Patton, 2002.

As with other design elements, the type of sample chosen for a PDP project evaluation will depend on many factors.

In the next section we turn our attention to methodological considerations related to data collection.

4.3 Data Collection Methods

As mentioned previously, the choice of data collection methods for a PDP project evaluation will depend upon the evaluation design and the resources available. In this section we discuss the selection and use of four typical data collection methods—surveys, observations, individual interviews, and focus groups—and present another method that is particularly applicable for collecting data on graduates working in related services, goal attainment scaling.

4.3.1 Surveys

There are two basic types of survey research: cross-sectional and longitudinal. **Cross-sectional surveys** collect data from a group of respondents at one point in time and are generally used for the following purposes:

- To examine current attitudes, belief, opinions, or practices;
- To compare two or more groups;
- To measure community needs (e.g., for related services provision);
- To evaluate a program; or
- To conduct a large-scale assessment of selected individuals or programs, such as a statewide or national survey.[104]

Cross-sectional surveys may be administered multiple times—for example program graduates 1, 2, and 3 years past graduation could be surveyed each year—but the group of respondents will differ each time (even if some of the same individuals respond to all of the surveys).

Longitudinal surveys follow selected individuals over time and are used to:

- Study **trends** in a population over time (e.g., attitudes among policy makers regarding public financing for related services for students with disabilities);

- Follow development or change in a **cohort** or subgroup of individuals who have been identified based on a specific characteristic (such as students who are deaf or hard of hearing) (Note: In a cohort design, different individuals may respond to each round of surveys, but all individuals must meet the cohort selection criteria to participate in the survey.); or

- Track development or changes in a specific group of individuals, or **panel**, over time. (Note: A panel survey follows the same individuals over time, thereby allowing the researcher to study

[104] Creswell, 2002.

actual changes in individuals. However, it may be hard to track each individual as time progresses, making panel surveys more costly and labor-intensive to conduct.)

The evaluator must decide, given time and resource constraints, if it is better to use an existing survey or develop a new one. Even if the existing survey is not exactly tailored to the specific study questions, it is generally easier to adapt an existing survey than to start developing a new one from scratch (Note: When adapting an existing survey the psychometric properties will change.). When done correctly, survey development is time-consuming and can be costly. Of course, the length and complexity of the survey will be determined by the study questions and by the type of analysis the study team would like to conduct.

4.3.1.1 Using an Existing Survey

As mentioned in Section 1, conducting surveys of preparation program graduates and their principals (presumably, supervisors in the case of related services professionals) is one effective way to gather information about whether preparation programs provided graduates with the skills needed to succeed on the job.

The Center to Improve Personnel Preparation Policy and Practice in Early Intervention and Preschool Education developed a survey of higher education programs preparing people to enter the fields of Early Intervention/Early Childhood Special Education.[105] Many of the items included in the survey can be used by evaluators of PDP projects to gather information about the practices of the different personnel preparation programs.

Many states and districts already are using administrator and teacher self-assessments to gather information about educators' perceptions of their professional strengths and areas of development, some of which are publicly available (e.g., the Rhode Island Model, http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Education-Eval-Main-Page/Teacher-Model-GB-Edition-II-FINAL.pdf). Other examples of existing surveys that measure teacher self-efficacy include:

- A modified Gibson and Dembo Teacher Efficacy Scale[106] for use in a special education resource-room context'[107]
- The Teachers' Sense of Efficacy Scale;[108] and
- The Teacher Self-Efficacy Scale.[109]

---

[105] Bruder & Stayton, 2004, http://www.uconnucedd.org/pdfs/projects/per_prep/PersonnelPrep_Briefing_Book.pdf
[106] 1984, cited in Coladarci & Breton, 1997.
[107] Coladarci & Breton, 1997.
[108] Moran & Hoy, 2001.
[109] Bandura, 2006.

It is also possible to use professional standards (e.g., the National Association of School Psychologists Professional Standards) as the foundation to develop a survey to assess educator practice (see Appendix B for links to professional standards and other educator resources). For example, Cochran et al. (2012) recently conducted a field validation of the Council for Exceptional Children's Division for Early Childhood early childhood special education /early intervention personnel standards.

Increasingly, surveys are being administered to students to gather information about their perceptions of the classroom instructional environment. The Tripod student survey instrument (http://tripodproject.org/about/background/), for example, "*assesses the extent to which students experience the classroom environment as engaging, demanding, and supportive of their intellectual growth.*"[110] The Tripod project also has teacher surveys that measure teachers' perceptions of pedagogy, teacher-student relations, and working conditions.

4.3.1.2 Developing a New Survey (or Adapting an Existing One)

In this section we highlight some important considerations related to developing/adapting a survey, rather than presenting an extensive discussion of the methodology. There are many good books about the process of developing surveys[111]—the interested evaluator is urged to review them for more detailed information about the topic. For simplicity, we talk here about developing a new survey, although the principles are generally the same for adapting an existing survey.

A key step in developing a survey is developing a **framework for the survey**. This includes, for each evaluation question that will be addressed by the survey, determining the item topics (i.e., what will be asked), the target population of interest (i.e., who will be asked), and the mode of administration (e.g., how they will be asked). Table 9 on the next page presents an example.

Item Topics. The goals of the study, the study questions, and the resources available will help to determine the topics that should be included in the survey. While it may be tempting to conduct an exhaustive survey covering all possible areas of interest, resource limitations and the time participants are willing to spend on the survey may force the evaluation team to make decisions regarding which outcomes are the most important to measure. Further, if additional data collection activities (e.g., observations or interviews) are planned, it may not be necessary to develop a lengthy survey.

Target Population. For each evaluation question the study team must decide who will be the best source of information, or target population (e.g., principals/supervisors, graduates, students, or all three) and determine how the sample will be selected (see Section 4.2 for a discussion of sampling).

Mode of Administration. There are two basic modes of survey administration: self-administered questionnaires and structured interview questionnaires. Self-administered questionnaires are usually provided in paper-and-pencil or online/electronic format, while structured interview questionnaires

[110] Jerald, 2012, p. 7.

[111] E.g., Czaja & Blair, 2005; Dillman, Smyth, & Christian, 2009; Groves, Flower, Couper, et al., 2004, and Harkness, Braun, Edwards, et al., 2010.

generally are administered in-person or over the telephone (see the sub-section on Interviews for more information). When developing the survey, the study team should determine which language will be the most appropriate for the target population (e.g., Vietnamese for parents of children living in a predominantly Vietnamese-immigrant community) and consider offering the survey in multiple languages. Further, before beginning to develop each item, the study team should decide whether the mode of administration will affect the respondents' decision to answer a question or influence the answer itself. For instance, students may be more likely to answer questions about their perceptions of their teacher in a self-administered survey than in a structured interview questionnaire administered in a group setting.

Table 9. Sample Survey Development Framework

Evaluation Question	Item Topics	Target Population	Mode of Administration	Type of Analysis
To what extent do graduates exit the program with the skills and knowledge necessary to perform at a high level?	Principal Survey • Principal/ supervisor reports of graduates' knowledge and skills • Principal/ supervisor reports of graduates' use of evidence-based practices • Principal/ supervisor reports of graduates' performance relative to professional standards • Principal/ supervisor reports of quality of service provided to students/children	Principals/ supervisors	Self-administered survey (paper-and-pencil or electronic)	Descriptive statistics
	Graduate Survey • Graduates' perceptions of their knowledge and skills • Graduates' perceptions of self-efficacy • Graduates' reports of their use of evidence-based practices • Graduates perceptions of their performance relative to professional standards • Graduates' perceptions of quality of service provided to students/children	Program graduates	Self-administered survey (paper-and-pencil or electronic)	Descriptive statistics
	Student/Client (of program graduate) survey • Students'/children's perceptions of quality of service provided to them	Students/ children	Self-administered survey (paper-and-pencil or electronic)	Descriptive statistics

Czaja & Blair highlighted three fundamental characteristics of a good survey questionnaire: it is a valid measure of the factors of interest, it convinces respondents to cooperate, and it elicits acceptably accurate information.[112] Figure 5 presents a visual representation of the process that respondents must go through to answer a given survey question. If the respondent cannot understand the question,

[112] Czaja & Blair, 2005, p. 65.

cannot remember the answer (or never knew it), does not want to give the answer, or cannot figure out *how* to give the answer, then the there is a problem with the item.[113]

Figure 5. The Process of Answering a Survey Question

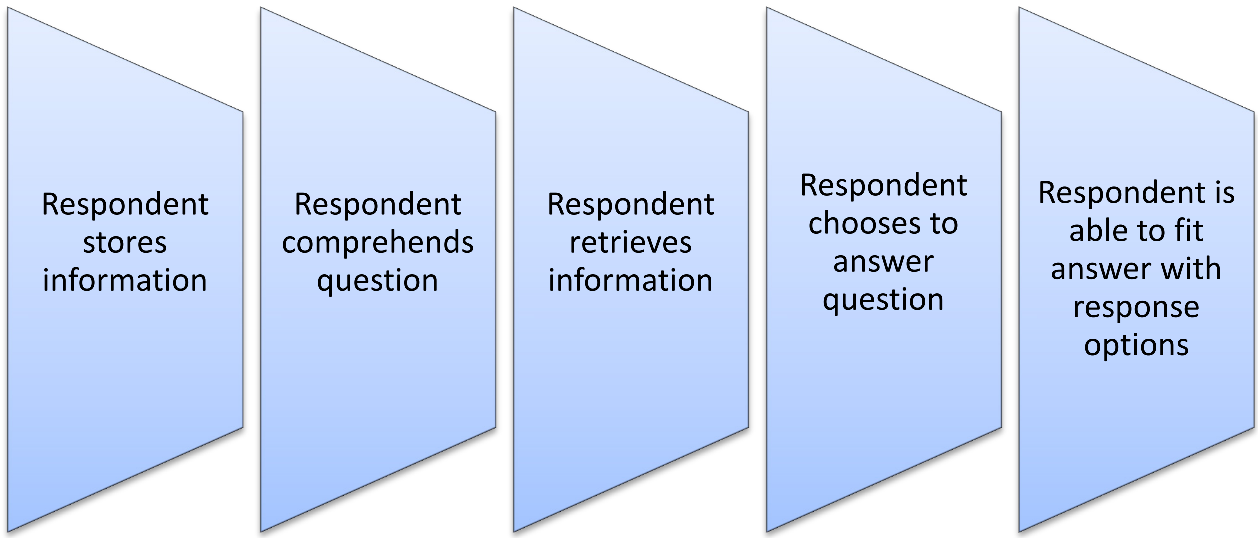


Table 10 presents some characteristics of a good survey item with examples of questions that might be asked of PDP scholars or graduates or the students of graduates.

[113] Fiore, Helba, Berkowitz, et al., 2012.

Table 10. Elements of a Good Survey Item, with Examples of Not Good and Good Items

Characteristic	Examples	
	Not Good	Good
Only asks one question at a time (avoids double-barreled questions)	To what extent do you think the academic coursework and field experience required by your PDP prepared you for what you would experience in the classroom?	To what extent do you think the field experience required by your PDP prepared you for what you would experience in the classroom?
Contains a clear threshold for answering "yes"	Have you seen your speech and language pathologist in the past month?	Have you met with your speech and language pathologist to practice fluency in the past month?
Provides a timeframe	How often does your supervisor observe you providing early intervention services to children?	How often each semester does your supervisor observe you providing early intervention services to children?
Provides a timeframe appropriate to the topic	How many times did were you absent from class during your personnel preparation program?	How many times were you absent from your field placement last semester?
Uses clear terminology and plain language	Are children with, or at risk for, developmental delays more likely to experience latency to learn a contingency?	Are children with, or at risk for, developmental delays more likely to experience delays in learning the relationship between a behavior and its consequences?
Gives exhaustive and mutually exclusive response options	In which of the following situations does your supervisor most frequently observe your work? • When you're working in a general education classroom • When you're working in a self-contained special education classroom • When you're working with your students	In which of the following situations does your supervisor most frequently observe your work? • When you're co-teaching in a general education classroom • When you're teaching in a self-contained special education classroom • When you're providing pull-out services to children with special needs

Czaja & Blair pointed out that that most refusals to complete a survey occur at the beginning. That is, the information provided in a cover letter or in the first parts of the survey helps the target respondent decide whether or not to complete the survey. As such, they recommend that the first question have all of the following characteristics:

- Relevant to the central topic
- Easy to answer
- Interesting
- Applicable to and answerable by most respondents
- Closed format.[114]

[114] Czaja & Blair, 2005, p. 94.

Another consideration when developing surveys is that the ordering of the items within a survey can affect responses. Whenever possible, it is better to place sensitive questions later in a survey so that respondents do not decide to terminate the survey early. It also is good to place items requesting demographic information (e.g., gender, occupation, ethnicity, disability status) at the end of the survey so as not to discourage any respondents from answering the survey questions.

In some cases the answer to one question can influence the answer to the following item. For example, if a graduate is asked to rate the different components of the PDP-funded project (e.g., the academic coursework, the field placement, the relationships among scholars in the program, the relationships among scholars and professors), his or her general assessment of the program in a subsequent question may be different than if the item order had been reversed.

The available response options also can affect a respondent's ability or willingness to answer. For instance, if a student has a "neutral" opinion about her experience with the occupational therapy services offered by her related services provider and is asked to choose between the response options "highly dissatisfied," "moderately dissatisfied," "moderately satisfied," or "highly satisfied," she may choose to skip the survey item.

Moreover, researchers have identified a number of problems with response options in the agree-disagree (A-D) format (i.e., strongly agree, agree, neither agree nor disagree, disagree, strongly disagree). Specifically, some of the **problems associated with agree-disagree items** include:

- A-D items are more cognitively difficult – Respondents have to understand the question as written and then translate their answer to the question into an A-D response format.
- A-D items are subject to acquiescence response bias – Respondents tend to agree with A-D items regardless of the content, especially when the respondent ability and motivation are low and the task difficulty is high. Culture also influences this bias (i.e., Latinos and respondents from collectivistic cultures tend to demonstrate this bias more frequently).
- It is often difficult to interpret the meaning of a "disagree" response – It can be difficult to know whether the respondents disagree with the wording of the item itself or with the subject of the item.
- A-D items often force respondents to think through double negatives to be able to respond – Respondents may have to consider whether they disagree with a negatively-worded statements, such as "Public officials don't care about people like me much."[115]

[115] Holbrook, 2013.

Whenever possible, we recommend using response options that are item-specific. For example, consider the following items; the first offers the A-D response options and the second offers item-specific options:

Option 1: "Sometimes my field experience seems completely unrelated to the topics covered in my PDP coursework." Do you AGREE STRONGLY, AGREE SOMEWHAT, NEITHER AGREE NOR DISAGREE, DISAGREE SOMEWHAT, or DISAGREE STRONGLY with this statement?

Option 2: "How often does your field experience seem completely unrelated to the topics covered in your PDP coursework?" ALWAYS, MOST OF THE TIME, ABOUT HALF OF THE TIME, SOME OF THE TIME, OR NEVER?

Although it may take more time to develop item-specific response options for a survey, the quality of the data obtained by these items is better than with A-D items. Additionally, studies have shown that item-specific response options do not take significantly more time to complete than A-D format items. There are times when using item-specific response options may not be an option, however, such as: (a) when comparing results of the survey to previous studies that used A-D items; or (b) when conducting longitudinal studies that have used A-D items in past data collections. In these situations, it may be better to continue the use of the agree-disagree format, keeping in mind the inherent limitations of these items. [116]

See http://www.360degreefeedback.net/media/ResponseScales.pdf for a variety of response options that can be used for developing a survey.

4.3.1.3 Reviewing and Testing the Survey

Whenever possible, a content expert should review the survey to see if the content is accurate and if the items adequately address the item topics laid out in the data analysis plan. Further, a methodological expert should review the survey to identify any issues with item construction or survey organization. If neither of these are viable options, the study team should carefully review the survey to look for any of the issues outlined above. Once again, there are many good books on survey development that can provide guidance on the survey review.[117]

Once the review has been completed, the evaluation team should conduct a pretest (or pilot test) of the survey. As Czaja and Blair pointed out,

> *In designing a questionnaire, we make many decisions and assumptions, some conscious, others not. Underlying the questionnaire draft is our judgment about what kinds of things respondents will know, what words they will understand, what sorts of information they can and will provide, and what response tasks they can perform. When we pose alternative choices to the respondent, we have in mind some notion of the appropriate dimensions for an answer; in fact, we start to*

[116] Holbrook, 2013; for more information see Saris, Revilla, Krosnick & Shaeffer, 2010.
[117] E.g., Czaja & Blair, 2005; Dillman, Smyth, & Christian, 2009; Groves, Flower, Couper, et al., 2004, and Harkness, Braun, Edwards, et al., 2010.

envision what our data will look like. Much of our effort is subtly informed by our feelings about how people will respond to our queries, by a belief that what we ask is sensible, by some vision of how the respondents' world works.[118]

Even after a thorough review is conducted, the pretest is the best way to know how the target population may respond to the survey. During the survey review issues may arise that can be specifically addressed in the pretest, such as whether respondents understand the terminology included in a particular item. The pilot may involve one-on-one administration of the survey in an interview situation (e.g., through a structured interview or cognitive interviewing[119]) or field testing the survey (i.e., administering the survey using the exact procedures planned for the overall study) with a small sample similar to the target population and then conducting follow up interviews, telephone calls, or even internet or email exchanges. A pretest protocol may be developed if more than one person is conducting the follow up with respondents.

The following are examples of pretest questions adapted from Czaja and Blair:[120]

- Were there any questions you were not sure how to answer? If yes, which ones were those? Why were you not sure how to answer the question?
- When I used the term *latency to learn a contingency*, what did you think I meant by that?
- When I asked the question about the quality of your PDP-funded program, what sorts of things did you consider?
- Are there any questions you think that many people would find difficult to answer? If yes, which ones were those? Why do you think people would have difficulty with the question?
- Were there any important things related to these issues that we failed to cover?

Czaja and Blair also offered four final suggestions for survey development:

- If possible, use multiple pretesting methods and multiple rounds of testing;
- Learn to assess survey questions by reading them aloud and listening carefully for any awkward or unnatural phrasing;
- Consider no question's wording sacrosanct—accept that sometimes the question must be reworded; and
- Look for examples of questionnaires written by experienced researchers.

[118] Czaja & Blair, 2005, p. 105.
[119] A cognitive interview is a one-on-one interview designed to determine the process respondents use in answering a question, or to identify problems respondents have in understanding or answering a question. Such interviews use a variety of laboratory techniques such as think aloud procedures or the paraphrasing of questions (Czaja & Blair, 2005).
[120] Czaja & Blair, 2005, p. 109.

Finally, remember: **When making changes to an existing survey, it is not appropriate to cite the psychometric properties of the original survey.** If this information is desired, the study team will need to calculate new reliability statistics and evaluate the validity of the revised survey.[121]

4.3.1.4 Potential Sources of Error in Surveys

Every survey contains some sources of error that reduce the accuracy of the data collected. This error can be large or small—and some error simply cannot be eliminated no matter what the researcher does, as in the case of a sample survey (which inherently includes error because it does not gather data from all members of the population). Nevertheless, some sources of error are more harmful than others, and the researcher should make efforts to minimize them.[122] These are briefly discussed below.

4.3.1.4.1 Low Response Rates

The response rate for a survey is the percentage of eligible respondents for whom survey data are obtained. There are multiple ways to calculate response rates but we will not focus on those here[123]; instead we focus on two types of response rates that can impact the results of the survey—the overall response rate (known as the unit response rate) and the item response rate. The unit response rate is the percentage of potential respondents who actually complete the survey, while the item response rate is the percentage of respondents who answer a particular item. A low item response rate tends to be less of a concern than a low unit response rate, but "*If [you] fail to obtain any information from some respondents, and for others fail to obtain complete information, [your] estimate and other analyses may be distorted, sometimes quite seriously.*"[124]

When response rates are low, the reliability and validity of the study conclusions are called into question. For example, if the study team sends a survey to all of the parents of students of the PDP project graduates (n = 150) and only 25% of the parents respond, to what extent can the study team be confident that the results of the survey are an accurate reflection of parents' perceptions of the PDP project graduates? If 90% of program graduates do not answer a series of questions related to the support they received from their supervisor during their field experience, how should the study team interpret those results? Are the omissions a result of poorly worded questions or did the graduates misunderstand the skip instruction in a previous question? Are the omissions indicative of graduates' unwillingness to answer questions they perceive to be too sensitive? If a survey pretest was conducted prior to administering the survey, any difficulties with item wording or skip patterns should have emerged. However, even when a pretest was conducted, it can be very difficult for a researcher to know how to interpret missing data from items not answered. Consequently, if either of these rates is low, it is

[121] See, for example, Dimitrov, 2012.
[122] Czaja & Blair, 2005.
[123] For more information, see Czaja & Blair, 2005.
[124] Ibid., p. 197.

important to determine whether there are systematic differences between respondents and non-respondents, since such differences might indicate bias in the data.

4.3.1.4.2 Differences Among Respondents and Non-Respondents

When examining differences among respondents and non-respondents, consider two primary questions:

- Are some demographic groups under- or over-represented among survey respondents?
- Are some of the survey items answered differently by members of different groups?

As discussed in Section 4.2, the study's sampling frame tells what demographic groups should be included in the study sample. Using demographic data from the respondents (collected through the survey or through administrative records), evaluators can analyze the response patterns to determine whether some groups responded to the survey or answered specific questions more than others. For example, were program graduates who had higher grade point averages (GPAs) at the end of their program more likely to respond to the survey than graduates who had lower GPAs at the end of their program? Or, were program graduates who had a specific field supervisor more likely to respond to questions about the quality of their field placement experience than graduates who had a different field supervisor?

Of course, the existence of differential response rates among respondents and non-respondents does not necessarily mean that the data obtained through the survey are biased. However, further investigation is warranted since such differences might affect the validity of the study's conclusions. "*To the extent that opinions and behaviors differ by subgroups, their overrepresentation or underrepresentation will affect results.*"[125]

If different patterns of responses among different groups of respondents are found, we recommend conducting a **non-response analysis** to check for response bias. One way to do this is to contact a few non-respondents (best if selected randomly) by telephone or by email to see if their answers differ substantially from respondents. There is no specific guideline for the number of such follow-ups that might be made, but if evaluators can document that the responses obtained from the non-respondents do not differ very much from the responses originally obtained, the evaluation team (and the study audience) will have greater confidence in the conclusions of the survey. Regardless of whether a non-response analysis is conducted, evaluators should clearly document in the evaluation report the limitations of the study conclusions related to a possible response bias in the survey.

[125] Czaja & Blair, 2005, p. 198.

4.3.1.5 Reducing Error in Surveys

As with every other aspect of the study, decisions about whether and how to reduce survey error must balance costs and available resources.[126] We have already talked about design elements that can improve the quality of a survey and we will not revisit them here.

The best way to avoid error related to non-response is to conduct extensive follow-up with the survey sample to increase the number of respondents. This can be done through telephone calls, emails, and even letters (although telephone calls are generally considered to be the most effective follow-ups). Keep in mind that every contact the study team has with the potential respondents—whether through email exchanges, voicemail recordings, or messages left with another person in the household—can affect whether or not the respondents decide to complete the survey. While developing the data collection plan for the evaluation, evaluators should consider the ways that respondents' willingness to participate in the study might be maximized. Taking time to "*[think] through the data collection process in simply a commonsense manner—but from the respondents' perspective—[will] produce many ideas, concerns and insights about how best to conduct data collection.*"[127]

We now turn to our discussion of another common data collection method, observations.

4.3.2 Observations

It is important to note that conducting observations that will produce high-quality data requires disciplined training and rigorous preparation.[128] Observational methods range from the highly qualitative field observations (e.g., field notes and journals) commonly utilized by anthropologists and ethnographers to the more quantitative observations conducted using rubrics, frameworks, or other types of observation instruments or checklists. For the purposes of this Toolkit, we will focus on the latter.

4.3.2.1 Considerations When Choosing/Developing Observation Protocols

The type of observation instrument, or protocol, used in the study will have a significant impact on the quality and consistency of data that are collected during the observation. Every observation protocol will have certain inherent limitations, so during the planning phase the evaluation team must consider which kinds of data need to be collected during the observations. Consider the following examples in Figures 6 and 7 below.

---

[126] Czaja & Blair, 2005.
[127] Ibid., p. 199.
[128] Patton, 2002.

Figure 6. Observation Protocol Example 1

Teacher's name/initials/pseudonym:____________________________ Grade:__________________

Observer:_______________________________________ Setting:_______________________________

Summary of Teacher's Observed Behavior	Implications

Figure 7. Observation Protocol Example 2

Teacher's name/initials/pseudonym:____________________________ Grade:__________________

Observer:_______________________________________ Setting:_______________________________

Observed Behavior	**Time Interval (Place an "x" in each box if the behavior was observed during the interval)**			
	10:00 - 10:05	10:06 - 10:10	10:11 - 10:15	10:16 - 10:20
Teacher follows the instructional model closely.				
Teacher differentiates instruction.				
Teacher models skills correctly.				
Teacher follows the steps of the correction procedures to provide immediate feedback.				

As can be seen, the types of data collected by the two instruments presented above will vary considerably. In the case of Figure 6, there is a strong possibility that the data collected by one observer will differ greatly from the data collected by another observer. In the case of Figure 7, there are many behaviors not included in the protocol that might be observed during a class period, but since they are not listed they will not be recorded. The example in Figure 6 provides data that are information-rich, yet possibly inconsistent—and requiring extensive qualitative analysis, while the example in Figure 7 provides data that are consistent, able to be analyzed quantitatively, yet possibly incomplete. Additionally, the type of training needed to be able to assess "implications" of a given behavior (Figure

6) is very different from the training needed to be able to accurately record specific instances of selected student behaviors (Figure 6). (Note: In the case of Figure 7, if the study wanted to gather additional data about how often certain behaviors occurred, rather than recording an “x” in the time interval, the observers could use a scale such as 0 = never, 1 = sometimes, 2 = frequently to give some measure of the frequency of behaviors during the class period.) When selecting or developing an instrument to use in the observations, these considerations should be kept in mind.

In many instances it is preferable to use an existing observation protocol that has been tested and that has guidelines for training observers in its use, such as the Classroom Assessment Scoring System (CLASS)[129] and the Classroom Climate Scale.[130]

The majority of states and districts use teacher observations as their primary measure of teacher effectiveness.[131] However, a study conducted in 12 districts across four states (which included surveys of 15,000 teachers and 1,300 administrators) found that such observations are **generally infrequent** (generally two or fewer observations per year) and **too brief** (60 minutes or less) to accurately measure teacher performance. Additionally, the observations were frequently conducted by untrained administrators and the results seldom were used to provide feedback to mediocre or poor performing teachers.[132]

The Bill & Melinda Gates Foundation is one of a number of organizations advocating improvements in the quality of classroom observations.[133] In addition, many states are taking steps to develop and implement classroom observation protocols that are more consistent and better measures of the instructional practices taking place within a classroom. Some of these protocols are available online (see, for example, the Indiana Department of Education’s RISE Evaluation and Development System, http://www.doe.in.gov/sites/default/files/evaluations/rise-handbook-2-0-final.pdf). Unfortunately, a study conducted by the National Comprehensive Center for Teacher Quality found that the majority of states and districts do not use observation protocols designed to measure the specific functions and distinct roles of special educators. Rather, most simply adapt an existing protocol to reflect specialized roles and responsibilities.[134] The Alabama Department of Education (DOE), for example, has modified its teacher observation instrument to reflect instruction toward alternative standards for teachers of students with low-incidence disabilities.

Whichever protocol the study team selects for the observations, it will be necessary to train the observers to ensure that the data they collect is consistent and reliable.

[129] Pianta, L Paro, & Hamre, 2008.
[130] McIntosh, Vaughn, Schumm et al., 1993.
[131] Holdheide, Goe, Croft, & Reschly, 2010.
[132] Weisberg, Sexton, Mulhern, & Keeling, 2009.
[133] Jerald, 2012.
[134] Holdheide et al., 2010.

4.3.2.2 Training the Observers

It is well known that human perception is highly selective. "*What people 'see' is highly dependent on their interests, biases, and backgrounds. Our culture shapes what we see, our early childhood socialization forms how we look at the world, and our value systems tell us how to interpret what passes before our eyes.*"[135] Simply being able to see and hear what is happening in a classroom or other setting does not mean that a person is able to conduct a high-quality observation. Different people frequently highlight and report different aspects of a given situation, so some training is needed to ensure that individuals conducting observations report the same incident with accuracy, authenticity, and reliability. As Patton[136] pointed out, training to become a skilled observer includes:

- Learning to pay close attention to what is happening in a given context (e.g., recording visual, auditory and behavioral cues);
- Practicing writing descriptively (e.g., documenting the setting and interactions among people being observed);
- Acquiring discipline in recording field notes or observational data (e.g., monitoring frequency of student behaviors during 5-minute time intervals);
- Knowing how to separate detail from trivia (e.g., documenting a significant disruption in the flow of instruction but not recording every side conversation taking place among students);
- Using rigorous methods to validate and triangulate observations (e.g., calculating inter-rater reliability among two or more observers); and
- Reporting (or at least acknowledging) the strengths and limitations of one's own perspective as an observer (e.g., recognizing that having a background in literacy might make an observer more attuned to differentiated instruction focused on literacy than that focused on math).

Further, all individuals who will be conducting observations should be trained in the use of the specific observation protocol employed by the study. It is important to provide observers with an instructional manual for each protocol that includes clear descriptions of the key terms and examples of what the behavior being observed "is" as well as what it "is not." Additionally, the observation protocol itself should include clear instructions to the observer as well as definitions of any terms that might be subject to differing interpretations by multiple observers. For example, the phrase "at inappropriate times" in Figure 7 above might mean one thing to one observer and mean something entirely different to another observer.

Prior to sending the observers into the field to conduct the observations, the study team should practice using the protocol, and should calculate the agreement between different raters (also known as inter-rater reliability, see section 4.3.2.6). Ideally, this would include opportunities to use the observation

[135] Patton, 2002, p. 260.
[136] Ibid.

protocol in a "real-life" setting, such as might occur during a pilot study. If this is not possible, video clips can serve as examples for training observers (Some are available online. See, for example http://www.nytimes.com/interactive/2010/03/07/magazine/20100307-teacher-videos.html#/readingaloud or http://www.youtube.com/watch?v=n16q9V9Pf2c). At a minimum, the study team should have a conversation about the use of the protocol, during which sample situations are presented to the team and a consensus is reached as to how the protocol would be used to record what is happening.

4.3.2.3 Deciding How Many Observations to Conduct and Who Will Conduct Them

Although there is no established standard for the number of observations to conduct, good practice suggests that educators should be observed multiple times.[137] Nevertheless, the number of observations will depend in large part upon the resources available, as well as the proximity of the study team to the classroom or professional setting. Likewise, the number of individuals conducting each observation will depend upon these same factors. When developing the evaluation plan the study team should calculate the expected cost of the observations in terms of travel (e.g., Is it necessary for the observer to book a hotel and rent a car?) and staff time (e.g., How many hours will it take for the individuals to complete the observations, including travel time?). It is also useful to consider the skills that will be required of the observers and if multiple observers are needed to complete one assessment. In a study of the effects of a commercial reading program on student performance in middle school classrooms, for example, all observations were conducted by a pair of observers.[138] One observer was a researcher who was highly trained in observation methodology as well as in the specific protocol. The other observer was a literacy specialist who was trained in the use of the observation protocol and whose content expertise was valuable in helping to identify whether specific teacher practices were in line with good literacy practice (even if they might not exactly align with the reading program). Together, these two observers generated complementary data that enhanced the study team's understanding of what actually was taking place within the classroom.

4.3.2.4 Obtaining Access to the Observation Site

The ability of the study team to conduct observations depends upon the teacher or other related-services professional, the school, and the district being willing to allow the observations to take place. The study team will need to contact the school district well in advance to obtain permission to conduct the observation. This may entail obtaining formal permission from the district's institutional review board (IRB) or human subjects review board (HSRB) or simply receiving permission from the district superintendent or his or her designee. In some cases the study team will need to receive permission from the parents of students. The school principal will also need to be contacted about the observations

[137] Bill & Melinda Gates Foundation, 2012.
[138] Dimitrov, Jurich, Frye, Lammert, Sayko & Taylor, 2012.

(see Section 3.1 for information on obtaining IRB and district approval and Section 3.2.2 for information on obtaining participants' consent. Appendices A.6, A.7, and A.8 include sample notification letters).

4.3.2.5 Conducting the Observations

On the day of the observation the study team should arrive to the setting early to check in with the appropriate people (e.g., the principal or other administrators) and to give time to move to the observation location (e.g., to navigate a busy hallway during the break between classes). During the observation, the study team should try to minimize disruptions to the classroom/setting as much as possible. The study team should also insure that no classroom events (such as fieldtrips) or school-wide events (such as picture day) will interfere with their observation.

It is possible to conduct video observations, rather than in-person observations, if the study team has limited resources or staff to support travel to the different study locations. Through its Measures of Effective Teaching (MET) project the Bill and Melinda Gates Foundation is supporting the use of video observations as one measure in a comprehensive evaluation of teacher practice.[139] Other researchers have pointed to the benefits of special education teachers using video observations.[140] The costs associated with the different types of video observations vary greatly, however, so even though they do not require individual travel, video observations may not be more economical.

4.3.2.6 Calculating Inter-Observer Agreement

When more than one person conducts the observations it is good practice to calculate inter-observer agreement (also known as inter-rater reliability) to determine the consistency of the ratings across observers. Inter-observer agreement calculations give the researcher information about the degree to which changes in the ratings of the dependent variable are due to actual changes taking place in the classroom or practice setting, rather than differences in the consistency of the ratings.[141]

Keep in mind that inter-observer agreement is not a measure of the *accuracy* of the observers' recordings; even if both observers are consistent in their ratings, they might not accurately record what took place during the observation. For this reason it is extremely important to train the observers carefully in the use of the observation protocol before sending them out to conduct observations. Kennedy (2005) pointed out three reasons why calculating inter-observer agreement is important:

- It can be used as a training standard for new observers;
- It allows the researcher to determine how consistent the observers are across observation occasions; and

[139] Bill & Melinda Gates Foundation, 2011.
[140] See, for example, Baecher & Connor, 2010 and Dymond & Bentz, 2006.
[141] Kennedy, 2005.

- It helps to avoid observer drift, which occurs when the people conducting the observations do not consistently apply the definitions laid out in the observation protocol.

Some researchers make video recordings of observations at the start of a study to use for training purposes and to help "recalibrate" observers periodically throughout the study. Another method to ensure that all raters stay calibrated throughout the period of data collection is to designate two raters to some percentage of observations (10 to 20%) and calculate the inter-rater reliability for those double-coded observations. Whether PDP evaluators are able to do this for their studies will depend on the resources available.

There are many ways to calculate inter-observer agreement, including correlational and complex statistical approaches. We focus here on simpler calculations that often are used in single-case research. We also include a method for calculating agreement using ratings of more than two levels (or more than two raters). These are briefly summarized below.[142]

4.3.2.6.1 Total Agreement Approach

A simple and popular approach has been to calculate total agreement between two observers. To do this, the researcher sums the total number of responses (or times that a behavior or event was observed) recorded by each observer. This typically generates a different total for each observer, a "smaller" and a "larger". To calculate total agreement, divide the smaller total by the larger total, and multiply the result by 100%. The formula is

$$\frac{S}{L}\,(100\%)$$

where *S* is the smaller total and *L* is the larger total.

This approach has three advantages:

- It is easy to conceptualize and calculate;
- It can be used to calculate inter-observer agreement in instances where observers have not accurately aligned their intervals (i.e., one observer began recording during the first interval, but the second observer did not begin until the second interval); and
- It is relatively sensitive to overall levels of responding (or occurrence).

A key limitation of the total agreement approach, however, is that it does not tell the researcher whether the observers actually agreed on the occurrence of individual instances of behavior. Consequently, there can be high levels of inter-observer agreement but the observers may have never

[142] For more information about calculating inter-observer agreement (including additional types of calculations) see Kennedy, 2005 and Watkins & Pacheco, 2000.

agreed that a specific behavior ever took place. This presents a challenge to interpretations of the observation data.[143]

4.3.2.6.2 Interval Agreement Approach

The interval agreement approach (also known as combined, point-by-point, or overall agreement), on the other hand, does take into account when behaviors or events actually occur. This type of calculation requires an observation protocol that uses an interval or event system of measurement to record behavior. Once the observation is complete, the recording of behavior is compared between the two observers on an interval-by-interval basis; if both observers recorded a behavior as occurring or not occurring in a particular interval, it is scored as an agreement. If only one observer recorded a behavior during an interval it is scored as a disagreement. Then the total number of agreements is divided by the total number of agreements plus disagreements, and the sum is multiplied by 100%. The formula is

$$\frac{A}{A+D}(100\%)$$

where *A* is the number of agreements and *D* is the number of disagreements.

Interval agreement is one of the most commonly-used measures of inter-observer agreement. However, when a behavior being observed occurs very frequently or very rarely, it is possible for there to be high levels of interval agreement even if the observers do not agree on the occurrence of a specific behavior. For example, if a behavior occurred only one time during an observation and the observers disagree on its occurrence, the interval agreement would still be very high. For this reason, researchers developed another measure of inter-observer agreement, discussed next.[144]

4.3.2.6.3 Occurrence/Non-Occurrence Agreement Approach

The occurrence/non-occurrence agreement approach is an even more stringent way to measure inter-observer agreement than the interval agreement approach. In this case, during each interval, agreement is calculated for each occurrence and non-occurrence of a behavior or event. The formula for this is the same as for interval agreement, except two separate calculations are conducted: one for occurrence and another for non-occurrence.

$$\frac{A\mathrm{O}}{A\mathrm{O}+D\mathrm{O}}(100\%)$$

where *A*O is the agreement on the occurrence and *D*O is the disagreement on the occurrence.

$$\frac{A\mathrm{N}}{A\mathrm{N}+D\mathrm{N}}(100\%)$$

[143] Kennedy, 2005 and Bryington, Palmer & Watkins, 2002.
[144] Kennedy, 2005.

where *A*N is the agreement on the non-occurrence and *D*N is the disagreement on the non-occurrence.

Each statistic is then reported separately. Since this is the most rigorous of the commonly used approaches to calculating inter-observer agreement, it is considered the preferred method.[145]

4.3.2.6.4 Cohen's Kappa

The kappa method indicates what proportion of agreement is above and beyond what would be expected by chance alone. It has a few advantages over the previous methods, including that it corrects for chance agreement and can be used with ratings on two or more levels (such as a 1-7 rating scale). It also has a modification that can be used for more than two different observers.[146] In addition, it allows for generalizability across different experimental conditions, for instance observations in different classrooms in different years, since it is not affected by the rates of behavior, as are the prior methods. Kappa ranges from -1.00 which indicates perfect disagreement, to 0.00 which indicates chance agreement, to 1.00 which indicates perfect agreement. A sample calculation for raters who rate 10 teachers on a behavior as low, medium, or high, is shown in Table 11 below.

Table 11. Sample Calculation of Cohen's Kappa Agreement

		Observer 1		
		Low	Medium	High
Observer 2	Low	2	1	1
	Medium	1	2	0
	High	0	0	3

As you can see in the Table, the ratings of the first observer agreed with those of the second observer in 7 cases. They both rated 2 teachers as low on this behavior, 2 teachers as medium on this behavior, and 3 teachers as high on this behavior. In the remaining 3 cases, the two raters disagreed on the rating of the teacher's behavior. Numbers such as these can be plugged into an online calculator[147] and the kappa will be obtained. In this case the kappa would be 0.55, meaning that the observers accounted for 55% agreement over what would be expected by chance. A kappa in this range would generally be considered only fair. Kappas between 0.40 and 0.59 are considered fair, those between 0.60 and 0.74 are good, and those above 0.75 are considered excellent.

---

[145] Bryington, Palmer & Watkins, 2002.

[146] See Bryington, Palmer & Watkins, 2002 for a discussion of this modification, called the Fleiss kappa and see https://www.easycalculation.com/statistics/cohens-kappa-index.php for an explanation and link to an online calculator.

[147] http://vassarstats.net/kappa.html and http://graphpad.com/quickcalcs/kappa1.cfm are online calculators that provide kappas, standard errors, and confidence intervals.

The primary limitations of the kappa are difficulty in calculating, which can be solved with online calculators, and inability to be used in a situation where all observers agree or disagree in all cases, which is unlikely to happen often, and would rather clearly be a case of perfect agreement or disagreement. Since the kappa is rigorous and provides generalizability over different conditions, it is the preferred method for calculating inter-rater reliability.

4.3.2.7 Analyzing Observational Data

Clearly, the type of observational protocol used will affect the type of analysis to be conducted. Returning to Figure 6 and Figure 7 above, the type of analysis will depend on the data recorded in the protocol. For instance, in Figure 6, if the observer records narrative information about what occurred during the class period, qualitative content analysis will be required. If, however, the observer records tallies of behaviors (e.g., student got up and walked around the classroom five times during the period) it will be possible to develop summary tables of the frequency of behaviors during the class period. In Figure 7, it is possible to tally the total number of intervals in which a student exhibited a certain behavior. It is also possible to calculate a proportion of the total class period in which the student was exhibiting the behavior. If the protocol asked the observer to record frequency of behaviors (such as with the aforementioned scale of 0 = never, 1 = sometimes, 2 = frequently), the analysis could provide a clearer picture of the level to which the student was exhibiting a certain behavior (e.g., frequently during 4 of the 5 intervals).

4.3.3 Individual Interviews

Individual interviews are the most costly and time-consuming interviews to conduct, as they require a researcher to speak directly with a respondent—either in-person, on the phone (e.g., direct call, teleconference, random digit dialing, or computer-assisted telephone interviewing [CATI]), or over the internet (e.g., through email, web chat or social media).

As with observations, interviews can be more qualitative (i.e., unstructured) or more quantitative (i.e., structured) in nature. Whether the interview is unstructured, structured, or semi-structured will affect the type of data analysis that will be required. Unstructured interviews generally will require extensive qualitative data analysis (see Section 4.4.4 for more information), while structured interviews will allow for more quantitative analytic approaches.

It is good practice to prepare an interview protocol in advance of the interviews and to practice asking the questions to someone familiar with the topic. This will help to clarify any issues that may arise with the question phrasing, highlight “difficult” or “sensitive” terms or topics, and identify any possible areas where misunderstandings may arise. It may be necessary to revise and re-test the interview protocol prior to conducting the interviews. In addition, all data collectors should be skilled in conducting interviews in a way that encourages respondents to answer the questions openly and honestly.[148]

[148] See Weiss, 1994, for an excellent discussion of the design and conduct of qualitative interviews.

Furthermore, to the extent possible (and as appropriate), the data collectors should try to cover all of the questions on the interview protocol with every potential respondent. Inconsistencies in the way that interviewers ask questions across respondents can lead to poor quality data and limit the ability of evaluators to draw conclusions or generate insights from their qualitative interview data.[149]

4.3.3.1 Unstructured Interviews

In an unstructured interview the interviewer allows the respondent to decide how she wants to answer a given question. These questions can be pre-determined or they can arise in the moment as probes to obtain additional information about a topic or an issue raised by the respondent; this generally depends upon the needs of the study and the skill of the interviewer.[150] In these types of interviews it is important to ask questions that encourage the respondent to elaborate, rather than to provide brief or "yes/no" answers. For example, consider the following questions:

- Question 1: Do you think that your preparation program provided you with the knowledge and skills necessary to be successful in your job?
- Question 2: In what ways do you think that your preparation program provided (or didn't provide) you with the knowledge and skills necessary to be successful in your job?

As can be seen, Question 1 allows the respondent simply to respond "yes" or "no," while Question 2 urges the respondent to list the ways that the program has prepared him to succeed in a job. After all of the interviews are conducted, the responses to Question 1 will provide little information beyond allowing the study team to tally the number of graduates who believe the program provided the requisite skills (i.e., the respondents who answered "Yes"). The responses to Question 2, in contrast, will allow the interviewer to gain a better understanding of the graduate's perceptions of the important (and not-so-important) aspects of the preparation program. Additionally, the open-ended responses will allow the interviewer to probe for additional details. In this case, after the interviews are completed, the study team can look for patterns across respondents with respect to which elements of the preparation program were most frequently cited as positive or negative influences on graduate performance on the job. While this information is generally formative in nature, the responses can help the study team to generate hypotheses about how certain elements of a preparation program affect graduate performance that could later be studied as part of a summative evaluation.

In unstructured interviews it is also important to avoid leading the respondent to answer a question in a specific way. Consider the following examples:

- Question 3: How satisfied are you with your preparation program?
- Question 4: How do you feel about your preparation program?

[149] Weiss, 1994; see also Patton, 2002.
[150] See, for example Creswell, 2002, and Patton, 2002.

For the purposes of an unstructured interview, Question 3 has a couple of issues. First, it does not encourage the respondent to elaborate; it would be easy for a respondent simply to answer “not very satisfied” or “very satisfied” to this question. Second, the wording of the question effectively limits the responses by identifying the dimension along which the respondent is expected to answer—level of satisfaction.[151] Question 4, on the other hand, both encourages the respondent to elaborate and allows the respondent to discuss any positive, negative, or neutral feelings about the program, whether those feelings relate to satisfaction, cost-effectiveness, quality, or some other dimension.

Significant skill is needed from the interviewer to conduct a successful unstructured interview. As with many skills, the ability to conduct a good unstructured interview can be trained, but again the study team must consider resources. Does the evaluation team include people with good interviewing skills or will the team need to receive training? How much time will be available to practice interviewing? For more information about considerations for conducting an unstructured interview, see Patton (2002).

4.3.3.2 Structured Interviews

In a structured interview, the interviewer asks the participant questions with close-ended response options. For example, an interviewer might ask the following question:

- Question 1: To what extent do you believe your preparation program gave you the knowledge you need to be successful in your job?
    a. Great extent
    b. Moderate extent
    c. Little extent
    d. Not at all

As can be seen, such structured questions do not allow the interviewer to gather additional information beyond that which is included in the interview protocol. Nevertheless, this inherent limitation also can be a strength in some situations, because much less skill is needed from the interviewer to conduct a structured interview compared to an unstructured one. Additionally, structured interviews require fewer resources (i.e., time and money) to collect and analyze the data, and responses can be compared relatively easily across respondents, allowing the study team to generate descriptive information about the interview responses such as tallies of specific responses and cross-tabulations.[152]

The semi-structured interview, discussed in the next section, provides additional flexibility and may be desirable when the study team wants to keep the interview relatively structured while allowing the respondent to elaborate on specific questions.

[151] Patton, 2002.
[152] Creswell, 2002.

4.3.3.3 Semi-Structured Interviews

Semi-structured interviews combine elements of the unstructured and structured interviews. Generally, semi-structured interviews are composed primarily of structured questions, with one or two unstructured questions that allow the respondent to elaborate on a particular topic or issue. For example, an interviewer might ask the following question:

- Question 1: To what extent do you believe your preparation program gave you the knowledge you need to be successful in your job?
  a. Great extent
  b. Moderate extent
  c. Little extent
  d. Not at all

  Please explain: ____
  ____
  ____
  ____
  ____

As in the case of structured interviews, less skill is needed to conduct a semi-structured interview compared to an unstructured one. Additionally, the semi-structured interview requires fewer resources to collect and analyze the data, although more than the structured interview.

4.3.3.4 Considerations for Collecting and Analyzing Interview Data

As with the other aspects of a study, interview protocols must be designed with the end in mind. The more unstructured an interview, the greater the burden on the study team in terms of data collection and analysis. When making decisions about what kinds of interview questions to ask, the study team must think carefully about what they are going to do with the data they collect. Although some researchers would say “More is better” when it comes to collecting data, there is no need to collect data that will not be used. If the general purpose of the interview is to generate a list of the key program elements that graduates perceive contributed to their success on the job, a structured or semi-structured interview will probably suffice. If the study team is interested in knowing how the field experiences of graduates from different fields (e.g., psychology, counseling, or secondary transition) compare, a semi-structured or unstructured interview may be better. However, the desire to ask in-depth, probing questions must be weighed against the study team’s ability to spend the time and money required to collect and analyze the resulting data.

We recommend recording the interview—after obtaining permission from the respondents—whenever possible. However, if no one on the study team will have the time or interest to listen to the recording (let alone transcribe it), there is no need to record the interview in the first place. Some benefits of recording an interview include:

- A recording device does not select the information to record (as human interviewers do),
- Recordings allow the study team to listen to the interview again to fill in any areas where the interviewer's notes may be unclear or incomplete, and
- A recording enables the study team to make a transcription of the interview at a later date for more in-depth analysis, if desired. (Note: Being able to record an interview does not mean that the interviewer should not take notes; a recording can always fail.)

Also, the respondent should be told that they can elect to turn the recording off after initially agreeing to record the interview, and may do so during particularly sensitive times or questions. If the respondent requests that information be "off the record," neither the written nor verbal notes taken during the interview can be used in data analysis.

For more information, Patton (2002) presents a very detailed and informative discussion of how to conduct interviews and collect and analyze interview data.

4.3.4 Focus Groups

A focus group[153] generally includes 6-10 individuals who have similar backgrounds, such as a cohort of graduates from a preparation program. When organizing a focus group it is important to remember that the discussion should be focused around a specific topic; group members should not be encouraged to explore a wide variety of topics. It is expected that the group members will influence and build upon each other's responses to the moderator's questions, but it is not primarily a discussion; the focus group is conducted in order to obtain answers to the questions outlined in a focus group protocol.[154]

As with all data collection activities, focus groups have benefits and drawbacks. Some of the **benefits of focus groups** relate to:

- **Cost-effectiveness**—It is a quicker way to get information from multiple people than an individual interview.
- **Data quality**—Interactions among participants serve as checks and balances on false or extreme views.

[153] Some researchers make a distinction between focus groups and group interviews, pointing out that the way a group interview is conducted is generally different from a focus group. That is, in group interviews the interviewer tends to have a greater role (e.g., generally by directing the line of questioning) while in focus groups the interviewer acts as more of a mediator of the group discussion (Patton, 2002). We do not believe this distinction is particularly important for our discussion here.
[154] Patton, 2002.

- **Consistency of viewpoints**—It is relatively easy to see the extent to which there is consistency or great diversity of individual views among group members.
- **Participant experience**—Focus groups are generally viewed as enjoyable for the group members.[155]

Conversely, some of the **drawbacks or limitations of focus groups** concern:

- **Interview questions**—The number of questions that can be asked and answered in a focus group is limited. (Note: For a group of 8 and a 1-hour focus group, plan to ask no more than 10 major questions.)
- **Individual responses**—The moderator must limit individual responses in order to hear from all focus group participants.
- **Moderator skills**—To successfully manage a focus group, the moderator must be adept at managing group processes so as to prohibit one or two individuals from dominating the group.
- **Minority viewpoints**—People whose views might be in the minority might choose not to speak up to avoid negative reactions of other group members.
- **Confidentiality**—It is not possible to ensure confidentiality of responses in focus groups.[156]

Social media and the internet make conducting focus groups with individuals spread out across multiple locations relatively easy. A focus group could be conducted over Skype or another similar internet service, through instant messaging, or through social media. This would greatly reduce the time and cost associated with conducting a focus group, as the group members and the mediator could participate virtually, rather than traveling to attend an in-person meeting.

There is no guideline regarding the number of focus groups to conduct—this decision, too, will be driven by the needs of the study and the available resources.

4.3.5 Goal Attainment Scaling (GAS)

Goal attainment scaling (GAS) is a technique to quantify the achievement (or lack of achievement) of goals set,[157] but it also qualifies as a data collection method, which is why it is included here. Initially developed by Kiresuk and Sherman (1968) to evaluate comprehensive community mental health programs, GAS is currently being used by professionals in a variety of fields. GAS has been used to evaluate the effectiveness of individual interventions in producing change, as well as to evaluate program effectiveness.[158] For example, GAS has been used to evaluate special education programs[159]

[155] Ibid.
[156] Ibid.
[157] Turner-Stokes, 2009.
[158] Sladeczek, Elliott, Kratochwill, Robertson-Mjaanes, & Stoiber, 2001.

and an early childhood intervention project.[160] It also has been used to set, monitor, and evaluate goals for individuals with learning disabilities,[161] cognitive disabilities,[162] autism spectrum disorders,[163] and traumatic brain injuries.[164]

The basic methodology of goal attainment scaling (GAS) involves (a) selection of the target behavior, (b) an objective description of a desired intervention outcome, and (c) the development of multiple descriptions of the target behavior.[165] The processes for developing the GAS goals and creating the scale are described in more detail below.

4.3.5.1 Developing the GAS Goals

In the context of GAS, goals have two characteristics. First, a goal is an intended future state, and will usually involve a change from the current situation—although maintaining the current state in the face of expected deterioration could also be a goal. Second, a goal refers to the result of actions undertaken by the actors involved in an intervention, such as a teacher and her students.[166] Good GAS goals are individualized; challenging, but realistic and achievable; able to be written without too much effort, time, or specific training; and flexible enough to cover most situations. Also, they allow for accurate, unambiguous determination of goal achievement. To facilitate this, four aspects should be considered and incorporated into the definition of the goal: the target activity, the support needed, quantification of performance, and the time period to achieve the desired state.[167] Bovend'Eerdt et al. (2009) presented a flowchart for writing goals in GAS that outlines a process for developing goals that takes these four aspects into consideration.

Some authors argue for the establishment of goals by one or two specified "goal selectors" who are responsible for selecting goals and creating a scale for individuals based on a pre-specified intervention plan.[168] Others call for a more collaborative approach, whereby members of a team (e.g., a counselor, a teacher, a parent, and a student) work together to identify the main problem areas and establish priority goals for achievement by an agreed date.[169] Whichever approach one takes, an explicit goal needs to be established and this goal will serve as the criterion for success of the intervention.[170]

---

[159] E.g., Carr, 1979; Maher, 1983.
[160] Barnett et al., 1999.
[161] Glover, Burns, & Stanley, 1994.
[162] Bailey & Simeonsson, 1988.
[163] Oren & Ogletree, 2000.
[164] Mitchell & Cusick, 1998.
[165] Kiresuk & Sherman, 1968; Morrison, 2012; Sladeczek et al., 2001.
[166] Wade, 2009.
[167] Bovend'Eerdt, Botell, & Wade, 2009; Morrison, 2012.
[168] Kiresuk & Sherman, 1968.
[169] Sladeczek et al., 2001; Turner-Stokes, 2009.
[170] Sladeczek et al., 2001.

4.3.5.2 Constructing the Goal Attainment Scale

The GAS ratings should include "sufficiently precise and objective descriptions to enable an unfamiliar observer to determine whether the [individual] lies above or below that point."[171] The literature generally agrees upon the use of a five-point scale ranging from the least favorable (-2) to most favorable (+2) outcome to measure each goal,[172] although some have argued for the use of a seven-point scale (-3 to +3) to allow for greater sensitivity to change in the achievement of goals.[173] Some scales are constructed such that the "0" rating represents "no change in behavior/performance," while in other scales the "0" rating represents the "expected outcome" of an intervention.[174] These decisions should be made in advance to ensure consistency of ratings for each goal. Additionally, Kiresuk and Sherman advise that "the scale points be stated in terms of events the presence or absence of which can be easily judged by a follow-up worker who has had no contact with the [intervention procedures]."[175]

In general, a goal attainment scale should look something like this:

Behavior/Performance Rating:
+2: Change Much More than Expected
+1: Change Somewhat More than Expected
0: Change At Expected level
-1: Change Somewhat Less than Expected
-2: Change Much Less than Expected

-OR-

Behavior/Performance Rating:
+2: Much Positive Change
+1: Some Positive Change
0: No Change (e.g., anchored to baseline data)
-1: Some Change in Wrong Direction
-2: Much Change in Wrong Direction

Roach and Elliott (2005) presented a list of characteristics or dimensions to facilitate the development of descriptions for the different GAS ratings including, among others, frequency (never, sometimes, very often, almost always, always), amount of support needed (totally dependent, extensive assistance, some assistance, limited assistance, independent) and development (not present, emerging, developing, accomplished, exceeding). They also created a simple template for developing goal attainment scales that can serve as a helpful example for anyone using GAS in an evaluation.

GAS offers the option to weight goals to account for their relative importance or difficulty. Turner-Stokes (2009) offered suggested weighting scales for goal importance and difficulty.

Importance	Difficulty
0 = not at all (important)	0 = not at all (difficult)
1 = a little (important)	1 = a little (difficult)
2 = moderately (important)	2 = moderately (difficult)
3 = very (important)	3 = very (difficult)

171 Kiresuk & Sherman, 1968, p. 445.
172 Kiresuk & Sherman, 1968; Kiresuk, Smith & Cardillo, 1994; Sladeczek et al., 2001; Turner-Stokes, 2009.
173 Bovend'Eerdt et al., 2009.
174 Kiresuk & Sherman, 1986; Kiresuk et al., 1994; Roach & Elliott, 2005; Turner-Stokes, 2009.
175 Kiresuk & Sherman , 1968, p. 447.

To determine the weight, simply multiply the importance and the difficulty:

Weight = Importance x Difficulty

The weights can then be incorporated into the mathematical formula for calculating standardized GAS scores (discussed below).

4.3.5.3 Collecting GAS Data

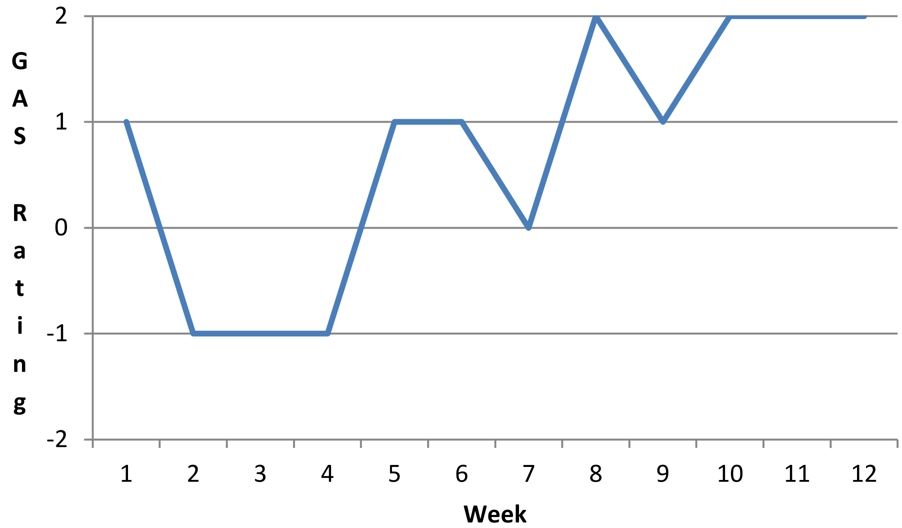
GAS can be applied to "any form of objectively determinable event."[176] For example, goal attainment scales can be developed for data collected through a psychometric instrument, public records, self-ratings, and autobiographical reports. In the context of an evaluation of a PDP project, data collected through student and teacher self-ratings, principal assessments, student IEPs, student learning objectives, and student assessment scores can all be measured using goal attainment scaling. "*Student GAS self-ratings can function as either self-monitoring (a form of direct assessment, completed as behavior occurs) or self-report (a less direct measure of an individual's perception of their behavior). GAS ratings completed by teachers and other adults function as informant reports, which are also indirect reports because they represent an observer's retrospective perceptions of behavior.*"[177]

Roach and Elliott (2005) recommended conducting initial assessments of goal attainment to establish a baseline for behavior/performance, and then assigning the score of 0 to the baseline data.[178] GAS data can be collected daily, weekly, monthly, or in other time intervals (e.g., every six months), depending on the nature of the intervention and the goal.

GAS ratings can be plotted and viewed as a measure of intervention-induced change.[179] Figure 8 below is an example of GAS ratings that have been plotted to show change over time.

[176] Kiresuk & Sherman, 1968, p. 447.
[177] Roach & Elliott, 2005, p. 9.
[178] See also Kratochwill, Elliott, & Rotto, 1995.
[179] Morrison, 2012; Roach & Elliott, 2005; Sladeczek et al., 2001.

Figure 8. Sample GAS Ratings for a Special Education Student



As can be seen in the above example, establishing a baseline when developing the GAS rating scale such that 0 represents the status at baseline will better illustrate changes in performance due to an intervention. Ratings of interventionists, teachers, parents, and even students can all be plotted on the same graph to give an overall picture of performance as perceived by the different individuals involved. The GAS ratings can also give a common language to communicate the effects of an intervention for a given time interval.[180]

4.3.5.4 Calculating a Standardized GAS Score

In some situations practitioners may want to compare GAS scores with other measures of intervention outcomes that are presented in *T*-scores. Kiresuk and Sherman (1968) developed a mathematical formula to derive standardized *T*-scores (with a mean of 50 and a standard deviation of 10) which allow for comparison across individuals and contexts. The standardized composite goal attainment score equals the sum of the goal attainment levels times the relative weights for each goal. It is calculated as follows:

$$\text{Overall GAS} = 50 + \frac{10\sum(W_i X_i)}{\sqrt{(1-\rho)\sum W_i^2 + \rho(\sum W_i)^2}}$$

Where W_i is the weight assigned to the *i*-th goal (if equal weights, $W_i = 1$), X_i is the numerical value achieved (between -2 and +2 on a five-point scale), ρ is the expected correlation of the goal scales. Cardillo and Smith (1994)[181] found that there are high correlations among weighted and unweighted scores, so it is appropriate and simpler to use unweighted scores (i.e., all scores have a weight of 1). In

[180] Sladeczek et al., 2001.
[181] Cited in Turner-Stokes, 2009.

addition, Kiresuk and Sherman (1968) found that, for practical purposes, ρ most commonly approximates 0.3, so the equation can be simplified to:

$$\text{Overall GAS} = 50 + \frac{10\sum(W_iX_i)}{\sqrt{0.7\sum W_i^2 + 0.3(\sum W_i)^2}}$$

To reduce the need for calculations, the book by Kiresuk et al. (1994) includes score calculation tables for easy reference and Turner-Stokes (2009) developed a GAS spreadsheet calculator in the form of an excel worksheet available online at:
http://www.kcl.ac.uk/lsm/research/divisions/cicelysaunders/resources/tools/gas.aspx.

However, Sladeczek et al. (2001) pointed out that “for most practitioners, such a transformation may be somewhat cumbersome, and the transformation to standardized *T* scores does not add additional information to what the overall goal attainment score is conveying (i.e., change due to the intervention; p. 51). Nevertheless, this transformation may be useful if practitioners want to compare scores across multiple interventions.

4.3.5.5 Strengths/Advantages of GAS

Goal attainment scaling can be a useful approach for measuring outcomes for PDP project evaluations because the data can be readily collected by special education teachers and related-services professionals in the course of their practice. “GAS promotes clearly operationalized intervention goals and on-going (i.e., time-series) evaluation of student progress, making it a potentially useful tool for special educators and school psychologists working within a responsiveness-to-intervention (RTI) model of special education identification.”[182] Additionally, GAS communicates expectations for student performance and enables the practitioner to apply systematic rigor to assessment while focusing on the particular needs of the student.[183] Further, the process of setting goals, in and of itself, may have a positive effect on intervention outcomes.[184]

[182] Roach & Elliott, 2005, p. 16.
[183] Morrison, 2012.
[184] Evans, 1981.

There are a number of **strengths or advantages of using GAS ratings** to monitor students' academic growth:

- Once the scale is created, collecting GAS data is time efficient
- GAS:
  - is flexible, allowing for the scaling of various content domains for a student (i.e., personalized-individualized);
  - is conceptually consistent with behavioral consultation;
  - requires minimal skills to collect data;
  - is a nonintrusive assessment method;
  - can be used as a self-assessment;
  - can be used by multiple informants across settings (e.g., home, school, community);
  - can be used repeatedly to monitor perceptions of intervention progress;
  - can be used to document perceptions of intervention outcomes;
  - is relatively inexpensive to implement once the scale is created; and
  - requires minimal skill to interpret data.[185]

4.3.5.6 Limitations/Disadvantages of GAS

Probably the biggest limitation of GAS for evaluators of PDP projects is the time and effort required to develop good goals and scale definitions. A significant amount of time is needed to construct a goal attainment scale and many evaluators may not have the resources available to devote to the task. Once the scale is developed it can be used repeatedly (assuming the goals are applicable in other situations), but the initial development of multiple scales to be used in a PDP project evaluation may not be feasible without significant resources dedicated to the task.

Another potential issue is the concern that graduates might choose to set goals that are relatively easy for their students to achieve, simultaneously lowering their expectations for student performance and giving an inaccurate picture of their students' abilities. Without the establishment of formal procedures for checking the rigor of goals set, it would be easy for graduates to collect GAS data that are overwhelmingly favorable.

Roach and Elliott pointed out other limitations or disadvantages to the use of GAS.

- There is limited published, empirical research on the school-based use of the method.
- GAS is a subjective summary of observations collected over time.
- Goals are not norm-referenced.

[185] Roach & Elliott, 2005.

- The guidelines for interpretation of performance are determined by parties involved with the intervention, thus subject to bias.
- GAS is a global (i.e., less discrete) accounting of behavior.[186]

Other disadvantages/limitations include:

- GAS has utility for monitoring and evaluating progress toward a specific goal, but it is not designed to establish an absolute level of functioning for clients. For example a student with autism may change a specific maladaptive behavior as a result of an intervention, but not have significantly increased his or her absolute level of adjustment, skills, or functioning;[187] and
- It is not appropriate to establish causal relationships between independent and dependent variables in GAS.[188]

We now turn our discussion to the various methods that might be used in analyzing data collected through a PDP project evaluation.

4.4 Data Analysis Methods

It is beyond the scope of this Toolkit to discuss the myriad methods of quantitative and qualitative data analysis. In this section we focus on basic methods of quantitative and qualitative analysis that can readily be applied in a PDP project evaluation. Before applying any analysis method, however, evaluators first have to identify, and respond to, the prevalence of missing data, discussed below.

4.4.1 Dealing With Missing Data

Almost every study will have at least some missing data, so evaluators must have a plan for dealing with missing data in the analysis. Addressing the problem of missing data is particularly important when conducting inferential statistical analyses, but having missing data in a study can potentially affect the results, no matter what type of analysis is being conducted. The strategy that is adopted to deal with missing data will depend in part on the nature of the data that are missing. Howell (2012) pointed out several reasons why data might be missing:

- **Data missing completely at random**—The probability that an observation (X_i) is missing *is unrelated* to the value of X_i or to the value of any other variables. If data are missing completely at random, the analysis will remain completely unbiased, since the estimated parameters are not biased by the absence of these data.

[186] 2005, p. 15.
[187] Smith & Cardillo, 1994.
[188] Sladeczek et al., 2001.

- **Data missing at random**—The "missingness" of the data does not depend on the value of X_i *after controlling for another variable*. That is to say, if missingness is correlated with other variables in the analysis, the data are considered missing at random. The presence of data that are missing at random can bias the estimates if steps to deal with the missing data are not taken.

- **Data missing not at random**—The probability that an observation (X_i) is missing *is related* to the value of X_i or to the value of another variable. When data are missing not at random it is necessary to write a model that accounts for the missing data.

It is not within the scope of this Toolkit to go into detail about the different methods to treat missing data. The U.S. Department of Education provides guidance on what to do with missing data in experimental studies (see http://ies.ed.gov/ncee/pubs/20090049/index.asp). For additional information, see Howell (2012), Allison (2001) and Baraldi and Enders (2010).

4.4.2 Quantitative Analysis

As we described in Section 3.4.1, prior to conducting quantitative analysis evaluators need to enter the data into a database and prepare them for analysis. Once all of the data are entered into the database and the analysis program has been selected (e.g., SPSS, SASS, STATA or R for quantitative analysis), it is time to begin exploring and conducting descriptive and inferential analysis of the data, discussed below.

4.4.2.1 Descriptive Analysis

Descriptive statistics provide information about the overall trends and distribution of the data. This includes reporting percentages or frequencies for nominal and ordinal data, as well as conducting exploratory analysis of interval and ratio data. Exploratory analysis includes looking for outliers in the data, determining the modality (e.g., unimodal or bimodal) of the data distribution, and calculating kurtosis and skewness.[189] The analysis then proceeds with measures of central tendency, measures of variability, and measures of relative standing (as appropriate for data type).

If desired, if the data distribution is flat or the data are skewed, it may be possible to conduct linear or non-linear transformations of the data at this time—for example to change a skewed distribution to a more normal one. Evaluators might consider conducting transformations if they wish to use the data with inferential statistics and need to satisfy the assumptions of specific statistical tests.

[189] For information on how to do these procedures in SPSS, see Dimitrov, 2010.

After exploring the data and examining the shape of the distribution, it is also good practice to calculate the following descriptive statistics:

- **Measures of central tendency**, which give the researcher an idea of where the majority of scores are located in a distribution, when working with interval or ratio data. These include the mean, the median, and the mode.
  - Mean: The arithmetic average of a set of scores.
  - Median: The score above and below which 50% of scores fall.
  - Mode: The most frequently occurring score in the distribution.

- **Measures of variability**, which indicate the spread of the scores in a distribution, also appropriate when working with interval or ratio data. Three key measures of variability are the range, the variance, and the standard deviation.
  - Range: The distance between the highest and lowest score in a distribution.
  - Variance: A measure of how far the scores in the distribution are spread out around the mean.
  - Standard deviation: The average distance of scores from the mean (the square root of the variance).

Depending on the type of analysis and the study questions, evaluators also may want to calculate **measures of relative standing**, which are statistics that describe how a particular score compares to a group of scores. Three common measures of relative standing include the *z*-score, the percentile, and the percentile rank—all assume the data are interval or ratio in nature.[190]

- *Z*-score: A standardized score that shows how many standard deviations a score falls above or below the mean. The z-score has a mean of 0 and a standard deviation of 1. Raw scores can be converted to z-scores so that scores from a variety of measures can be compared.

- Percentile: The score below which a certain percent of scores fall.

- Percentile rank: The percent of scores that fall below a given score.

For some of the study questions (e.g., To what extent do PDP graduates report satisfaction with their field experience?), frequency distributions or descriptive analyses may be all that are needed. For more complex questions, however—particularly those related to the summative evaluation questions—evaluators likely will need to conduct some type of inferential statistical analysis, discussed briefly in the next section.

[190] Creswell, 2002.

4.4.2.2 Inferential Statistical Analysis

Inferential statistics are divided into two types: parametric and non-parametric. *"Parametric statistics are statistical tests based on the premise that the population from which samples are obtained follows a normal distribution and the parameters of interest to the researcher are the population mean and standard deviation."*[191] Parametric statistics are based on certain assumptions, including:

1. The variables are measured in interval or ratio scales.
2. Scores from any two individuals are independent—one person's score is not dependent on another person's score.
3. The distribution of population scores is normally distributed.
4. When two or more groups are involved in the study, each group representing a different population, the populations have equal variances.[192]

These assumptions typically hold true when large numbers of individuals are included in the sample; when the sample is small, however, one or more of these assumptions can be violated, thereby affecting the inferences that can be drawn from the analyses.

Non-parametric statistics have fewer assumptions than parametric statistics, making them more appropriate when the sample is small or when the data are not normally distributed. In fact, the only assumption that applies to non-parametric statistics is the assumption of independence of scores (number two in the list above). Non-parametric tests are typically not as sensitive as parametric tests, however, so whenever possible we encourage evaluators to use parametric tests. Additionally, the typical parametric tests are able to withstand violations of some of the assumptions, so in many cases it is reasonable to use a parametric test even when some of the assumptions presented above are violated. We recommend evaluators without a strong quantitative analysis background consult a statistician for help determining which statistic to use. When designing or identifying data collection instruments, evaluators should consider the types of data that will be produced by each instrument so that they can be sure to collect enough of the right data for the specific analysis method chosen.[193]

Creswell (2002) highlighted seven factors that go into the decision of what type of statistic to use:

1. Whether the evaluation wants to compare groups [or performance at different time periods] or relate one or more variables;
2. The number of independent variables;
3. The number of dependent variables;

[191] Creswell, 2002, p. 237.
[192] Ibid.
[193] Ibid.

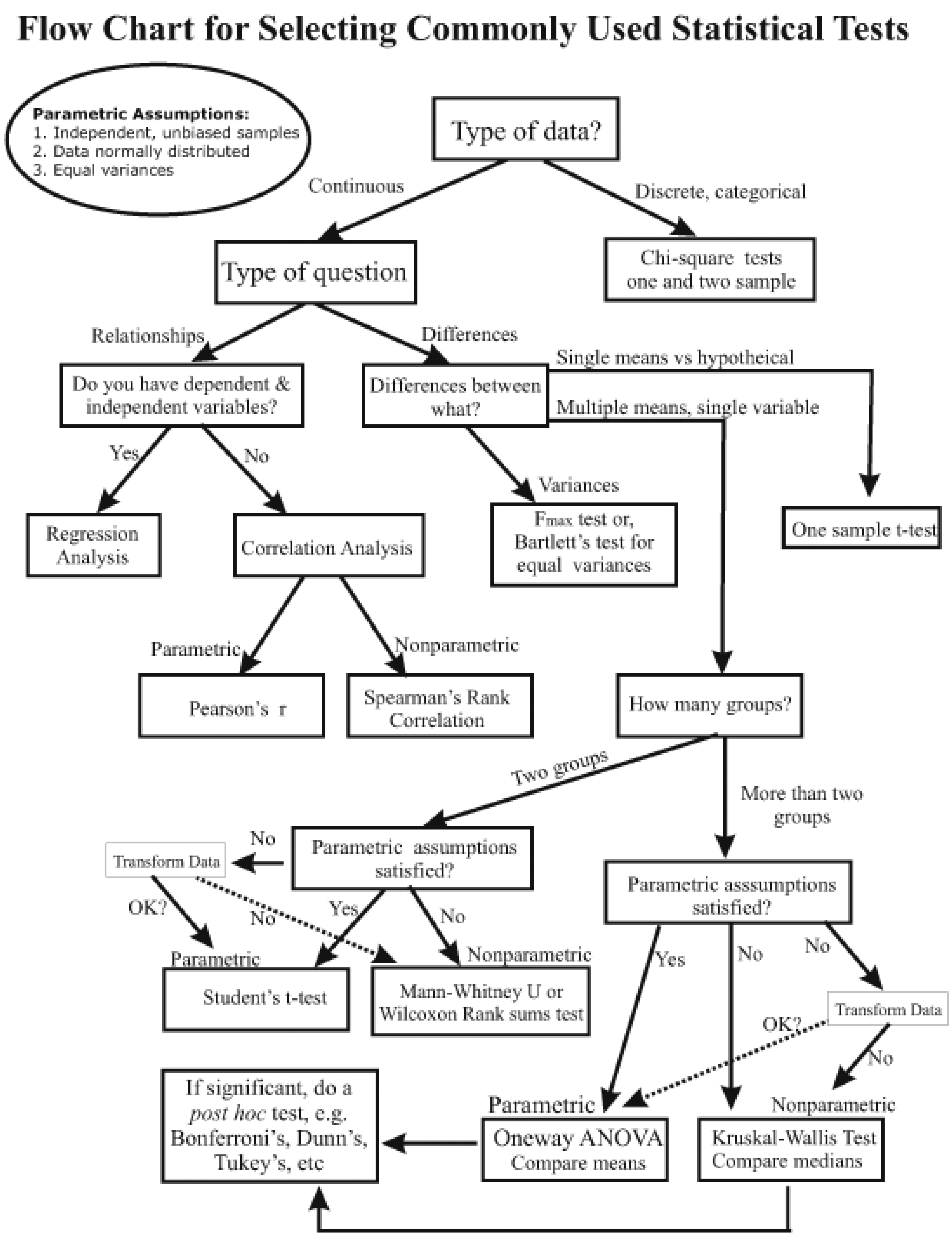
4. Whether covariates will be included in the analysis, and if so, how many;
5. The scale of measurement for the independent variable;
6. The scale of measurement for the dependent variable; and
7. Whether the data are normally distributed, bimodal, or skewed.

Carefully considering each of these factors will help evaluators to select the most appropriate and feasible test to conduct. Table 11 outlines some common parametric and non-parametric statistical tests that might be used in an evaluation and Figure 9 presents a decision tree to further help with the decision making. Box 2 presents an example of how this might be done in an evaluation.

Table 12. Common Parametric and Non-Parametric Statistical Tests

Type of Question/Hypothesis	Parametric Test	Non-parametric Test
Group comparison (2 groups)	*T*-test (independent samples) Also known as Student's *t*-test	Mann-Whitney *U* test
Within-subject comparison	*T*-test (dependent samples)	Wilcoxon Signed-Rank Test
Group comparison (3+ groups)	Analysis of variance (ANOVA)	Kruskall-Wallis test
Group comparison (3+ groups, multiple measures)	Repeated measures ANOVA	Friedman's Chi-square (χ^2) test
Relate variables	Pearson product moment correlation (Pearson *r*)	Spearman's *rho* rank correlation
Within-group comparison of frequencies		Chi-square (χ^2) goodness of fit test
Comparison of frequencies between groups or variables		Chi-square (χ^2) test for association

Figure 9. Decision Tree for Inferential Statistics



Source: http://abacus.bates.edu/~ganderso/biology/resources/statistics.html#whichtest

Box 2. Determining which Inferential Test to Use

Using the decision tree for inferential statistics above, let's examine one of the study questions from the sample Evaluation Plan presented in Appendix A.5 to determine which type of test might be used.

Sample Question: To what extent do graduates demonstrate in the workplace the high quality skills and knowledge needed to improve outcomes for children with disabilities? We will compare graduates of our program to those from a non-PDP program.

1. Type of data. The data we are using for our outcome measures is teacher workplace performance. To develop this outcome variable a continuous score was derived by summing the following data sources:
    a. the results of the questions in a supervisor survey related to teacher performance on the Council for Exceptional Children (CED) standards; and
    b. ratings from classroom observations. (Alternately the evaluator could maintain these as two dependent variables and conduct two analyses.)
2. Type of question. We are concerned with the differences, if any, between our groups of graduates.
3. Differences between what? We will have two means, one for each graduate group, on the variable of interest, the teacher workplace performance score we created.
4. How many groups? We have our PDP graduates and the comparison group of non-PDP SPED teachers, so analysis will be comparing performance of two groups.
5. Are the parametric assumptions satisfied? Let's consider each assumption in turn.
    a. Are the samples independent and unbiased? Yes.
    b. Is the data normally distributed? Descriptive data indicates "Yes."
    c. Do the groups have equal variance? Descriptive data indicates "Yes."

Based on this information, an independent samples *t*-test (Student's t-test) could be used.

For additional information on the use of these tests, Hinkle, Wiersma, and Jurs (2003) and Dimitrov (2010) are good reference books. Additionally, Rice University, the University of Houston Clear Lake, and Tufts University have developed an online statistics book that is free and available to the public at http://onlinestatbook.com/2/index.html.

4.4.2.3 Multilevel Analysis

An important feature of most educational research and evaluations is that the data typically have a multilevel structure. That is, data that are collected at one level (e.g., student outcome data) are frequently "nested" within clusters at different levels (e.g., within classrooms, within schools). Multilevel analysis examines relationships between variables measured at different levels of the multilevel data structure.[194]

T**wo basic multilevel designs are frequently used in education**:

- **The hierarchical design**—Entire clusters are assigned to treatment (e.g., entire schools or classrooms are assigned to treatment or control groups); and
- **The block design**—Individuals within the same cluster are assigned to two different treatments (e.g., students assigned to treatment and control groups within the same school).

It is beyond the scope of this Toolkit to delve into this topic; however, evaluators planning a PDP project evaluation that features an experimental or quasi-experimental design and intends to measure the outcomes of students of PDP graduates should consider using multilevel analysis to account for the clustering of students within schools. Evaluators without training in this subject should consult a statistician for guidance on how to design the evaluation and conduct the multilevel analyses. Hox (2010) provides a nice introduction to multilevel analysis; Bryk and Raudenbush (1988); Raudenbush (1997) and Schochet (2009) provide more technical details on how to conduct multilevel analysis.

In the next section we move our discussion to analysis of data in single-case designs.

4.4.3 Analysis of Data in Single-Case Designs

Evaluation of outcomes in single-case studies can be conducted through a variety of means, most typically including visual analysis of the data or calculation of summary statistics such as the percentage of non-overlapping data points[195] and the standard mean difference effect size. Additionally, it is possible to conduct a randomization test, which is a specific type of statistical analysis of data from single-case or small *n* studies (e.g., studies with fewer than 10 participants).[196] We briefly discuss these below. For more information about data analysis in single-case studies, see Barton & Reichow (2012); Horner & Spaulding (2010); Kennedy (2005); Kratochwill, Hitchcock, Horner, et al. (2013); Kratochwill & Levin (2010); and Todman & Dugard (2001).

---

[194] Hox, 2010 and Hedges & Rhoads, 2013.
[195] Scruggs, Mastropieri & Casto, 1987.
[196] Todman & Dugard, 2001.

Box 3 Example Using Single Case Design and Goal Attainment Scaling to Measure Teacher Outcomes

In May 2012 Julie Morrison from the University of Cincinnati made a presentation to the 325K PDP Working Group about her experience using goal attainment scaling and single-case designs to evaluate the performance of nearly 100 interns enrolled in a PDP. In that study, interns were expected to collect data on student performance for each of six interventions. Then, they were required to:

- conduct a visual analysis of the data for all six interventions;
- submit GAS data for each of the six interventions; and
- calculate effect size (ES) and percentage of non-overlapping data (PND) as a supplement to the GAS score, where such calculations were appropriate.

At the end of each school year, the study team had GAS data for nearly 600 interventions and effect size and percentage of non-overlapping data for a subset of these interventions.

4.4.3.1 Visual Analysis of Single-Case Data

As data are collected they are plotted in a graph (as in Figure 3 and Figure 4 in Section 4.1.4 above) and patterns are studied over the course of the study. Ideally, a person should be able to look at a graph of data collected from a single-case design and understand what the data represent. When conducting visual analysis of the data, the researcher is assessing the following aspects of the data:

- Level
- Trend
- Variability
- Immediacy
- Consistency
- Percentage of Overlap

When examining **changes in level**, the researcher calculates the mean or the median for the data within a given condition (e.g., for the A and B phases of an A-B-A-B design). This allows for the estimation of the central tendency of the data during a particular part of the experiment as well as a comparison of patterns between phases. The last few data points in each phase "*contain the most essential information regarding the level of behavior before a phase change*"[197] because they give clues to any changes in pattern. If these last few data points within a given phase appear to be showing an upward (or downward) trend, there is reason to consider that any changes in level in the subsequent phase may be part of a natural progression that started in the prior phase.

[197] Kennedy, 2005, p. 197; see also the WWC standards for single-case designs (Kratochwill, Hitchcock, et al., 2010).

Changes in trend of the data refers to the best-fit straight line that can be placed over the data within a phase. Two elements of the trend are important: slope and magnitude. The slope tells the direction of the change within a phase while the magnitude indicates the size or extent of the slope (generally estimated qualitatively as high, medium, or low). (Note: The greater the slope, the less meaningful changes in level are to the general estimate of the data pattern.) To judge the trend of the data visually, the researcher simultaneously estimates the slope and the magnitude of the data.[198] It is also possible to estimate the trend quantitatively; see Kennedy (2005) for more information.

The extent of **changes in variability** shows the degree to which individual data points deviate from the general trend of the data. To estimate variability, the researcher examines the degree to which the data points are spread out from the best-fit straight trend line. Variability is typically referred to as high, medium, or low.[199]

When looking at the **immediacy** of an effect, the researcher looks at how quickly a change is noticeable when moving from one phase to the next. This is done by examining last 3 data points in one phase and the first 3 in the next to see the magnitude of the change.

When examining **consistency** of the data patterns, the researcher looks at the extent to which data patterns are similar in similar phases.

We talk about a common method for calculating **percentage of overlap** in the next section.

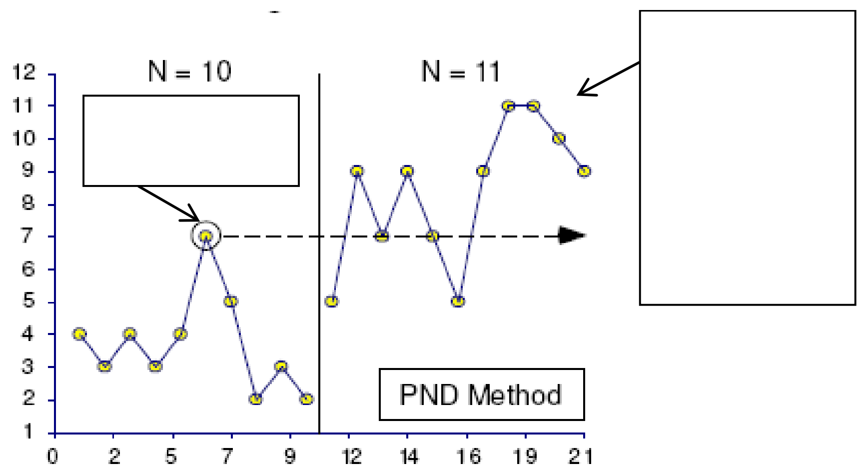
For information on how to create single-case design graphs using Microsoft Excel, see Barton & Reichow (2012).

4.4.3.2 Percentage of Non-overlapping Data Points

The calculation of percentage of non-overlapping data points (PND) involves identifying the highest point in the baseline data and determining the percentage of data points in the subsequent intervention phases that exceed the highest baseline data point.[200] Figure 10 illustrates how this is conducted. In this example the PND = 7/11 = 60%.[201]

[198] Kennedy, 2005.
[199] Ibid.
[200] Scruggs et al., 1987.
[201] Wendt, 2009.

Figure 10. Calculating Percentage of Non-Overlapping Data Points



Source: Wendt, 2009

Scruggs, Mastropieri, Cook and Escobar (1986) outlined specific guidelines for interpreting PND scores. In general, the higher the percentage of non-overlapping data points, the more effective the treatment. If a study includes multiple single-case experiments, the PND scores are aggregated by calculating the median, rather than the mean. The median is used in this case because scores in single-case designs are not usually distributed normally and it is less subject to outliers than the mean. While the PND is relatively easy to interpret, it has two major limitations. First, it ignores the majority of the baseline data and may be subject to ceiling effects as the study participant reaches the top of the performance scale. Next, it lacks sensitivity and ability to discriminate performance as the PND nears 100%.[202]

4.4.3.3 Standard Mean Difference Effect Size

An effect size is a descriptive statistic that provides an estimate of the magnitude, or strength, of group differences or of the relationship among variables. Effect sizes are commonly calculated in meta-analyses to summarize findings in a specific area of research and to allow for comparisons across studies. Further, the American Psychological Association (2009) recommends reporting effect sizes along with the results of significance tests and confidence intervals to give more information about the magnitude or importance of a study's findings. There are many ways to calculate effect sizes, including regression based and non-regression based effect sizes. Each type of effect size calculation has advantages and disadvantages, but not all are considered user-friendly for practitioners in terms of their

[202] Wendt, 2009.

simplicity of calculation and interpretation.[203] In this section we focus on a particular type of non-regression-based effect size—the standard mean difference.[204]

The **standard mean difference** (SMD) effect size is calculated by determining the difference between the mean baseline score and mean intervention score and then dividing by a standard, such as the standard deviation of the baseline.[205] There are two variations of SMD. Specifically, SMD may be calculated using the mean for all baseline and intervention data points (SMDall) or it may be calculated using the mean from only the last three data points of each phase (SMD3) and then multiplying by 100. This latter calculation is also known as the percentage reduction or the mean baseline reduction. Olive and Smith (2005) recommend the use of SMDall to complement visual analysis because it yields similar results to other summary statistics, it is relatively simple to calculate, and it does not have some of the same assumptions as regression-based effect size models (i.e., regression-based models are based on the assumption that linearity exists).[206]

Cohen (1988) suggested that a small effect size is $d = 0.2 – 0.49$, a medium effect size is $d = 0.5 – 0.79$, and a large effect size is $d \geq 0.8$. This standard has been widely used by researchers and evaluators in the field of education, despite criticisms of its application to education interventions. Rather than using Cohen's broad categorization of the magnitude of effect sizes across intervention research as a whole, we recommend evaluators consult Lipsey et al. (2012), who have outlined specific norms for effect sizes for educational interventions.

Campbell (2004) pointed out two statistical problems that affect the calculation of effect sizes in single-case studies. First, single-case data are auto-correlated, meaning that observations are temporally ordered and usually not independent. This violates one of the basic assumptions of most statistical tests of significance. Second, effect sizes can be confounded by a trend in the data, thereby causing the effect size to be under or overestimated.[207] Another problem with using only the SMD effect size is that valuable information inherent in the visual analysis of the data is lost (i.e., data variability, trend magnitude and direction, mean levels and shifts, and embedded cycles within phases).

4.4.3.4 Randomization Tests

Single-case and small-*n* designs cannot use the same parametric statistical procedures that are typically used in larger studies, since the smaller the sample size the less confidence a researcher can have that the parametric assumptions are met.[208] Instead, non-parametric analyses are recommended for studies with a relatively small number of participants. These non-parametric tests are generally based on rankings of scores (e.g., the Mann-Whitney U and Wilcoxon T non-parametric alternatives to the

[203] Campbell, 2004; Morrison, 2012.
[204] For more information on effect sizes see, for example, Campbell, 2004; Cohen, 1988; Ellis, 2010; and Kelley & Preacher, 2012.
[205] Busk & Serlin, 1992, cited in Morrison, 2012.
[206] Morrison, 2012.
[207] West & Hepworth, 1991, cited in Campbell, 2004.
[208] Siegel & Castellan, 1988, cited in Todman & Dugard, 2001.

independent samples and dependent samples *t*-tests, respectively), but they lack sensitivity to real treatment effects in studies with very small numbers of participants. However, in some cases, rather than using non-parametric tests, the researcher can conduct randomization tests, which (a) do not discard information in the data by reducing them to ranks, and (b) can provide valid statistical analysis of data from a wide range of single-case and small-*n* designs.[209]

While we cannot present a full discussion of randomization tests here, we want to point out that randomization tests can increase confidence in the causal inferences that are made in single-case and small-*n* studies.[210] Further, they are not particularly difficult to conduct, although they do require the researcher (or teacher or other professional) to follow specific procedures.

The basic principle of randomization tests is that some aspect of the experimental design must be randomized. However, randomization tests do not require students to be randomly assigned to specific treatments (or controls) or to certain classrooms; instead, the timing of the introduction of the intervention can be randomly selected. For an A-B single-case design, for example, the researcher randomly selects the point (e.g., day or week) at which to switch between the baseline and intervention phases. The same is true for a multiple baseline design (with some additional requirements). As with any type of analysis there are limitations to the interpretation of the data from randomization tests, but they can provide information about the statistical effects of an intervention. Additionally, randomization tests can be conducted using software such as Microsoft Excel or SPSS. Interested evaluators are urged to consult Kratochwill & Levin (2010) and Todman & Dugard (2001) for more information.

4.4.4 Qualitative Analysis

Unless the data collection instruments are completely structured, quantitative types of surveys, observations, and interviews, it is likely that significant amounts of qualitative data will be gathered during the course of the evaluation. It is beyond the scope of this Toolkit to go into detail about the many different types of qualitative inquiry and analysis approaches.[211] Instead, in this section we make some recommendations for preparing the qualitative data for analysis, briefly discuss the major phases of qualitative analysis, and highlight some benefits and drawbacks of using qualitative analysis software programs. Interested evaluators are urged to consult Miles and Huberman (1994) and Patton (2002) to learn more about how to conduct a variety of qualitative analyses.[212] Maxwell (2005) is a good source of information about how to design a qualitative study (e.g., as part of a mixed-methods evaluation).

[209] Todman & Dugard, 2001.
[210] Ibid.
[211] Patton, 2002, provides a nice discussion of a variety of orientations and theoretical approaches to qualitative inquiry and analysis. Denzin & Lincoln, 2005, is another valuable resource for those interested in learning more about qualitative inquiry.
[212] Two online resources are Schutt, 2011, and Suter, 2012.

4.4.4.1 Preparing Qualitative Data for Analysis

Just as with quantitative data analysis, the first step in qualitative analysis is preparing the data. We recommend evaluators follow these steps to prepare the data for analysis:

- **Enter the data into a spreadsheet or qualitative data analysis software program to facilitate analysis (e.g., to allow searches for particular words or phrases).** – This is particularly helpful when there are large numbers of study participants or when large quantities of qualitative data are expected. Some online survey providers (e.g., Survey Monkey) allow individuals to download a database file with the individual responses to each item, making the process of creating the database very easy.

- **Develop decision rules for how to handle any data problems (e.g., missing data, incomplete or misspelled answers).** – Here, the term "data problems" specifically refers to problems of missing or erroneous data, not data that might not seem to fit a researcher's preconceived ideas of what the data "should" look like. An example of when a decision rule may be necessary is when a person gives a response to Question 5 on a self-administered survey that actually looks like it should have been a response to Question 6 (while leaving Question 6 blank).

- **Conduct a preliminary check for errors, missing data or other problems with the data.** – When respondents are responsible for entering their answers to survey questions into a database—for example, in an online survey—and the study team plans to use search terms to facilitate data analysis, it is particularly important to check whether or not respondents have misspelled key terms (e.g., when a person enters "*congitive"* instead of "cognitive").

- **Apply the decision rules to address any errors.** – Whenever possible, go back to the original respondent for correction or clarification of any errors; if that is not possible the decision rules will provide guidance and will ensure consistency in the handling of data problems. It is important to note that when working with qualitative data, all responses—especially those that might be considered "outliers" in quantitative analysis—potentially can provide valuable insight into the subject of study. For this reason, the qualitative analyst should be cautious in the application of decision rules to deal with data problems. We recommend against deleting responses.

Once the data have been prepared, the major phases of analysis can begin (although in reality the initial phases of data analysis begin during data preparation, as the analyst reviews the questions and begins making sense of the responses). In the next sections we briefly discuss the major phases involved in qualitative analysis: data reduction, data display, and conclusion drawing and verification.[213] It is important to note that these are not discrete phases that happen in sequence. Rather, qualitative analysis is an iterative process that involves "concurrent flows of activity."[214]

---

[213] This framework comes from Miles & Huberman, 1994; see Berkowitz, 1997, for an overview of the approach.
[214] Miles & Huberman, 1994, p. 10.

Berkowitz (1997) highlighted some **important questions that the qualitative analyst should keep in mind during the data analysis process**:

- What patterns and common themes emerge in responses dealing with specific items? How do these patterns (or lack thereof) help to illuminate the broader study question(s)?
- Are there any deviations from these patterns? If yes, are there any factors that might explain these atypical responses?
- What interesting stories emerge from the responses? How can these stories help to illuminate the broader study question(s)?
- Do any of these patterns or findings suggest that additional data may need to be collected? Do any of the study questions need to be revised?
- Do the patterns that emerge corroborate the findings of any corresponding qualitative analyses that have been conducted? If not, what might explain these discrepancies? [215]

These questions are not specific to any one phase of data analysis and therefore should be addressed as appropriate throughout the study.

4.4.4.2 Data Reduction

"Data reduction refers to the process of selecting, focusing, simplifying, abstracting, and transforming the [qualitative] data."[216] In fact, data reduction occurs even before any data are collected, as the researcher makes decisions about the conceptual framework for the study, the research questions, data collection approaches, and data sources. As Freeman put it, "*Before I could access what it is I wanted to know, I had to first figure out what I meant by know, who I thought should do the knowing, and where I thought this knowledge could be found*."[217]

Data reduction generally relates to deciding which aspects of the data should be emphasized; this decision should be guided by the need to address the salient evaluation questions.[218] For example, if the evaluation is focused on the quality of graduates of a PDP for related-services providers, the study team may choose to focus the analysis only on the responses that pertain to students' perceptions of the quality of the related services they receive and not on students' in-class experiences with special education teachers.

Data reduction also involves transforming the data in a way that facilitates analysis and understanding. This may include selecting specific quotes to highlight, summarizing or paraphrasing long passages, or developing a narrative summary of a larger pattern of responses across multiple respondents or sites. In

[215] 1997, p.2.
[216] Miles & Huberman, 1994, p. 10.
[217] 2000, p. 360.
[218] Berkowitz, 1997.

some situations it might be helpful to identify "quantities" in the data, such as the number of PDP graduates who mentioned that they had a particularly challenging field experience. However, Miles and Huberman (1994) cautioned against simply using tables of numbers to illustrate the variety of qualitative responses to a question represented in the data.

For more in-depth discussion of data reduction, see Miles and Huberman (1994); Berkowitz (1997) provides a general overview of the approach.

4.4.4.3 Data Display

Data display is the next major phase of qualitative analysis and it involves arranging the data in an "organized, compressed assembly of information that permits conclusion drawing and action."[219] "A display can be a long piece of text, or a diagram, chart, or matrix that provides a new way of arranging and thinking about the more textually embedded data."[220] Data displays allow the analyst to begin identifying systematic patterns and interrelationships within cases (also known as intra-case analysis) and between cases (also known as inter-case analysis). [221]

Intra-case analysis examines patterns in the data across all of the sources within a case, such as scholars, graduates and faculty associated with one PDP institution or campus. Table 12 presents an example of a data display for one campus involved in a PDP project evaluation. Inter-case analysis examines patterns across different cases within one study, such as among respondents in three campuses involved in a PDP project evaluation. Table 13 presents a similar display that incorporates information from multiple campuses involved in the evaluation. Berkowitz (1997) pointed out that using data displays in this way allows for a relatively quick recognition of patterns in the data, even before the individual responses are completely analyzed. Additionally, it allows for the identification of differences in responses across the different groups of respondents.

[219] Miles & Huberman, 1994, p. 11.
[220] Berkowitz, 1997, p. 4.
[221] Berkowitz, 1997.

Table 13. Qualitative Data Matrix for Campus A Showing Responses for Scholars, Graduates & Faculty

What was done to improve scholars' skills and knowledge?			
Respondent Group	Activities Named	Which most effective?	Why?
Scholars	• Structured seminars • Field experience • In-class activities • Homework	• Structured seminars • Field experience	• Concise way of communicating information • Practical experience applying knowledge & developing skills
Graduates	• Field experience • Informal exchanges with other scholars	• Field experience • Informal exchanges with other scholars	• Practical experience applying knowledge & developing skills • Sharing practical advice on how to handle specific situations
PDP Faculty	• Structured seminars • Field experience • In-class activities • Homework • Online discussion boards	• Field experience • Homework	• Practical experience applying knowledge & developing skills • Allowed scholars to develop skills & knowledge in low-stakes environment

Source: Adapted from Berkowitz, 1997.

Table 14. Qualitative Data Matrix Comparing Responses at Three PDP Campuses

Participants' views of activities to improve scholars' skill and knowledge at three campuses			
Respondent Group	Activities Named	Which most effective?	Why?
Campus A	• Homework • Online discussion boards • Field experience • In-class activities	• Structured seminars • Field experience	• Concise way of communicating information • Practical experience applying knowledge & developing skills
Campus B	• Field experience • Structured seminars • Homework	• Field experience • Homework	• Practical experience applying knowledge & developing skills • Allowed scholars to develop skills & knowledge in low-stakes environment
Campus C	• Field experience • In-class activities • Structured seminars • Homework • Informal exchanges with other scholars	• Field experience • Informal exchanges with other scholars	• Practical experience applying knowledge & developing skills • Sharing practical advice on how to handle specific situations

Source: Adapted from Berkowitz, 1997.

As can be seen in the tables above, the field experience was consistently rated by all respondents as the most effective activity for improving scholars' skills and knowledge. In this example, all respondents gave the same reason for choosing this activity, but the reason for identifying an activity as the most effective differed across respondents. Data displays allow the qualitative analyst to easily identify these similarities or differences.[222] This is achieved by what Glaser and Strauss (1967) called the "'method of constant comparison,' an intellectually disciplined process of comparing and contrasting across instances to establish significant patterns, then further questioning and refinement of these patterns as part of an ongoing analytic process."[223]

Once the analyst identifies patterns of responses and similarities and differences among different respondent groups, the next step is to delve more deeply into the reasons behind the different responses, which is explored in the next phase of qualitative analysis: conclusion drawing and verification.

4.4.4.4 Conclusion Drawing and Verification

Conclusion drawing involves taking a step back to consider what the analyzed data mean and to assess how they relate to the evaluation questions under study. Verification entails going back to the data as often as necessary to check or validate emergent conclusions.[224] An essential element of this process is testing the conclusions for their plausibility, their sturdiness, their "confirmability"—in other words, their validity.[225] It is important to note that "validity" in this context is different from the conception of "validity" as part of quantitative analysis. Specifically, in the qualitative context, the "validity" of the analysis relates to "whether the conclusions being drawn from the data are credible, defensible, warranted, and able to withstand alternative explanations."[226] Miles and Huberman outlined **13 tactics for testing or confirming findings in qualitative analysis[227]**:

1. Checking for representativeness;
2. Checking for researcher effects;
3. Triangulating across data sources and methods;
4. Weighting the evidence;
5. Checking the meaning of outliers;
6. Using extreme cases;
7. Following up surprises;
8. Looking for negative evidence;
9. Making if-then tests;
10. Ruling out spurious relations;

[222] For detailed information on how to develop a variety of data displays, see Miles and Huberman (1994).
[223] Berkowitz, 1997, p. 8.
[224] Ibid.
[225] Miles & Huberman, 1994, p. 10.
[226] Berkowitz, 1997, p. 8.
[227] 1994, pp. 262-276.

11. Replicating a finding;
12. Checking out rival explanations; and
13. Getting feedback from informants.

It is beyond the scope of this Toolkit to discuss all of these tactics in detail. For more information about these tactics, and about standards for judging the quality of conclusions drawn from qualitative data, consult Miles and Huberman (1994).[228] Berkowitz (1997) provides a more general discussion of the process of conclusion drawing and verification.

In the next section we briefly highlight some of the more popular qualitative data analysis software programs and point out some of the advantages and disadvantages of using them.

4.4.4.5 Qualitative Data Analysis Software Programs

Depending on the resources available, the study team may want to invest in qualitative data analysis software, such as the commonly-used proprietary tools NVivo, ATLAS.ti, Ethnograph, HyperRESEARCH, QDA Miner, MAXQDA, Qualrus, Dedoose or one of the open-source programs such as Transana and Coding Analysis Toolkit (CAT; Suter, 2012). These tools can facilitate complex analysis of large quantities of qualitative data in less time than might be required to analyze the data by hand or through a spreadsheet, although significant time nevertheless may be required to set up or tailor the program to the evaluator's needs (See Schutt, 2011, for an example of how qualitative analysis might be conducted using some common software programs). Some of the **features of qualitative analysis software** include:

- Project management tools (e.g. that allow multiple analysts to work with data in different files following consistent analysis protocols established at the outset);
- Communication and memo writing tools (e.g., that allow researchers to incorporate comments and memos in the overall analysis);
- Collaboration tools (e.g., that allow researchers to share their analyses, comments, and memos; some even calculate inter-rater reliability);
- Graphing features (e.g., that allow a researcher to generate visual displays of findings to facilitate understanding of the connections among the different themes or data sources); and
- Query, search, and report functions (e.g., that can search for specific words and phrases and generate reports on their frequency).

[228] pp. 262-280.

As can be seen, these types of software can provide valuable benefits to the qualitative analyst. However, some researchers have raised concerns about the use of software because of issues related to:

- The time required to learn and master the software;
- An increased focus on coding as the primary analysis strategy;
- An increased focus on breadth and volume, rather than depth and meaning; and
- An increased focus on queries and text searches that may detract from the nuance and context in the data.[229]

Of course, no software program can take the place of good training in qualitative analysis, which requires significant creativity, intellectual discipline, and analytic rigor on the part of the analyst. The software can facilitate analysis, but a strong foundation in qualitative research techniques is required to ensure the quality of the conclusions reached by the study team.

[229] E.g., St. John & Johnson, 2000.

Appendix

Appendix A. Templates, Sample Forms

A.1. CIPP Logic Model Template

A logic model provides a starting point for developing an effective evaluation plan. CIPP's goal is to assist projects in developing logic models that are precise and that include features and content that give the models utility for evaluation purposes. The logic models serve as both ends and means. As ends, they are stand-alone representations of projects that provide an overall visual summary. They also provide descriptive information that can be used to catalogue or compare features across multiple projects, if needed. As means, the use of logic models is central to (1) defining outcomes that are meaningfully connected to project activities and (2) supporting evaluations so that the process will improve projects' overall performance.

CIPP's Logic Model Scheme

A project logic model portrays a project's overall plan. It serves to clarify the relationships among a project's goals, activities, and outputs and to lay out the connections between them and the project's expected outcomes. Therefore, a project logic model depicts a program theory and accompanying hypotheses, highlighting (1) the resources or inputs dedicated to an effort, (2) the planned activities to be carried out with those resources, and (3) the specific outputs and outcomes the activities will generate. Evaluation can then be viewed as a test of the logic model's hypotheses, and a logic model can be used by evaluators and the grantee to refine and guide data collection and analysis for assessing both process and performance.

We find it helpful to begin with a summary chart that contains the information that will populate the logic model. The chart outlines the OSEP priority, assumptions, external factors/context, and inputs. It then displays, in table format from left to right, the project's goals and objectives, strategies/activities, outputs, and outcomes. From the chart, a logic model is prepared. The logic model is less comprehensive than the chart in its content but it uses lines and arrows to connect specific project elements and provides a dynamic display. Both the chart and the logic model are continuously updated as the content of specific elements changes, such as when planned activities are revised or when unintended outcomes occur. The logic model will also change as the relationships among the components develop over time, mostly likely by becoming more complex and interactive. For CIPP, we use the following definitions of the logic model components:

Goals/Objectives – The *goals* capture the overarching purposes of the project. Goals make clear the anticipated impact on systems or individuals. Goals imply gaps or deficits that will be remedied when the project produces its long-term outcomes. *Objectives*, if used in a logic model, are targeted sub-goals.

Strategies/Activities – *Strategies* are the broad approaches to addressing the goals. They include multiple activities. *Activities*, which may or may not be listed in the logic model, are the specific actions funded by the grant or supported by other resources under the umbrella of the project.

Outputs – *Outputs* are the direct results of the project activities, including project products and programs. Most outputs will be quantifiable, including tallies of the number of products and programs or counts of the customer contacts with those products and programs.

Direct/Intermediate Outcomes – *Direct outcomes* are what customers do or become as a result of outputs. Usually, direct outcomes are changes in the customers' actions or behaviors based on knowledge or skills acquired through project outputs. *Intermediate outcomes* result either directly from outputs or indirectly through direct outcomes. They generally come later in time than direct outcomes and often represent a step between direct outcomes and long-term outcomes.

Long-term Outcomes – *Long-term outcomes* are the broadest project outcomes and follow logically from the direct and intermediate outcomes. They are the results that fulfill the project's goals. Outputs, direct outcomes, and intermediate outcomes all contribute to the achievement of the long-term outcomes. Although the long-term outcomes represent fulfillment of the purpose of the project, they may or may not represent the achievement of a desired larger project impact. That is, the project may have an anticipated impact that is beyond the immediate scope of the project, either temporally or conceptually, and thus beyond the scope of the logic model.

The Summary Chart

The summary chart contains all the information that will populate the logic model, plus additional details about strategies, activities, and outputs. The chart begins with the OSEP priority and then states assumptions about how and why the project will be successful. External factors and context provide a brief description of the environment in which the project will be operating. Inputs are specific resources available to the project. The table itself displays, from left to right, the project's goals and objectives, strategies/activities, outputs, and outcomes. Note that strategies/activities are aligned with outputs but that goals and outcomes may cut across multiple strategies/activities.

The Logic Model

The content of the logic model is taken entirely from the summary chart, but the logic model content is condensed to fit into the flow chart format. In the logic model, lines and arrows are used to depict the temporal and causal connections among the various project elements. Not surprisingly, multiple lines or arrows come to or from most of the boxes, indicating the complexity of the relationships that are expected. Also depicted are the anticipated results in the form of direct, intermediate, and long-term outcomes. The outcomes are themselves interconnected. Thus, direct outcomes, as well as outputs, lead to the higher level, more distal outcomes.

Summary Chart

OSEP Priority: OSEP's stated purpose for the [project] is to . . . .

Assumptions: The [project] is managed by [institution], which has a long history of . . . . This reputation along with the renewed emphasis by OSEP on . . . will create interest in . . . . Specifically, stakeholders focused on . . . will find the [project's] products and services essential in . . . .

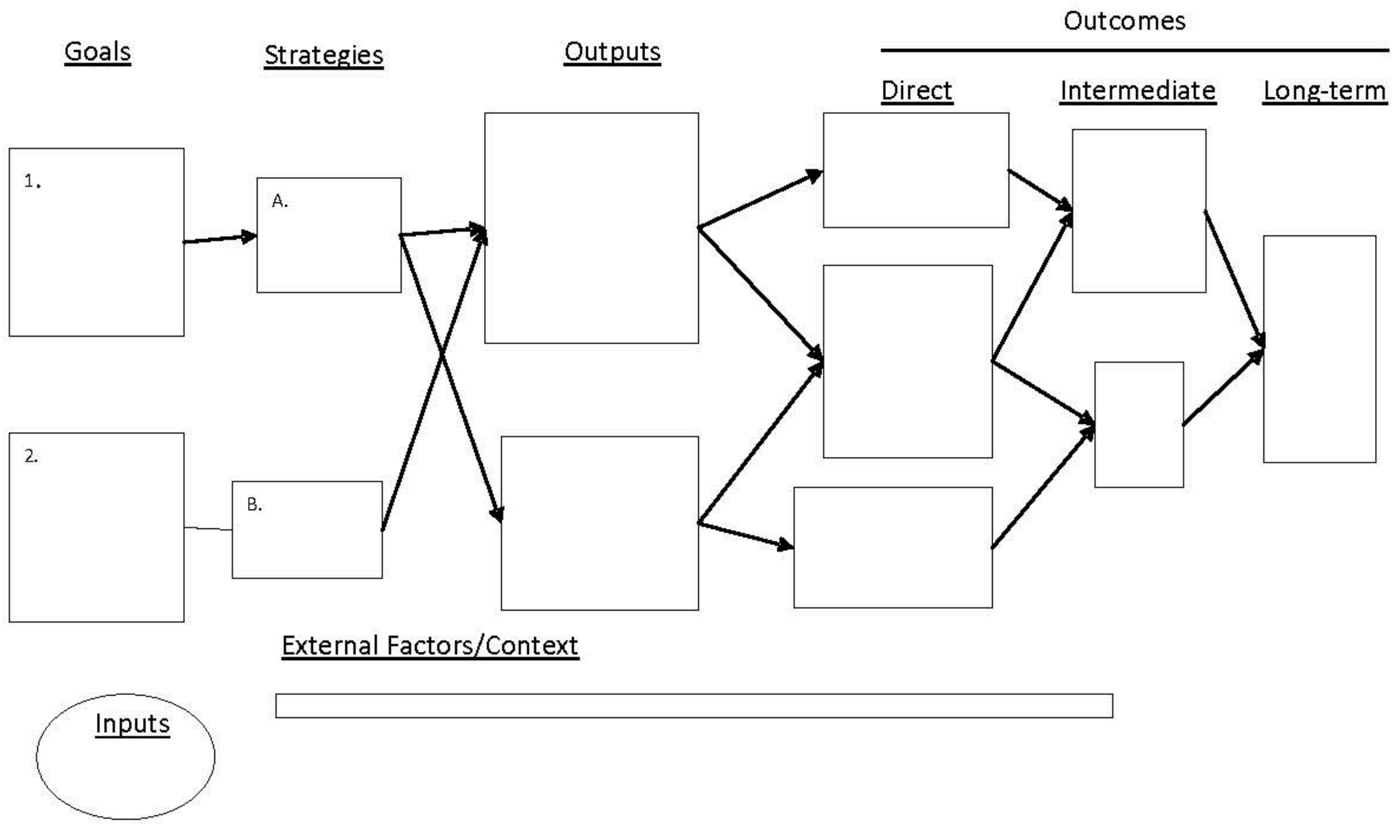
External Factors/Context: Other federal initiatives; OSEP policy environment; [institution's] accumulated experience and visibility….

Inputs: OSEP funding, experienced project staff, lessons learned from past experience, research-based policy and practices, …

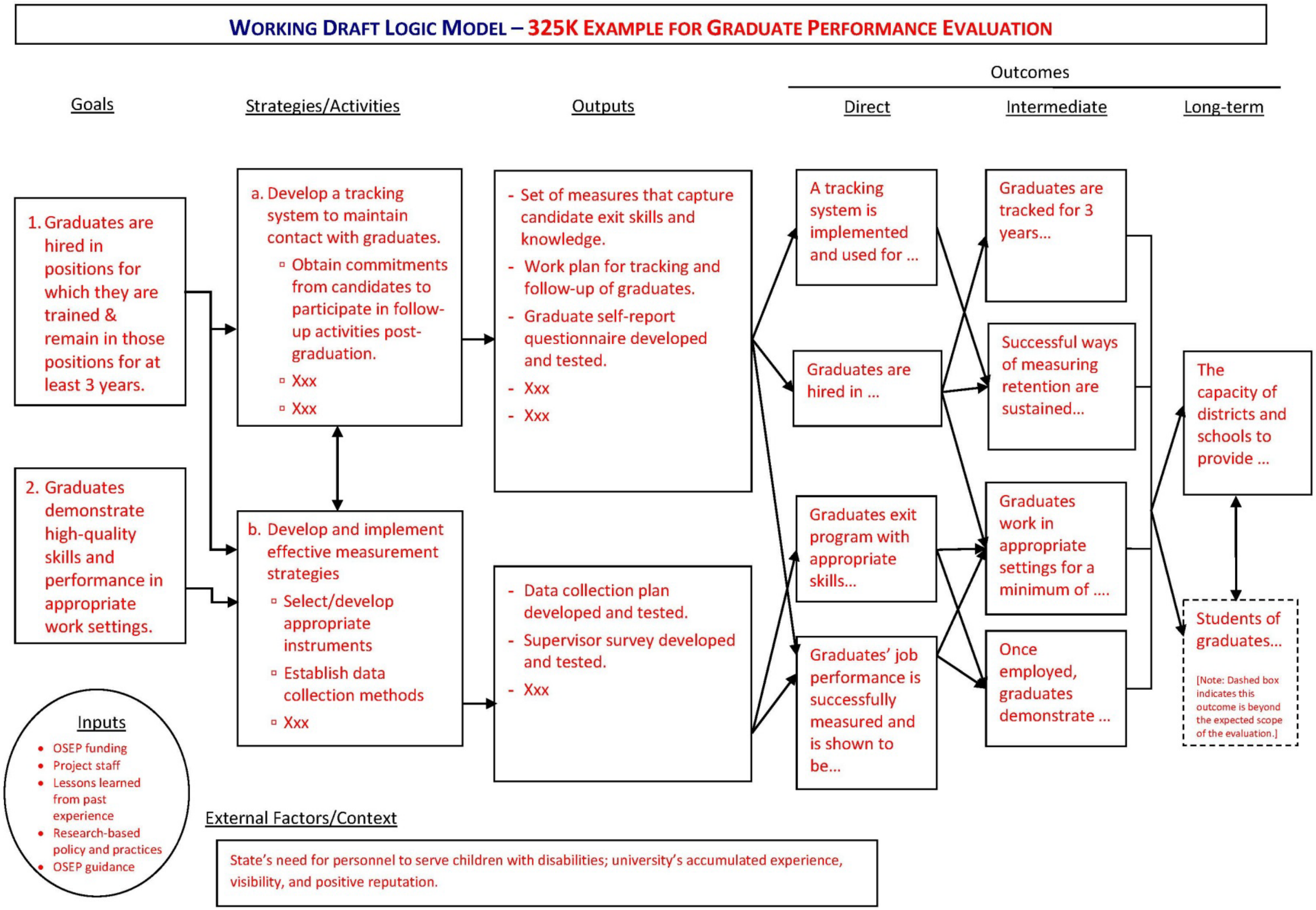
Table A.1.1. Template: Summary Chart Depicting Goals/Objectives, Strategies and Activities, Outputs and Outcomes

<table>
<tr><th>Goals/Objectives</th><th>Strategies and Activities</th><th>Outputs</th><th>Outcomes</th></tr>
<tr><td rowspan="3">1.
2.
3.
4.</td><td>A.
(1)
(2)
(3)</td><td>•
•
•</td><td rowspan="3">Direct:
•
•
•
•
Intermediate:
•
•
Long-term:
•
•</td></tr>
<tr><td>B.
(4)
(5)
(6)
(7)</td><td>•
•
•
•
•</td></tr>
<tr><td>C.
(8)
(9)</td><td>•
•
•</td></tr>
</table>

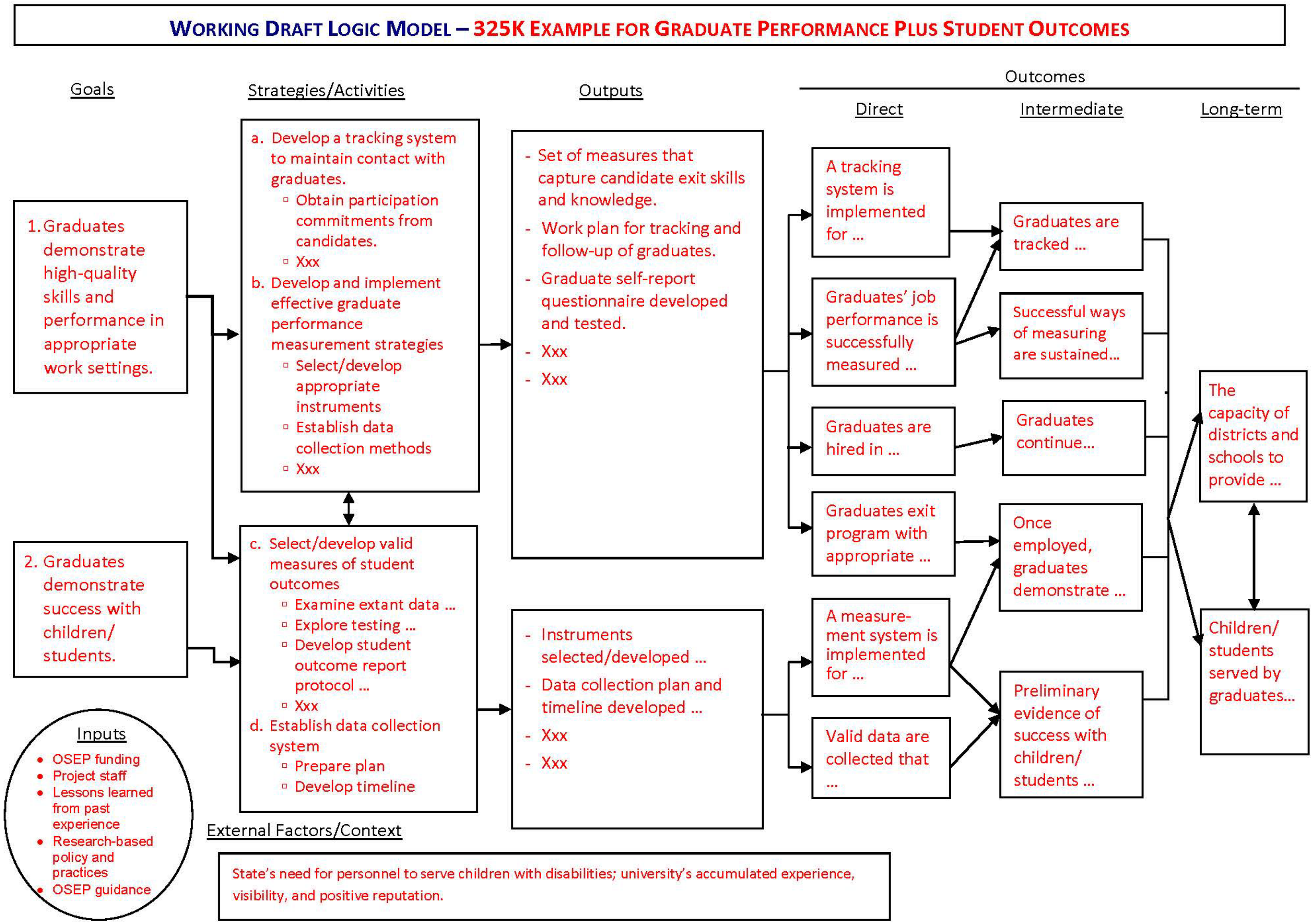
LOGIC MODEL TEMPLATE



A.2. Sample Logic Model Illustrating a Project Focused Primarily on Graduate Performance Outcomes



A.3. Sample Logic Model Illustrating a Project Focused on Graduate Performance and Student Outcomes



Westat

Resource Toolkit

A.4. CIPP Summative Evaluation Plan Template

Note: *Instructions/Guidelines for completing the template are in italics.*

Introductory paragraph—one paragraph overview of the project.

Need, Goals, and Activities

Paragraph stating the need that the selected project was established to address. Can be followed by a textbox that succinctly states that need.

NEED:

Outline showing the goals, strategies, and activities.

GOAL #1:

STRATEGY:

ACTIVITY:

ACTIVITY:

STRATEGY:

ACTIVITY:

GOAL #2:

STRATEGY:

ACTIVITY:

STRATEGY:

ACTIVITY:

ACTIVITY:

GOAL #3:

STRATEGY:

[Project] Program Theory

Paragraph providing a general intro to the program/project theory of change.

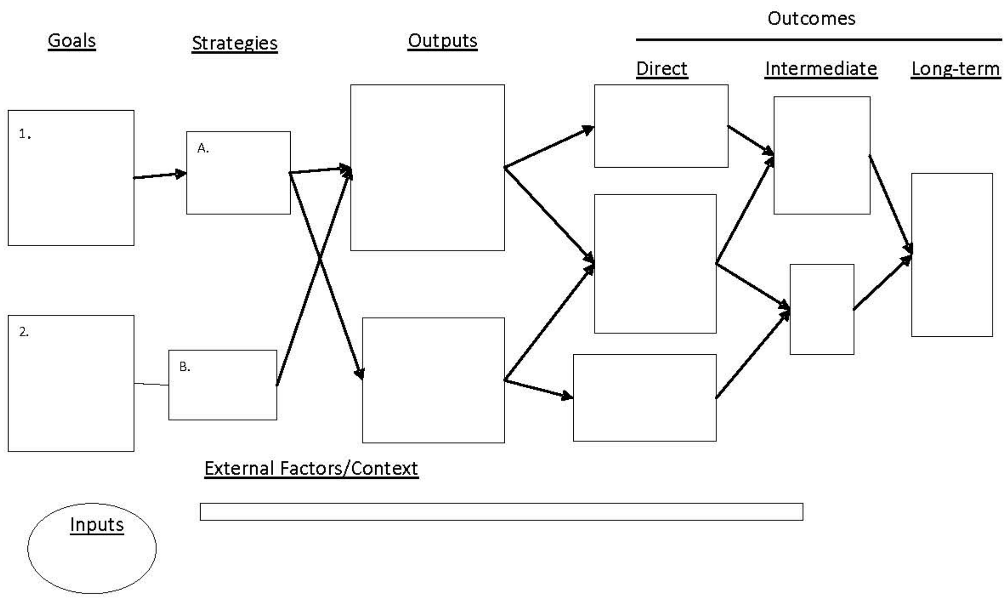
Paragraph introducing Table A.4.1, which presents the alignment of project goals with strategies/ activities, outputs, and outcomes. The table should also include the top-level summative evaluation question associated with each goal.

Table A.4.1. Summative evaluation questions aligned with [project] goals, strategies/activities, outputs, and outcomes. *[Mostly comes from the Summary Chart prepared for the project logic model, with the addition of the highest level summative evaluation questions.]*

Goals	Strategies/Activities	Outputs	Summative Outcomes	Summative Evaluation Questions
1.	▪ ▫ ▫	- - -	• •	A.
2.	▪ ▫	- - -	• •	B.

Paragraph introducing the logic model as a graphical representation of the content of Table A.4.1.

Figure A.4.1. [Project] logic model. *Insert completed logic model here.*



Summative Evaluation Design

Restatement of top-level evaluation questions from Table A.4.1, by outcome level.

Evaluation Questions Related to Direct Outcomes

A.

B.

C.

Evaluation Questions Related to Intermediate Outcomes

B.

C.

Evaluation Questions Related to Long-Term Outcomes

D.

Paragraphs presenting an overview of the evaluation approach that will be applied to addressing each question.

Data Collection

Paragraphs presenting an overview of data collection plans, starting with data collection related to the project's formative evaluation.

Introduction to Table A.4.2. Table A.4.2 presents the summative evaluation questions in relation to outcomes, data collection activities, and the data collection instruments that are specific to types of respondents.

Discuss status of instruments—already in use, under development, planned by the project, needing to be developed or modified.

Discuss need for baseline, whether baseline data already exist, and how baseline data collection might work.

Table A.4.2. Summative evaluation questions in relation to outcomes, data collection activities, and instruments.

Summative Evaluation Question	Outcome	Data Collection Instrument	Type of Data Collection
A.	•		
	•		
B.	•		
	•		

Sampling

Discuss need for sampling, if any. Discuss the likelihood of a need for a sampling plan and some of the specifics if known. Include a sampling plan, if available.

Analysis Approach

Descriptive Analyses

Discuss need for descriptive analyses—that is, where they will be used. Present specific techniques corresponding to the specific summative evaluation questions and data collections. Identify specific "studies" if they are sufficiently discrete.

Statistical Analyses

Discuss need for statistical analyses—that is, where they will be used. Present specific techniques corresponding to the specific summative evaluation questions and data collections. Identify specific "studies" if they are sufficiently discrete.

Point out where pre-post data will be used and why only posttest data may be available in some cases. Name the statistical tests, if known. Discuss limitations, or at least the major limitations.

Data Sources

Include an appendix with a table of instruments and the actual instruments, if available.

Data Sources for Descriptive Analyses

List specific instruments or data bases to be used to collect data or as sources of data.

Data Sources for Statistical Analyses

List specific instruments or data bases to be used to collect data or as sources of data.

Data Collection Schedule

Ongoing Assessments

Describe the timeline for any assessments or data collections that are ongoing and relevant to the summative evaluation. These will likely be the data collections for the formative evaluation or some type of data base that is regularly updated.

One-time Data Collections

Describe the timeline for any assessments or data collections that are one time and relevant to the summative evaluation.

Pre-Post Data Collections

Describe the timeline for pre-post—this may be experimental, quasi-experimental, time series or some other rigorous design, or it may be just a repeat of data collections—assessments or data collections that are relevant to the summative evaluation.

Prepare a simple table like Table A.4.3 or a Gantt chart, as illustrated in *Section 2.2.5.6**. It also may be helpful to prepare a table such as Table A.4.4, which provides information on the status of data collection instruments and activities.*

Table A.4.3. Summative evaluation data collection schedule.

Evaluation Activity	First Data Collection	Additional Data Collections

Table A.4.4. [Project] Instrument Information Table

Evaluation Question	Data Source	Possible instrument/ protocol	What is the status of the instrument/ protocol? E=exists UD=under development TBD= to be developed	Have any data been collected with this instrument (if it exists)? If so, when?	If data were collected, was sampling used?	When are future data collection(s) planned for this instrument/ protocol?	Will data collection require sampling?
A.							
B.							

A.5. Sample Evaluation Plan: Graduate Performance and Student Outcomes

Note: *Instructions/Guidelines for completing a section are in italics.* Examples are provided in red.

Need, Goals, Activities, Outcomes, and Evaluation Questions

Instructions: Prepare a sentence or a brief paragraph stating the need that the project was established to address. The need is related to the project purpose but is stated in terms of the deficit that the project will fill.

Example: Schools and other service providers have a need for fully qualified speech therapists who can work with bilingual secondary students with disabilities. These therapists must be proficient in evidence-based practices. The need for therapists from diverse racial/ethnic backgrounds and therapists with disabilities is especially strong.

Fill in Table A.5.1, using however many rows you require. Table A.5.1 aligns project goals with strategies/activities, outputs, and outcomes. The table should also include the top-level summative evaluation questions associated with each goal. Definitions of the elements of the table are provided below, while examples of each for one goal are provided within the table. Please note that the examples provided are not intended to be comprehensive or cohesive.

Goals/Objectives – The goals capture the overarching purposes of the project. Goals make clear the anticipated impact on systems or individuals. Goals imply gaps or deficits that will be remedied when the project produces its long-term outcomes. Objectives, if used, are targeted sub-goals.

Strategies/Activities – Strategies are the broad approaches to addressing the goals. They include multiple activities. Activities are the specific actions funded by the grant or supported by other resources under the umbrella of the project. Although listing activities under broader strategies is preferred, activities alone may be listed.

Outputs – Outputs are the direct results of the project activities, including project products and programs. Most outputs will be quantifiable. They include tallies of the number of products and programs or counts of the customer contacts with those products and programs.

Direct/Intermediate Outcomes – Direct outcomes are what customers do or become as a result of outputs. Usually, direct outcomes are changes in the customers' actions or behaviors based on knowledge or skills acquired through project outputs. Intermediate outcomes, which result either directly from activities or indirectly through direct outcomes, often represent a step between direct outcomes and long-term outcomes. Often, intermediate outcomes are changes in the knowledge, skills, or behavior of persons touched by the project's direct customers. All important outputs will have one or more direct or intermediate outcomes.

Long-term Outcomes – Long-term outcomes are the broadest project outcomes and follow logically from the direct and intermediate outcomes. They are the results that fulfill the project's goals. Outputs, direct outcomes, and intermediate outcomes all contribute to the achievement of the long-term outcomes. Although the long-term outcomes represent fulfillment of the purpose of the project, they may or may not represent the achievement of the desired project impact. That is, the project may have an anticipated impact that is beyond the immediate scope of the project, either temporally or conceptually, and thus beyond the scope of the evaluation plan. Furthermore, long-term outcomes may be the target of more than one goal.

Evaluation Questions – Evaluation questions frame the way data will be summarized or analyzed to address the overarching issue of whether the project's goals have been addressed. The data may be collected specifically for evaluation, collected for other purposes (such as research) but useful for evaluation, or extant. In effect, evaluation questions connect evaluation data to the goals. Evaluation questions here should be top-level questions only.

Table A.5.1. Example: Evaluation questions aligned with [project name] goals, strategies/activities, outputs, and outcomes.

Goals	Strategies/Activities	Outputs	Outcomes	Evaluation Questions
Example				
1. Graduates demonstrate high-quality skills and performance in appropriate work settings.	▪ Develop a tracking system to maintain contact with graduates. ▫ Develop a reliable and efficient system that will allow graduates to be followed for a minimum of 3 years post-completion. ▫ Obtain commitments from candidates to participate in follow-up activities post-graduation. ▪ Develop a plan for systematically obtaining data on the performance of graduates and their students. ▫ Assemble a set of measures for determining skills and knowledge of candidates at time of program completion ▫ Work in conjunction with districts and agencies where graduates are employed (or likely to be employed) to develop valid and practical measures for determining graduate performance. ▫ Work in conjunction with districts and agencies to determine the most valid and efficient means for collecting data on the performance of children/students that graduates serve. ▪ Implement the tracking system and integrate with other data collection plans	- Set of measures that capture candidate exit skills and knowledge assembled. - Work plan for tracking and follow-up of graduates. - Work plan for the collection of valid data on the performance of graduates and of their students.	Direct • Graduates exit the program having demonstrated the skills and knowledge to perform at a high level. • Graduates are tracked for 3 years. • Data on the performance of graduates are collected for 3 years. Intermediate • Graduates are working in appropriate settings for a minimum of 3 years. • Graduates are providing high quality services.	A. Are we collecting valid data on the performance of our graduates? B. Do the expected number and percentage of graduates work in appropriate settings for 3 years? C. To what extent do graduates exit the program with the skills and knowledge necessary to perform at a high level? D. To what extent do graduates demonstrate in the workplace the high quality skills and knowledge needed to improve outcomes for children with disabilities?

Goals	Strategies/Activities	Outputs	Outcomes	Evaluation Questions
2. Graduates demonstrate success with children/ students.	▪ Select/develop valid measures of student outcomes ▫ Examine extant data availability. ▫ Explore testing options. ▫ Develop graduate reporting protocol. ▪ Establish data collection system ▫ Prepare plan ▫ Develop timeline ▪ Develop analysis plan	- Instruments selected or developed and tested. - Data collection plan and timeline. - Analysis plan.	<u>Direct</u> • A measurement system is implemented. • Valid data are collected. <u>Intermediate</u> • Evidence of success with children/students served by graduates. <u>Long-term</u> • Students of graduates demonstrate improved outcomes.	E. To what extent do graduates demonstrate success with child/students with disabilities?

Data Sources and Collection by Evaluation Question

Fill in Table A.5.2, using however many rows you require. Table A.5.2 presents the summative evaluation questions in relation to outcomes, data sources, and data collection strategies. The first row provides an example, using Evaluation Question c from the Goal I example in Table A.5.1. You will provide more detail about data and data collections as indicated in the sections that follow the table.

Table A.5.2. Example: Evaluation questions in relation to outcomes, instruments, and data collections.

Summative Evaluation Questions	Outcomes	Instruments or Datasets	Modes of Data Collection	Comparison Data Collection Planned (✓ if yes)	Timeframe for Initial Data Collection	Timeframe for Recurrence of Data Collection, If Any
Example						
C. To what extent do graduates exit the program with the skills and knowledge necessary to perform at a high level?	Direct • Graduates exit the program having demonstrated the skills and knowledge to perform at a high level.	▪ Comprehensive examines ▪ Observation protocols for field placement supervisors ▫ Short form ▫ Extended observation form	- Tests - Observation protocol	☐	Final semester of training	None
E. To what extent do graduates demonstrate success with children/students with disabilities?	Direct • A measurement system is implemented. • Valid data are collected. Intermediate • Evidence of success with children/students served by graduates.	▪ State testing program ▪ Instruments for measuring non-academic student outcomes. ▪ Protocols for extraction of extant data on student outcomes. ▪ Dataset of graduate performance data and corresponding student outcome data.	- Testing. - Review of extant data.	✓	18 months after start of grant	Annually for 3 years

Special Data Collection Considerations

Instrument or Dataset Status

Instructions: List and discuss status of each instrument—ready to use, needing modification, under development, planned by the project. Discuss availability of each data set planed for use—access to the data set and timeliness of the data availability.

Example: The project will need to develop a tracking system capable of keeping up with contact information and employment status of each graduate. The system will need to be easily updatable and will include logs of contacts and attempted contacts with the graduates as well as alternative contacts who may be of assistance in locating a graduate.

Comparison Data

Instructions: Discuss need for baseline and, if needed, whether baseline data already exist or how baseline data collection might work. Identify possibilities for experimental or quasi-experimental designs—discuss planned use of comparison groups and whether random assignment will be used. Describe any available extant datasets that could be used in addition to or instead of collected data.

Example: To determine the retention rates for our graduates (Evaluation Question B) in comparison to graduates of other programs, the project has made an arrangement with Districts X and Y, where 60 percent of our graduates are employed, to use administrative records of employment of speech and language therapists. To protect personnel confidentiality, the data will be provided to us in masked form in two datasets. The first dataset will be the employment status of our graduates. The second dataset will be the corresponding statistics for all other recently employed speech and language therapists. We will use the second dataset as our comparison data.

To address Evaluation Question E, the project has also made an arrangement with Districts X and Y to use administrative records of progress on a relevant speech/language assessment. To protect child confidentiality, the data will be provided to us in aggregated or masked form in two datasets. The first dataset will be the goal attainment scores for the students our graduates serve. The second dataset will be the corresponding statistics for all other special education students receiving speech and language services. We will use the second dataset as our comparison data. When possible without risking the exposure of individual students' identities, the data will include individual student descriptors, such as sex, SES, and disability category.

Sampling

Instructions: Discuss need for or consideration of sampling, if any, and present some of the specifics if known.

Example: To minimize costs to the project and burden on field supervisors, in addressing Evaluation Question C we will employ sampling. The project will ask field placement supervisors to follow an extended observation protocol for a representative sample of our enrolled candidates during their final semester of training. These data will be collected weekly. The random sample will be stratified to ensure that it is representative of key characteristics of our candidates: sex, race/ethnicity, disability status.

Sampling will also be employed when the performance of children served by graduates is measures for Evaluation Question E. The project will ask currently enrolled candidates to manage a data collection of social language (pragmatics) for the children being served by graduates. We will use the Language Use Inventory (LUI) to collect data twice a year, with permission of the parents or guardians, who will be the respondents. The random sample will be stratified to ensure that it is representative of the demographics of the children our graduates served—that is, it will be stratified by grade, sex, and race/ethnicity to ensure that a sufficient number of students with each of these characteristics are represented to allow for generalization for these characteristics. The LUI is a standardized instrument with well-known characteristics, which will allow an accurate estimation of needed sample sizes by our statistician.

Analysis Approach

Descriptive Analyses

Instructions: Discuss need for descriptive analyses—that is, for which evaluation questions will descriptive analyses be used. Present specific techniques corresponding to the evaluation questions and data collections.

Example: Analysis will generate descriptive statistics related to Evaluation Question D on various measures of graduate performance. For example, analysis of observation data will identify the percentage of graduates using specific evidence-based practices at or above a specified threshold level. Additionally, cross tabulations will be used to identify factors associated with the strength of the results. For example, cross-tabulated data may show that a higher percentage of graduates whose field placement was in Setting 1 were using evidence-based practices compared to those whose field placement was in Setting 2. Other predictors, such as hours in field placement, the university supervisor who oversaw the field placement, and the point of time in the course of study when the field placement occurred, will also be examined.

Statistical Analyses

Instructions: Discuss need for statistical analyses—for which evaluation questions. Present specific techniques being considered, if these are known, corresponding to the evaluation questions and data collections. Identify specific "studies" if they are sufficiently discrete.

Example: The project will use statistical analysis to partly answer Evaluation Question E. For graduates working in Setting 3, a two-group comparison will be possible using pre and post scores from the administrative records of relevant speech/language assessments. One group will be children project graduates served; the other group will be comparable children served by graduates from other programs. Scores across different assessments will be standardized as effect sizes and will be analyzed as change scores for individual children. We will use a one-sided t-test to compare the two groups. Children will be matched on pretest scores, race, SES, testing accommodations, and disability category.

A.6.1 Sample Notification Letter for Districts with Research Approval Office/Department

Dear <Sal> <SupFName> <SupLName>:

I am writing to inform you that my organization, <EvaluatorName> is planning to conduct site visits and interviews with special educators and administrators at <SchoolName(s)> in your district. We have already completed the research application required by your district and received approval to conduct this study. We have been contracted by <GranteeName> to conduct an evaluation as part of a federal requirement to evaluate the performance of personnel preparation programs that receive funding from the Personnel Development Program in the U.S. Department of Education's Office of Special Education Programs. This evaluation involves gathering data to assess the performance of special education teachers and related-services providers who graduated from <GranteeName> in recent years, based in part on measures of educator practice and student achievement.

We plan to conduct <NumberObservations> of site visits to the schools where these <GranteeName> graduates are currently working. These site visits will consist of:

- an interview with the school principal or his/her designee,
- individual or group interviews with teachers or service providers, and
- observations of educator practice.

The interviews and observations should each last between <InterviewLength> and <ObservationLength>. There is no need to prepare or provide any documentation.

The data we collect during the visits will not be used to evaluate the schools' or your district's performance. It will be aggregated with data collected on <GranteeName> graduates working in other schools and districts as part of a report on the performance of <GranteeName>program as a whole. All data collected for this study will be kept confidential, except as required by law. A report will be delivered to <GranteeName> with results aggregated across all respondents.

Thank you in advance for your district's cooperation and participation in this important study. Feel free to contact me directly with questions or issues. I can be reached by calling <EvaluationDirectorTelephone> or by emailing <EvaluationDirectorEmail>.

Sincerely,

<EvaluationDirector>

<EvaluationDirectorContactInformation>

A.6.2 Sample Request Letter for Districts without Research Approval Office/Department

Dear <Sal> <SupFName> <SupLName>:

We are requesting permission to conduct site visits at schools in your district where graduates of <GranteeName> are currently working. My organization, <EvaluatorName>, is conducting a study under contract with <GranteeName> to evaluate the performance of graduates of their personnel preparation program. This evaluation is part of a federal requirement to monitor the performance of personnel preparation programs that receive funding from the Personnel Development Program in the U.S. Department of Education's Office of Special Education Programs. This evaluation involves gathering data to assess the performance of special education teachers and related-services providers who graduated from <GranteeName> in recent years. **Your district's participation is important for <GranteeName> to ensure that its program maintains high standards so that its graduates continue providing high quality instructional services to children like those in your district.**

These site visits will consist of:

- an interview with the school principal or his/her designee,
- individual or group interviews with teachers or service providers, and
- observations of educator practice.

We plan to conduct <NumberInterviews> and <NumberObservations> over the course of <StudyLength>. The interviews and observations should each last between <InterviewLength> and <ObservationLength>. There is no need to prepare or provide any documentation.

The data we collect during the visits will not be used to evaluate the schools' or your district's performance. It will be aggregated with data collected on <GranteeName> graduates working in other schools and districts as part of a report on the performance of <GranteeName> program as a whole. All data collected for this study will be kept confidential, except as required by law. A report will be delivered to <GranteeName> with results aggregated across all respondents.

Thank you in advance for considering participating in this important study. Please indicate the decision of your district on the enclosed form. Feel free to contact me directly with questions or issues. I can be reached by calling <EvaluationDirectorTelephone> or by emailing <EvaluationDirectorEmail>.

Sincerely,

<EvaluationDirector>

<EvaluationDirectorContactInformation>

A.6.3 Sample District Response Form

I give district permission for the evaluation study conducted by <EvaluatorName> to take place during the current school year. This study is designed to evaluate the graduates of <GranteeName> and will include site visits to participating schools.

____
Name

____
Date

____
Position

____
District

____
Phone Number

A.7. Sample School Notification Letter

Dear <Sal> <PrinFName> <PrinLName>:

I am writing to ask your permission for my organization, <EvaluatorName>, to visit your school and interview you and selected staff in <VisitTimeframe>. We have been contracted by <GranteeName> to conduct a study as part of a federal requirement to evaluate the performance of personnel preparation programs that receive funding from the Personnel Development Program in the U.S. Department of Education's Office of Special Education Programs. **We realize that your time is valuable and we will do our best to make your participation in this study as easy as possible.** Your help is very important to our efforts to improve the quality of personnel preparation programs, and subsequent outcomes for students. This study involves gathering data to assess the performance of special education teachers and related-services providers who graduated from <GranteeName> in recent years, based in part on measures of educator practice and student achievement. We have already received district approval to conduct this study.

Our plan is to conduct <NumberObservations> of site visits to your school. These site visits will consist of:

- a brief interview with you or your designee about the performance of each graduate of <GranteeName> working at your school,
- individual or group interviews with the graduates at your school, and
- observations of the graduates' practice.

The interviews and observations should each last between <InterviewLength> and <ObservationLength>. There is no need to prepare or provide any documentation prior to our visit.

The data we collect during the visits will not be used to evaluate your or your school's performance. It will be aggregated with data collected on <GranteeName> graduates working in other schools as part of a report on the performance of <GranteeName> program as a whole. All data collected for this study will be kept confidential, except as required by law. A report will be delivered to <GranteeName> with results aggregated across all respondents.

We understand that making staff and activities available to the site visitor will require time and effort from you and your staff. We appreciate your help with our efforts to improve training programs and outcomes for students. We will be in contact with you in the near future to arrange the details, date, and schedule of the visit to your school.

We appreciate your willingness to cooperate and provide information to help <GranteeName> improve the quality of its personnel preparation program. Feel free to contact me directly with questions or issues. I can be reached by calling < EvaluationDirector Telephone> or by emailing < EvaluationDirector Email>.

Sincerely,

<EvaluationDirector>

< EvaluationDirector ContactInformation>

A.8. Sample Passive Consent Form for Graduates

Dear <Sal> <GradFName> <GradLName>:

I am writing to tell you about an important study of the <GranteeName> personnel preparation program that you graduated from in <GraduationYear>. This study is part of a federal requirement to evaluate the performance of personnel preparation programs that receive funding from the Personnel Development Program in the U.S. Department of Education's Office of Special Education Programs. **Your participation in this study is extremely important** to our efforts to improve the quality of personnel preparation programs, and subsequent outcomes for students.

As part of this evaluation, my organization, <EvaluatorName>, would like to visit your school to observe your practice and interview you and your personnel supervisor. We have already completed the research application required by your district and received approval to conduct this study.

We expect to conduct the interviews and observations in <VisitTimeframe>. The interview and observations should each last between <InterviewLength> and <ObservationLength>. There is no need to prepare or provide any documentation prior to our visit.

The data we collect during the visits will be kept confidential, except as required by law, and will be aggregated with data collected on <GranteeName> graduates working in other schools to develop a report on the performance of the <GranteeName>program as a whole. A report will be delivered to <GranteeName> with results aggregated across all respondents—no identifying information about you or your school will be included in the report.

We understand that making yourself available to participate in this study will require time and effort and we greatly appreciate your willingness to cooperate. We appreciate your help with our efforts to improve training programs and outcomes for students. We will be in contact with you in the near future to arrange the details, date, and schedule of the visit.

We hope you will be willing to cooperate and provide information to help <GranteeName> improve the quality of its personnel preparation program. **If you do not wish to participate, please notify me** at < EvaluationDirector Telephone> or by emailing < EvaluationDirector Email>.

Sincerely,

<EvaluationDirector>

< EvaluationDirector ContactInformation>

A.9. Sample Active Consent form for Graduates

Dear <Sal> <GradFName> <GradLName>:

I am writing to ask you to participate in study of the <GranteeName> personnel preparation program that you graduated from in <GraduationYear>. This study is part of a federal requirement to evaluate the performance of personnel preparation programs that receive funding from the Personnel Development Program in the U.S. Department of Education's Office of Special Education Programs. **Your participation in this study is extremely important** to our efforts to improve the quality of personnel preparation programs, and subsequent outcomes for students.

As part of this evaluation, my organization, <EvaluatorName>, would like to visit your school to observe your practice and interview you and your personnel supervisor. Additionally, we ask that you complete a brief survey related to your perceptions of the quality of the training and support you received from <GranteeName> and your perceived self-efficacy as a <Profession>. We have already completed the research application required by your district and received approval to conduct this study.

We expect to conduct the interviews and observations in <VisitTimeframe>. The interview and observations should each last between <InterviewLength> and <ObservationLength>. There is no need to prepare or provide any documentation prior to our visit.

The data we collect during the visits will be kept confidential, except as required by law, and will be aggregated with data collected on <GranteeName> graduates working in other schools to develop a report on the performance of the <GranteeName>program as a whole. A report will be delivered to <GranteeName> with results aggregated across all respondents—no identifying information about you or your school will be included in the report.

We understand that making yourself available to participate in this study will require time and effort and we greatly appreciate your willingness to cooperate. We appreciate your help with our efforts to improve training programs and outcomes for students. **If you agree, please complete the form below** and return it to us. We will be in contact with you in the near future to arrange the details, date, and schedule of the visit and to provide information on how to access the survey. We appreciate your willingness to cooperate and provide information to help <GranteeName> improve the quality of its personnel preparation program. Feel free to contact me directly with questions or issues. I can be reached by calling < EvaluationDirector Telephone> or by emailing < EvaluationDirector Email>.

Sincerely,

<EvaluationDirector>

Study Permission Form

By returning this form, I ***agree*** to participate in this study.

(Please Print) My name is:_______________________________School:________________________________

Signature:____ Date:_______________

A.10. Sample Passive Consent Form

Dear Parent or Guardian:

Your child is being asked to complete a survey as part of a study of the performance of the special education personnel preparation program operating at <GranteeName>. This survey will gather information from your child about the performance of his or her special education teacher or related services provider. Please read this form for information about the survey, and for instructions on how to withdraw your child. *If you do not want your child to complete the survey, you must notify your school.*

Survey Content. The survey gathers information on <include brief description of survey contents>. You may examine the questionnaire in the school office or at the following Web site <WebSite>.

All data collected by the surveys will be kept confidential, except as required by law. The data we collect will be combined with the results of surveys of students in other schools and districts. A report summarizing the overall results of the surveys will be delivered to <GranteeName> with results aggregated across all survey respondents.

It is Voluntary. Your child does not have to take the survey. Students who participate only have to answer the questions they want to answer and they may stop answering questions at any time.

It is Anonymous. No names will be recorded or attached to the survey forms or data. The results will be made available only to the researchers using strict confidentiality controls.

Administration. The survey will be administered on <DateSurvey>. It will take about <SurveyTime> to complete and will be administered in your child's <ClassName> class.

Potential Risks. There are no known risks of harm to your child.

Potential Benefits. No direct benefits to your child are expected, however, the study is expected to help improve the quality of training for special education professionals.

For Further Information. The survey was developed by <SurveyDeveloper>. If you have any questions about this survey call me at <EvaluationDirectorTelephone> or by email me at <EvaluationDirectorEmail>. If you have question about your rights related to study participation, call the district at <INSERT NAME AND PHONE NUMBER OF DISTRICT CONTACT>.

If you do not want your child to participate, you may contact: <INSERT CONTACT INFORMATION (E.G., ADDRESS, PHONE NUMBER, E-MAIL).> [*Note: We recommend using a single point of contact*.]

Study Withdrawal Form

By returning this form, I ***do not give permission*** for my child to participate in this study.

(Please Print) My child's name is:________________________________ Grade:________

Teacher's name or Class subject:__________________________________

Signature:____ Date:________________

A.11. Sample Active Consent Form

Dear Parent or Guardian:

Your child is being asked to complete a survey as part of a study of the performance of the special education personnel preparation program operating at <GranteeName>. This survey will gather information from your child about the performance of his or her special education teacher or related services provider. Please read this form for information about the survey, and for instructions on how to withdraw your child. *You must sign this form and return it to your school if you give your permission for your child to participate in this study.*

Survey Content. The survey gathers information on <include brief description of survey contents>. You may examine the questionnaire in the school office or at the following Web site <WebSite>.

All data collected by the surveys will be kept confidential, except as required by law. The data we collect will be combined with the results of surveys of students in other schools and districts. A report summarizing the overall results of the surveys will be delivered to <GranteeName> with results aggregated across all survey respondents.

It is Voluntary. Your child does not have to take the survey. Students who participate only have to answer the questions they want to answer and they may stop taking it at any time.

It is Anonymous. No names will be recorded or attached to the survey forms or data. The results will be made available only to the researchers using strict confidentiality controls.

Administration. The survey will be administered on <DateSurvey>. It will take about <SurveyTime> to complete and will be administered in your child's <ClassName> class.

Potential Risks. There are no known risks of harm to your child.

Potential Benefits. No direct benefits to your child are expected, however, the study is expected to help improve the quality of training for special education professionals.

For Further Information. The survey was developed by <SurveyDeveloper>. If you have any questions about this survey call me at <EvaluationDirectorTelephone> or by email me at <EvaluationDirectorEmail>. If you have question about your rights related to study participation, call the district at <INSERT NAME AND PHONE NUMBER OF DISTRICT CONTACT>.

Study Permission Form

By returning this form, I ***give permission*** for my child to participate in this study.

(Please Print) My child's name is:________________________________ Grade:____________

Teacher's name or Class subject:__________________________________

Signature:____ Date:________________

Appendix B. Professional Standards and Other Educator Resources

Listed below are some links to professional organizations and standards for the different PDP specializations.

Special Education Teachers

Council for Exceptional Children (CEC): https://www.cec.sped.org/

Standards at: http://www.cec.sped.org/Standards/Special-Educator-Professional-Preparation/CEC-Initial-and-Advanced-Preparation-Standards

Special Education Resources for Special Educators: http://serge.ccsso.org/

Standards for Licensing General and Special Education Teachers of students with Disabilities: A resource for state dialog http://serge.ccsso.org/pdf/standards.pdf

National Association of Special Education Teachers: https://www.naset.org/

Board certification in special education: https://www.naset.org/2457.0.html

American Academy of Special Education Professionals: http://aasep.org/

International Association of Special Education: http://www.iase.org/

School Psychologists/School Counselors

National Association of School Psychologists (NASP): http://www.nasponline.org/

Standards for Credentials and Graduate Preparation at: http://www.nasponline.org/standards/2010standards.aspx

Guide for Performance-Based Assessment, Accountability, and Program Development in School Psychology Training Programs (2nd edition). Under related resources at: https://www.nasponline.org/Documents/Guide%20for%20performance%20based.pdf

Speech, Hearing and Language Pathologists

American Speech-Language-Hearing Association (ASHA): http://www.asha.org/

Performance Assessment of Contributions and Effectiveness of Speech-Language Pathologists: http://www.asha.org/Advocacy/state/Performance-Assessment-of-Contributions-and-Effectiveness/

Early Childhood/Early Intervention

Early Childhood Special Education/Early Intervention (birth to age 8) Professional Standards with CEC Common Core October 2008: http://community.fpg.unc.edu/sites/community.fpg.unc.edu/files/instructors/CONNECT-7-Standards.pdf

Alignment of 2009 CEC Initial Common Core, Initial Special Education Professionals in Early Childhood Special Education/Early Intervention (Birth to Eight) (DEC), and 2009 NAEYC Standards for Early Childhood Professional Preparation Programs: http://higherlogicdownload.s3.amazonaws.com/SPED/d221b435-851b-48e2-8872-e54c36c2e03b/UploadedImages/About%20DEC/Prof%20Dev/NAEYC%20DEC%20alignment%20FINAL%2012%2011.pdf

National Association for the Education of Young Children: http://www.naeyc.org/

> Standards for initial and advanced early childhood professional preparation: http://www.naeyc.org/caep/files/caep/NAEYC%20Initial%20and%20Advanced%20Standards%2010_2012.pdf

Council for the Accreditation of Educator Preparation: http://caepnet.org/

> Standards at: http://caepnet.org/standards/introduction

The Early Childhood Personnel Center provides links to the standards of many other organizations: http://ecpcta.org/national/

Teachers of English Learners (ELs)

U.S. Department of Education Office of English Language Acquisition (OELA): http://www2.ed.gov/about/offices/list/oela/index.html

National Clearinghouse for English Language Acquisition (NCELA): http://www.ncela.us/

Teachers of English to Speakers of Other Languages: http://www.tesol.org/

Educating English Language Learners: Building Teacher Capacity Roundtable Report Volume III State Requirements for Pre-service Teachers of ELLs: http://www.ncela.us/files/uploads/3/EducatingELLsBuildingTeacherCapacityVol3.pdf

Moughamian, A. C., Rivera, M. O., & Francis, D. J. (2009). Instructional models and strategies for teaching English language learners. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Retrieved from http://centeroninstruction.org/instructional-models-and-strategies-for-teaching-english-language-learners.

California Department of Education Research Overview Papers:
http://www.cde.ca.gov/sp/cd/ce/documents/dllresearchpapers.pdf

Secondary Transition Specialists

National Alliance for Secondary Education and Transition (NASET): http://www.nasetalliance.org/

National Standards & Quality Indicators for secondary education and transition:
http://www.nasetalliance.org/docs/NASET_8-pager.pdf

National Standards & Quality Indicators: Transition Toolkit for systems improvement:
http://www.nasetalliance.org/docs/TransitionToolkit.pdf

National Center on Secondary Education and Transition: http://www.ncset.org/

Appendix C. Recommended Readings on Research/Evaluation Methodology

Baecher, L. H., & Connor, D. J. (2010). "What do you see?" Using video analysis of classroom practice in a preparation program for teachers of students with learning disabilities. *Insights on Learning Disabilities 7*(2), 5-18.

Barton, E. E., & Reichow, B. (2012). Guidelines for graphing data with Microsoft PowerPoint for Office 2007. *Journal of Early Intervention, 34*, 129-150.

Brewer, J., & Hunter, A. (2006). *Foundations of multimethod research: Synthesizing styles.* Thousand Oaks, CA: Sage.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 1-76). Chicago: Rand-McNally.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Czaja, R., & Blair, J. (2005). *Designing surveys: A guide to decisions and procedures* (2nd ed.). Thousand Oaks, CA: Sage.

Dillman, D., Smyth, J., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys* (3rd edition). Hoboken, NJ: John Wiley & Sons.

Dimitrov, D. M. (2010). *Quantitative research in education: Intermediate and advanced methods* (2nd ed.). New York: Whittier Publications.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6(1)*, 24-67. DOI: 10.1080/19345747.2012.673143

Frechtling, J., with M. Melvin, D. Rog, & E. Johnson. (2010). *The 2010 user-friendly handbook for project evaluation*. Rockville, MD: Westat & Directorate for Education and Human Resources, Division of Research and Learning in Formal and Informal Settings, National Science Foundation.

Groves, R., Flower, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Harkness, J., Braun, M., Edwards, B., & Johnson, T. P. (2010). *Survey methods in multinational, multiregional, and multicultural contexts.* Hoboken, NJ: John Wiley & Sons.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: U.S. Department of Education, Institute for Education Sciences. Retrieved from: http://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.), (Quantitative Methodology Series). New York: Routledge.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston: Pearson Education.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. (2010). *Single-case design technical documentation, Version 1.0 (Pilot)*. Washington, DC: U.S. Department of Education, Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*, 26-38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 122-144.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.). Applied Social Research Methods Series, Vol. 41. Thousand Oaks, CA: Sage.

Maxwell, J. A., & Miller, B. A. (2010). Categorizing and connecting strategies in qualitative data analysis. In S. N. Hesse-Biber & P. Leavy. (Eds.). *Handbook of emergent methods*. New York: The Guilford Press.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.) Thousand Oaks, CA: Sage.

Raudenbush, S. W., et al. (2011). *Optimal Design Software for multi-level and longitudinal research* (Version 3.01) [Software]. Retrieved from: http://hlmsoft.net/od/

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests.* Mahwah, NJ: Lawrence Erlbaum Associates

U.S. Department of Education. (2013). *What Works Clearinghouse Procedures and standards handbook* (Version 3.0). Washington, DC: U.S. Department of Education, Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19#

U.S. General Accounting Office, GAO. (2012). *Designing evaluations: 2012 revision* (Report No. GAO-12-208G). Washington, DC: Author. Retrieved from http://www.gao.gov/assets/590/588146.pdf

U.S. General Accounting Office, GAO. (1992). *Quantitative data analysis: An introduction* (Report to Program Evaluation and Methodology Division, Report No. GAO/PEMD-10.1.11). Washington, DC: Author. Retrieved from http://www.gao.gov/special.pubs/pe10111.pdf

Willis, G. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage.

Appendix D. Outline of Presentation on Data Tracking Systems from July 2012 OSEP Project Directors' Conference

Jill Lammert, Ph.D., Westat
Center to Improve Project Performance (CIPP)
Office of Special Education Programs

Slide 1: Database Considerations

- What kind of database to use?
    - Simple database (e.g., MS Excel)
    - Relational database (e.g., MS Access)
- How to input the data?
    - Manual entry
    - Web form linked to the database
- How to update the data?
    - Point-in-time through manual entry
    - Real-time through web forms
- How to analyze the data?
    - Analyze from within the database (e.g., queries, reports)
        - Descriptive analyses
    - Export to statistical analysis program (e.g., SPSS/SASS/STATA)
        - Descriptive and inferential analyses

Slide 2: Data Access Considerations

- Linking performance of program completers to student achievement requires a <u>unique ID</u> for each graduate <u>linked</u> to student achievement data
    - Individual students (each with own unique ID) - preferable
    - Average student performance by class/grade/school
- Some states have existing teacher evaluation systems that provide information on teacher performance or value-added (e.g., FL, LA)
- Must obtain district approval to access student data
    - Need to have a secure way to transfer and store data
- Obtaining access to student data requires parental consent if individual students can be identified (which may be more likely with special populations)

Slide 3: Other System Considerations

- Web-based platforms enable multiple users to input and access the data simultaneously
    - A relational database can be set-up so that it always reflects what is in the system, rather than having to be updated from a stand-alone database)
- Automatic reminders can facilitate timely data collection (e.g., database generates emails to graduates after a certain time if they haven't entered their student data)
- Pull-down boxes within the database reduce risk of incorrect data entry
- Consider creating "checkpoints" to control data entry (e.g., person inputting data must check box for "obtained consent" before being able to enter student data)

Slide 4: Types of Data to Track

- Preparation program characteristics
- Program completer characteristics
- Program completer performance data
- Characteristics of the program completers' students
- Student performance data
- Comparison data (if available)
- Data collection activities, timelines, and responsibilities

Slide 5: Preparation Program Characteristics

- Institution name
- Program size
- Certification(s) offered
- Course requirements (e.g., # of credits, courses required/offered, student-teaching requirements)

Slide 6: Program Completer Characteristics

- ***Unique ID number***
- Demographic info (e.g., gender, ethnicity)
- Contact information (e.g., first and last name, email, mailing address, telephone, alternate contact information)
- Pre-admission performance (e.g., admission test scores, college GPA)
- Degrees held
- Prior teaching experience
- Preparation program grades/coursework
- Participation in professional development
- Student-teaching assignment
- Prep. program professor supervising student-teaching placement
- Point-in-time during program of student-teaching assignment
- Teaching certification(s)/ endorsement(s)
- Teaching assignment (e.g., school, district, class assignment, teaching position, other teaching responsibilities, support provided, grade level, # of students)

Slide 7: Program Completer Performance Data

- Quality of services provided
    - Observations or videos of classroom teaching
        - Evidence-based practice, data-driven instruction
    - Review of artifacts
        - Exemplary lessons, portfolios
    - Surveys
        - Administrators, students, parents
- Student achievement growth
    - Individual students taught by the graduate
    - Students of a team of teachers (or co-teachers)
    - Grade-level or school-wide average achievement

Slide 8: The Program Completers' Students

- ***Unique student ID (preferable)***
- Student demographics
  - Grade
  - SPED classification
  - Gender
  - Age
  - Ethnicity
  - Eligibility for free and reduced-price meals
  - English learner status

Slide 9: Student Performance Data

- Achievement data
  - State assessments/alternate assessments (testing accommodations)
  - Interim/benchmark assessments
  - Formative (teacher-made tests)
  - Student course grades
  - Student learning objectives (for non-tested subjects)
- Non-academic student outcomes
  - Behavioral (e.g., problem behaviors, disciplinary referrals)
  - Emotional/psychological (e.g., motivation, engagement)

Slide 10: Comparison Data

- Teacher-level (all need unique teacher IDs)
  - Non-program completers (e.g., those who started but didn't finish the preparation program)
  - Other teachers in the completers' schools
  - Prior graduates of the preparation program
- Student level (must be linked to teachers)
  - Students of program completers
  - Students of comparison teachers (see above)

Slide 11: Assignment of Responsibilities

- Contacting graduates
- Contacting graduates' schools
- Obtaining consent
- Developing data collection instruments
- Collecting data
- Inputting data into the database
- Updating data
- Preparing data for analysis

Slide 12: Timelines for Data Collection

- Contact with graduates
  - Initial contact
  - Follow-up contacts
- Administration of surveys/questionnaires/tests
  - First administration
  - Subsequent administrations
- Conduct of observations/interviews
  - First
  - Subsequent

Appendix E. Validity Threats

Threats to Internal Validity

Threats to internal validity relate to whether the study results can actually be attributed to the variables included in the study, or whether some confounding variables might be affecting the outcomes. The principle threats to internal validity include[230]:

- **History**: When study participants experience an event during the study period that might influence their performance. This might occur if all teachers at a school, including new program graduates, are offered an intensive professional development course by their school district in the summer prior to the academic year under study. In this situation it would be difficult to separate the effects of the summer training program from the training provided by the personnel preparation program.
- **Maturation**: During a lengthy study, biological or cognitive maturation of study participants may affect the outcomes of the study. This is particularly an issue when studying young children or assessing performance of skills that can be expected to change as a result of cognitive maturation, such as when assessing the motor skills of children beginning in infancy through age two.
- **Pretesting**: While pretests are common in experimental and quasi-experimental studies, the exposure of study participants to a pretest may influence the results of the posttest due to a practice effect or simply being more aware of the topic under study. An example of this might be when students are given a spelling test to assess their spelling ability, followed by direct instruction in spelling and a posttest to observe changes in the students' performance. One way to avoid this is to use alternate forms of a test for the pre- and the posttests.
- **Measuring instruments**: Using different instruments to collect data, such as standardized tests, observation rubrics, or teacher self-report surveys, can affect the accuracy of scores. This would be of particular concern in cases when different groups of students or teachers are assessed using different measurement instruments, such as two different standardized math tests, and then their results are compared.
- **Statistical regression to the mean**: When a study participant performs extremely well or extremely poorly on a particular test or other measure, it is common that his or her performance will be less extreme on a subsequent test or measure. This is called regression toward the mean and it occurs because the initially observed extreme positive or negative scores contain relatively large (positive or negative) random error that will probably not be as large with subsequent measures, thus causing the subsequent score to be closer to the mean.
- **Differential selection**: In studies that have treatment and control groups, pretest differences between the groups, such as initial reading ability, will carry over to the posttest. If these pretest differences are not accounted for in the analysis—for example, by using a pretest measure of reading ability as a covariate in the analysis—it will not be possible to determine how much of

[230] Dimitrov, 2010; see also Shadish et al., 2002.

the group differences seen at the posttest are due to the pre-existing differences among the groups and how much are due to the treatment.

- **Attrition**: When some study participants drop out of one or more groups in study before it is completed, it is typically not a random process. For example, the parents of a lower-performing student may withdraw their child from an intervention if they fear that the intervention might have an adverse effect on their child's self-confidence. This is also known as *experimental mortality*. Attrition is more of a threat to the internal validity of experimental studies, since quasi-experimental studies already account for the fact that there are differences between the groups under study. Whenever possible, you should collect baseline data for all groups under study and maintain records of the individual participants in the study (e.g., using class rosters to track student enrollment from fall to spring) so that you can examine whether the analytic sample—the sample that remained in your study throughout the entire study period—contains group differences on key variables (such as initial ability) that may call into question the results of your study.
- **Interaction among factors**: Some of the threats to internal validity mentioned above may interact, thus producing additional confounding effects on the results of the experimental study.

Threats to External Validity

Threats to external validity affect your ability to generalize your results to persons, settings, treatments, and outcomes not directly included in the study. The major threats to external validity include[231]:

- **Interaction of selection biases with experimental treatment**: When a treatment is more effective for participants with particular characteristics, the findings cannot be generalized. For example, a computer-based mathematics intervention will likely be more effective for students who already have experience working with computers than for students who have relatively little prior exposure to computers.
- **Reactive effect of pretesting**: When participants take a pretest they become aware of and sensitized to the issues targeted by the treatment. Therefore the post-treatment results may not be generalized to a population of participants that has not received a pretest. This might occur if at the beginning of a student's senior year you administer a pretest measure designed to gauge a student's awareness of available postsecondary options and then administer a similar posttest measure at the end of the student's senior year.
- **Reactive effects of experimental procedures**: Participants in a study often react to the presence of observers and experimental procedures, thereby altering their behavior. This makes it difficult to generalize the findings to persons who are exposed to a treatment in normal settings. An example of this might be when an evaluator installs a video camera in a pull-out classroom to record the interactions of a student with a speech and language pathologist.

[231] Dimitrov, 2010; see also Shadish et al., 2002.

- **Multiple-treatment interference**: When participants in a study are exposed to multiple treatments (or variations of the same treatment), the effect of the second (and any subsequent) treatment might be confounded with residual effects of the preceding treatment. Consequently, the overall outcome of the treatments will depend, among other things, on the sequence in which they were introduced. This might be a factor when a school psychologist is conducting applied behavioral analysis using an alternating treatment single-case design.[232]

[232] See Kennedy, 2005, for more information on the different types of single-case designs.

Appendix F. Works Cited

Allison, P. D. (2001). *Missing data.* Thousand Oaks, CA: Sage Publications.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Baecher, L. H., & Connor, D. J. (2010). “What do you see?” Using video analysis of classroom practice in a preparation program for teachers of students with learning disabilities. *Insights on Learning Disabilities, 7*(2), 5-18.

Bailey, D. B., & Simeonsson, R. J., Jr. (1988). Investigation of use of goal attainment scaling to evaluate individual progress of clients with severe and profound mental retardation. *Mental Retardation*, 26, 289-295.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In T. Urdan & F. Pajares. (Eds.), *Self-efficacy beliefs of adolescents.* Greenwich, CT: Information Age Publishing.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37. DOI: 10.1016/j.jsp.2009.10.001

Barnett, D. W., Daly, E. J. III, Hampshire, E. M., Hines, N. R., Maples, K. A., Ostrom, J. K., & Van Buren, A. E. (1999). Meeting performance-based training demands: Accountability in an intervention-based practicum. *School Psychology Quarterly*, 14, 357-379.

Barth, R. P., Guo, S., & McCrae, J. S. (2008). Propensity score matching strategies for evaluating the success of child and family service programs. *Research on Social Work Practice, 18*(3), 212-222. DOI: 10.1177/1049731507307791

Barton, E. E., & Reichow, B. (2012). Guidelines for graphing data with Microsoft PowerPoint for Office 2007. *Journal of Early Intervention, 34*, 129-150.

Berkowitz, S. (1997). Analyzing qualitative data. In J. Frechtling & L. Sharpe (Eds.). *User-friendly handbook for mixed method evaluations*. Arlington, VA: National Science Foundation. Retrieved from: http://www.nsf.gov/pubs/1997/nsf97153/

Berry, A., Petrin, R. A., Gravelle, M. L., & Farmer, T. W. (2012). Issues in special education teacher recruitment, retention, and professional development: Considerations in supporting rural teachers. *Rural Special Education Quarterly, 30*, 3-11.

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Retrieved April 5, 2016, from: http://files.eric.ed.gov/fulltext/ED540960.pdf

Bill and Melinda Gates Foundation. (2011). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project*. Retrieved April 5, 2016, from: https://docs.gatesfoundation.org/Documents/preliminary-findings-research-paper.pdf

Billingsley, B. S. (2003). *Special education teacher retention and attrition: A critical analysis of the literature* (COPSSE Document No. RS-2). Gainesville, FL: University of Florida, Center on Personnel Studies in Special Education.

Billingsley, B. S. (2004). Special education teacher retention and attrition: A critical analysis of the research literature. *The Journal of Special Education, 38,* 39-55.

Boller, K., Atkins-Burnett, S., Malone, L. M., Baxter, G. P., & West, J. (2010). *Compendium of student, teacher, and classroom measures used in NCEE evaluations of educational interventions. Volume I. Measures selection approaches and compendium development methods* (NCEE 2010-4012). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Bovend'Eerdt, T. J. H., Botell, R. E., & Wade. D. T. (2009). Writing SMART rehabilitation goals and achieving goal attainment scaling: A practical guide. *Clinical Rehabilitation, 23*, 352-361.

Brewer, J., & Hunter, A. (2006). *Foundations of multimethod research: Synthesizing styles.* Thousand Oaks, CA: Sage.

Brownell, M. T., & Jones, N. D. (2015). Teacher evaluation in special education: Approaches, supporting research, and challenges confronted. *Journal of Special Education Leadership, 28*(2), 63-73.

Bruder, M. B., & Stayton, V. (2004). *The Center to Improve Personnel Preparation Policy and Practice in Early Intervention and Preschool Education, Briefing Book*. Farmington, CT: University of Connecticut. Retrieved from: http://www.uconnucedd.org/pdfs/projects/per_prep/PersonnelPrep_Briefing_Book.pdf

Bryington, A. A., Palmer, D. J., & Watkins, M. W. (2002). The estimation of interobserver agreement in behavioral assessment. *The Behavior Analyst Today, 3*(3), 323-328.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*(1), 65-108.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum Associates

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs, *Behavior Modification, 28*(2), 234-246. DOI: 10.1177/0145445503259264

Cardillo, J. E., & Smith, A. (1994). Reliability of goal attainment scores. In T. J. Kiresuk, A. Smith, & J. Cardillo (Eds.), *Goal attainment scaling: Applications, theory and measurement* (pp. 173-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carr, R. A. (1979). Goal attainment scaling as a useful tool for evaluating progress in special education. *Exceptional Children,* 46, 88-95.

Cochran, D. C., Gallagher, P. A., Stayton, V. D., Dinnebeil, L. A., Lifter, K., Chandler, L., K., et al. (2012). Early childhood special education and early intervention personnel preparation standards of the division for early childhood: Field validation. *Topics in Early Childhood Special Education, 38*(1), 38-51. DOI: 10.1177/0271121412436696

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Coladarci, T., & Breton, W. A. (1997). Teacher efficacy, supervision, and the special education resource-room teacher. *The Journal of Education Research, 90*(4), 230-239. Retrieved from: http://www.jstor.org/stable/27542097

Conrad, K. J., & Conrad, K. (1994). Reassessing validity threats in experiments: Focus on construct validity. In K. Conrad. (Ed.), *Critically evaluating the role of experiments* (pp.5-25). San Francisco: Jossey-Bass.

Council for Exceptional Children. (2012). *The Council for Exceptional Children's position on special education teacher evaluation*. Arlington, VA: Author.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., Plano Clark, V., Gutmann, M., & Hanson, W. (2003). Advances in mixed method design. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in the social and behavioral sciences.* Thousand Oaks, CA: Sage.

Czaja, R., & Blair, J. (2005). Designing surveys: A guide to Decisions and Procedures (2nd ed.). Thousand Oaks, CA: Sage.

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2005). *The Sage handbook of qualitative research* (3rd ed.). Thousand Oaks, CA: Sage.

Dimitrov, D. M. (2012). *Statistical methods for validation of assessment scale data in counseling and related fields*. Alexandria, VA: American Counseling Association.

Dimitrov, D. M. (2010). *Quantitative research in education: Intermediate and advanced methods* (2nd ed.). New York: Whittier Publications.

Dimitrov, D. M., Jurich, S., Frye, M., Lammert, J. D., Sayko, S., Taylor, L. Q. (2012, February). *Year one evaluation report/impact study: Illinois Striving Readers*. Arlington, VA: RMC Research Corporation.

Doherty, K. M., & Jacobs, S. (2013, October). *State of the states 2013. Connect the dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67. DOI: 10.1080/19345747.2012.673143

Donmoyer, R. (2012a). Attributing causality in qualitative research: Viable option or inappropriate aspiration? An introduction to a collection of papers. *Qualitative Inquiry, 18*(8), 651-654. DOI: 10.1177/1077800412455012

Donmoyer, R. (2012b). Can qualitative researchers answer policymakers' What-Works question? *Qualitative Inquiry, 18*(8), 662-673. DOI: 10.1177/1077800412454531

Doran, G. T. (1981). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review, 70*, 35-36.

Dymond, S. K., & Bentz, J. L. (2006). Using digital videos to enhance teacher preparation. *Teacher Education and Special Education: The Journal of the Education Division of the Council for Exceptional Children, 29*(2), 98-112. DOI: 10.1177/088840640602900202

Ellis, P. D. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. United Kingdom: Cambridge University Press.

Evans, H. M. (1981). Psychiatric patient participation in goal setting as related to goal attainment. *Dissertation Abstracts International, 41*(12-B, Pt 1), 4658–4659.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94,* 24-28.

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. Washington, DC: National Academy of Education.

Fiore, T., Helba, C., Berkowitz, S., Jones, M., & Newsome, J. (2012). *Making personnel development effective: Using outcome data for program improvement.* Presented at the U.S. Department of Education Office of Special Education Programs Project Directors Conference, July 23, 2012, Washington, DC.

Frechtling, J. A. (2007). *Logic modeling methods in program evaluation*. San Francisco: Jossey-Bass.

Freeman, M. (2000). Knocking on doors: On constructing culture. *Qualitative Inquiry, 6*, 359-369.

Gehrke, R. S., & McCoy, K. (2007). Considering the context: Differences between the environments of beginning special educators who stay and those who leave. *Rural Special Education Quarterly, 26,* 32-40.

Glover, S., Burns, J., & Stanley, B. (1994). Goal attainment scaling as a method of monitoring the progress of people with severe learning disabilities. *Mental Handicap, 22*, 148-150.

Goe, L. & Holdheide, L. (2011). *Measuring teachers' contributions to student learning growth for nontested grades and subjects.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Holdheide, L., & Miller, T. (2011). *A practical guide to designing comprehensive teacher evaluation systems: a tool to assist in the development of teacher evaluation systems.* Washington, DC: National Comprehensive Center for Teacher Quality.

Goldhaber, D. (2013). What do value-added measures of teacher preparation programs tell us? The Carnegie Knowledge Network. Retrieved from: http://www.carnegieknowledgenetwork.org/briefs/teacher_prep/

Goldrick, L., Osta, D., Barlin, D., & Burn, J. (2012). *Review of state policies on teacher induction.* Santa Cruz, CA: New Teacher Center.

Greene, J. C. (2006). Toward a methodology of mixed methods social inquiry. *Research in the Schools, 13*(1), 93-98.

Harris, D. N., & Herrington, C. D. (2015). Editors' Introduction: The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher, 44*, 71-76.

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student test scores (Angoff lecture). Princeton, NJ: Educational Testing Service.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: U.S. Department of Education Institute for Education Sciences. Retrieved on April 5, 2016, from: http://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Heinrich, C., Maffioli, A., & Vazquez, G. (2010). *A primer for applying propensity-score matching: Impact evaluation guidelines* (Technical Notes No. IDB-TN-161). Washington, DC: Inter-American Development Bank.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Holbrook, A. L. (2013). *Agree-disagree questions: Problems and some solutions.* Webinar presented as part of the American Association for Public Opinion Research (APPOR) Online Continuing Education Program, October 23, 2013.

Holdheide, L. (2011). *Challenges in evaluating special education teachers: linking teacher effectiveness to student outcomes.* Urban Special Education Leadership Collaborative, October 27, 2011. Washington, DC: National Comprehensive Center for Teacher Quality.

Holdheide, L., Browder, D., Warren, S., Buzick, H., & Jones, N. (2012). *Using student growth to evaluate educators of students with disabilities: Issues, challenges, and next steps: A forum of state special education and teacher effectiveness experts and researchers*. Retrieved on April 5, 2016, from: http://files.eric.ed.gov/fulltext/ED543814.pdf

Holdheide, L. R., Goe, L., Croft, A., & Reschly, D. J. (2010). *Challenges in evaluating special education teachers and English language learner specialists.* Washington, DC: National Comprehensive Center for Teacher Quality.

Horner, R., & Spaulding, S. (2010). Single-case research designs (pp. 1386–1394). In N. J. Salkind (Ed.), *Encyclopedia of research design*. Thousand Oaks, CA: Sage.

Howell, D. C. (2012, December 12). *Treatment of missing data — Part 1.* [Web log post]. Retrieved from: https://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.), (Quantitative Methodology Series). New York: Routledge.

Jerald, C. (2012). *Ensuring accurate feedback from observations: Perspectives on practice*. Seattle: Bill & Melinda Gates Foundation. Retrieved April 5, 2016, from: http://www.gatesfoundation.org/college-ready-education/Documents/ensuring-accuracy-wp.pdf

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*(2): 137-152. DOI: 10.1037/a0028086.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston: Allyn and Bacon.

Kiresuk, T. J., & Sherman, R. E. (1968). Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. *Community Mental Health Journal, 4*(6), 443-453.

Kiresuk, T. J., Smith, A., & Cardillo, J. E. (Eds.) (1994). *Goal attainment scaling: Applications, theory, and measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Knight, S., Edmondson, J., Lloyd, G., Arbaugh, F., Nolan, J., Whitney, A., & McDonald, S. (2012). Examining the complexity of assessment and accountability in teacher education*. Journal of Teacher Education, 63*(5), 301-303.

Kratochwill, T. R., Elliott, S. N., & Rotto, P. C. (1995). Best practices in school-based behavioral consultation. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-III* (pp. 519-538). Washington, DC: UNational Association of School Psychologists.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. (2010). *Single-case design technical documentation, Version 1.0 (Pilot)*. Washington, DC: Institute for Education Sciences. Retrieved from: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*, 26-38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 122-144.

Lammert, J. D. (2007). *The effects of assistive technology on sound-letter recognition and reading ability*. Unpublished manuscript.

Lazarus, S. S., & Heritage, M. (2016). *Lessons learned about assessment from inclusion of students with disabilities in college and career ready assessments*. Minneapolis, MN: National Center on Educational Outcomes and the National Center on Systemic Improvement.

Leko, M. M., & Smith, S. W. (2010). Retaining beginning special educators: What should administrators know and do? *Intervention in School and Clinic, 45*, 321-352.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. A., et al. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from: http://ies.ed.gov/ncser/pubs/20133000/pdf/20133000.pdf

Loeb, S., & Candelaria, C. A. (2015). *How stable are value-added estimates across years, subjects and student groups?* The Carnegie Knowledge Network. Retrieved from: http://www.carnegieknowledgenetwork.org/briefs/value-added/value-added-stability/

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review, 29*(6), 530-558. DOI: 10.1177/0193841X05275596

Maher, C. A. (1983). Goal attainment scaling: A method for evaluating special education services. *Exceptional Children, 49*, 529-536.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2nd ed.) (Applied Social Research Methods Series, Vol. 41). Thousand Oaks: Sage.

Maxwell, J. A. (2012). The importance of qualitative research for causal explanation in education. *Qualitative Inquiry, 18*(8), 655-661. DOI: 10.1177/1077800412452856

Maxwell, J. A. (2011). *A realist approach for qualitative research.* Thousand Oaks, CA: Sage.

Maxwell, J. A. (2004a). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3-11.

Maxwell, J. A. (2004b). Using qualitative methods for causal explanation. *Field Methods, 16*, 243-264.

McCaffrey, D. F., & Buzick, H. (2014). *Is value-added accurate for teachers of students with disabilities?* Stanford, CA: Carnegie Knowledge Network. Retrieved from: http://www.carnegieknowledgenetwork.org/briefs/teacher_disabilities/

McIntosh, R., Vaughn, S., Schumm, J. S., Haager, D., & Lee, O. (1993). Observations of students with learning disabilities in general education classrooms. *Exceptional Children, 60*(3), 249-261.

McLeskey, J., Tyler, N., & Flippin, S. (2003). *The supply of and demand for special education teachers: a review of research regarding the nature of the chronic shortage in special education (COPSSE Document Number RS-1).* Gainesville, FL: University of Florida, Center on Personnel Studies in Special Education.

Messick, S. (1989). Validity. In R. L. Linn (Ed)., *Educational measurement* (3rd ed., pp 13-103). Phoenix, AZ: The Oryx Press.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks: Sage.

Mitchell, T. & Cusick, A. (1998). Evaluation of a client-centered pediatric rehabilitation programme using goal attainment scaling. *Australian Occupational Therapy Journal, 45,* 7-17.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Method Research, 1*(48), 48-76. DOI: 10.1177/2345678906292462

Morrison, J. Q. (2012). Goal attainment scaling and single-case designs for service delivery accountability: A presentation and dialogue with the OSEP 325k Work Group. Webinar presentation May 2, 2012.

Nance, E., & Calabrese, R. L. (2009). Special education teacher retention and attrition: The impact of increased legal requirements. *The International Journal of Educational Management, 23*(5), 431-440.

Nichols, S. M. C., Bicard, S. C., Bicard, D. F., & Casey, L. B. (2008). A field at risk: The teacher shortage in special education. *Phi Delta Kappan, 89*, 597-600.

Noell, G. H., Brownell, M. T, Buzick, H. M., & Jones, N. D. (2014). Using educator effectiveness measures to improve educator preparation programs and student outcomes (Document No. LS-1). Retrieved from University of Florida, Collaboration for Effective Educator, Development, Accountability, and Reform Center website: http://ceedar.education.ufl.edu/wp-content/uploads/2014/09/LS-1_FINAL_08-27-14.pdf

Northeast Regional Resource Center, Learning Innovations at WestEd & University of the State of New York, The New York State Education Department. (2004, November). *Keeping quality teachers: The art of retaining general and special education teachers.* Albany, NY: Author.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313-324.

Oren, T., & Ogletree, B. T. (2000). Program evaluation in classrooms for students with autism: Student outcomes and program processes. *Focus on Autism and Other Developmental Disabilities, 15,* 170-175.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.) Thousand Oaks, CA: Sage.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS).* Baltimore, MD: Paul H. Brookes.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 38*, 399-416.

Quenemoen, M., Perie, M., & Kearns, J. (2010). *Scoring alternate assessments based on alternate achievement standards: A proposed typology of AA-AAS scoring practices.* Lexington, KY: National Alternate Assessment Center.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173-185.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213.

Raudenbush, S. W., et al. (2011). *Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01)* [Software]. Retrieved from: http://hlmsoft.net/od/

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1): 5-29.

Reybold, L. E., Lammert, J. D., & Stribling, S. M. (2012, November 30). Participant selection as a conscious research method: Thinking forward and the deliberation of 'emergent' findings. *Qualitative Research*, DOI: 1468794112465634

Roach, A. T., & Elliott, S. N. (2005). Goal attainment scaling: An efficient and effective approach to monitoring student progress. *Teaching Exceptional Children, 37*(4), 8-17.

Saris, W., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods, 4*, 61-79.

Schutt, R. K. (2011). Chapter 10. Qualitative data analysis. *Investigating the social world: The process and practice of research* (7th ed.). Thousand Oaks: Sage. Retrieved from: http://www.sagepub.com/upm-data/43454_10.pdf

Schochet, P. Z. (2009). *Technical methods report: The estimation of average treatment effects for clustered RCTs of education interventions* (NCEE 2009-0061 rev.). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Schochet, P. Z. (2008). Statistical power for randomized assignment evaluation of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62-87.

Scriven, M. (2008). A summative evaluation of RCT methodology: And an alternative approach to causal research. *Journal of Multidisciplinary Evaluation, 5*(9), 11-24.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research methodology: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders, 11*, 260-271.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). Singapore: McGraw-Hill.

Sladeczek, I. E., Elliott, S. N., Kratochwill, T. R., Robertson-Mjaanes, S., & Stoiber, K. (2001). Application of goal attainment scaling to a conjoint behavioral consultation case. *Journal of Education and Psychological Consultation, 12*(1), 45-58.

Smith, A., & Cardillo, J. E. (1994). Perspectives on validity. In T. J. Kiresuk, A. Smith, & J. E. Cardillo (Eds.), *Goal attainment scaling: Applications, theory, and measurement* (pp. 243–272). Hillsdale, NJ: Erlbaum.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). Optimal design version 2.0 [Software].

St. John, W. & Johnson, P. (2000). The pros and cons of data analysis software for qualitative research. *Journal of Nursing Scholarship: An Official Publication of Sigma Theta Tau International Honor Society of Nursing/Sigma Theta Tau, 32*(4): 393–397. PMID11140204.

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education, 3*(3), 285-289. DOI: http://dx.doi.org/10.4300/JGME-D-11-00147.1

Sundeen, T., & Wienke, W. D. (2009). A model for recruiting and preparing new rural special educators: Mining an untapped potential. *Rural Special Education Quarterly, 28*, 3-10.

Suter, W. N. (2012). Chapter 12. Qualitative data, analysis, and design. *Introduction to educational research: A critical thinking approach* (2nd ed.). Thousand Oaks: Sage. Retrieved from: http://www.sagepub.com/upm-data/43144_12.pdf

Tashakkori, A., & Creswell, J. W. (2007). Editorial: The new era of mixed methods. *Journal of Mixed Methods Research, 1*(3), 3-7. DOI: 10.1177/2345678906293042

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Thornton, B., Peltier, G., & Medina, R. (2007). Reducing the special education teacher shortage. *The Clearing House, 80*, 233-238.

Tschannen-Moran, M., & Woolfolk Hoy, A. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education, 17*, 783-805.

Turner-Stokes, L. (2009). Goal attainment scaling (GAS) in rehabilitation: A practical guide. *Clinical Rehabilitation, 23*, 362-370.

Tyler, C. E. (2011). *Designing high quality evaluation systems for high school teachers: Challenges and potential solutions.* Washington, DC: Center for American Progress.

Tyler, C. E. (2008). Recruiting teachers for hard-to-fill positions. *Leadership, 37*, 37-38.

U.S. Department of Education. (2014). *What Works Clearinghouse Procedures and standards handbook* (Version 3.0). Washington, DC: U.S. Department of Education, Institute for Education Sciences.

Retrieved March 29, 2016, from: http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

U.S. Department of Education (2015, March). Teacher shortage areas nationwide listing: 1990-1991 through 2015-2016. Washington, DC: Author. Retrieved from: http://www2.ed.gov/about/offices/list/ope/pol/tsa.pdf

Wade, D. T. (2009). Editorial. Goal setting in rehabilitation: An overview of what, why and how. Clinical Rehabilitation, 23(4), 291-295.

Watkins, M. W. & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education, 10*(4), 205-212.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Weiss, R. S. (1994). *Learning from strangers: The art and method of qualitative interview studies*. New York: The Free Press.

Wendt, O. (2009). *Calculating effect sizes for single-subject experimental designs: An overview and comparison*. Presented at the Ninth Annual Campbell Collaboration Colloquium, Oslo, Norway, May 18, 2009. Retrieved December 9, 2012 from: http://www.campbellcollaboration.org/artman2/uploads/1/Wendt_calculating_effect_sizes.pdf

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 609-662.

Yin, R. K. (1994). *Case study research: Design and methods* (2nd ed., Applied Social Research Methods Series, v. 5.). Thousand Oaks, CA: SAGE.